**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Aplicações de técnicas de análise multivariada em experimentos agropecuários
usando o software R**

Simone Daniela Sartorio

Dissertação apresentada para obtenção do título de Mestre
em Agronomia. Área de concentração: Estatística e
Experimentação Agronômica

**Piracicaba
2008**

Simone Daniela Sartorio
Licenciada em Matemática

Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R

Orientador:
Prof. Dr. **CÉSAR GONÇALVES DE LIMA**

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Sartorio, Simone Daniela

Aplicações de técnica de análise multivariada em experimentos agropecuários usando o software R / Simone Daniela Sartorio. - - Piracicaba, 2008.
130 p.

Dissertação (Mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2008.
Bibliografia.

1. Agropecuária 2. Análise de conglomerados 3. Análise multivariada 4. Análise de variância 5. Software I. Título

CDD 519.53
S249s

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

DEDICATÓRIA

Às pessoas que mais AMO:
meus pais, Romilda e Antonio,
minhas irmãs, Débora e Mariane,
minha avó, Sebastiana
e meu namorado, Fernando
OFEREÇO.

AGRADECIMENTOS

Primeiramente agradeço a Deus, que me guiou durante o Mestrado, estendendo sempre suas mãos nos momentos difíceis, me dando força, coragem e saúde pra vencer os obstáculos.

Em especial ao professor César, pelo conhecimento compartilhado, confiança e apoio, tornando possível a realização deste trabalho. Aos professores Jacinta (FZEA) e Gerson (ESALQ) que sempre estiveram prontos a ajudar, cedendo dados e informações, e a todos os professores do departamento que direta ou indiretamente ajudaram na concretização deste trabalho, em especial aos professores Carlos Tadeu, Clarice, Roseli, Décio e Silvio.

Ao aluno Saulo da Zootecnia (FZEA) pela ajuda na classificação dos artigos. Ao Faria (pós-doutorando/ESALQ) pelas conversas proveitosas sobre o tema e sobre o R, e pelos materiais que me disponibilizou. Aos companheiros do grupo *R-stat* que também me ajudaram com o uso do software.

A todos os meus grandes amigos e companheiros, os velhos e os que fiz durante mais esta etapa, pelos maravilhosos momentos que passamos juntos, pela força, alegria e grande ajuda que sempre me deram. Em especial a: Marina, Renata, Mirian, Lúcio, Vanderly, Lucymary, Juliana, Wilson, Pâmela, Josiane, Andréia, Giovana, Fernanda, Ângela, César, Júlio, Édila, Cássio, Michelle, Raphael e Alexsander (irrigação). Não posso esquecer a minha ‘maninha’ Karen (genética), pela grande companhia, força e também pela ajuda com o abstract.

As secretárias Solange e Luciane que sempre me ajudaram tirando dúvidas e resolvendo as pendências burocráticas.

A minha família, que abriu mão de momentos importantes para que eu pudesse realizar este trabalho. Em especial aos meus pais Romilda e Antonio, que sempre me ajudaram nos momentos difíceis, me incentivando e me dando todo apoio que só eles sabem dar.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de estudo.

E a todos aqueles que acreditaram em mim, o meu **MUITO OBRIGADA!!!**

Apesar dos contratempos, TUDO valeu a pena!

SUMÁRIO

RESUMO	7
ABSTRACT	8
1 INTRODUÇÃO	9
2 REVISÃO DE LITERATURA	12
2.1 A Estatística Multivariada	12
2.2 Algumas Considerações e Conceitos Importantes	14
2.2.1 Os Dados Multivariados	14
2.2.2 O Tamanho da Amostra	15
2.2.3 Observações Perdidas ou Incompletas	16
2.2.4 A Distribuição Normal Multivariada	17
2.2.5 Outliers Multivariados	19
2.2.6 Estatísticas Descritivas	20
2.2.7 Modelo Linear Multivariado	24
2.3 As Técnicas Multivariadas	26
2.3.1 Análise de Componentes Principais (<i>Principal Components Analysis</i>)	27
2.3.2 Análise de Correspondência (<i>Correspondence Analysis</i>)	29
2.3.3 Análise de Correlação Canônica (<i>Canonical Correlation Analysis</i>)	31
2.3.4 Análise de Variância Multivariada (<i>Multivariate Analysis of Variance</i>)	33
2.3.5 Análise de Agrupamentos (<i>Cluster Analysis</i>)	36
2.3.6 Análise Fatorial (<i>Factor Analysis</i>)	40
2.3.7 Análise Discriminante (<i>Discriminant Analysis</i>)	45
3 METODOLOGIA	51
3.1 Suporte Computacional	51
3.2 Material e Métodos - Experimento I	52
3.2.1 A Análise de Agrupamentos (AA)	53
3.2.1.1 Métodos de Agrupamento	54
3.2.1.1.1 Métodos Hierárquicos	54
3.2.1.1.2 Métodos Não-Hierárquicos	57
3.2.1.2 Validação e Interpretação dos Resultados	58
3.2.2 Resultados e Discussão - Experimento I	59
3.3 Material e Métodos - Experimento II	67

3.3.1 Análise de Variância Multivariada (MANOVA)	69
3.3.1.1 Comparações Múltiplas	73
3.3.2 Resultados e Discussão - Experimento II	74
3.4 Material e Métodos - Experimento III	84
3.4.1 Análise de Componentes Principais (ACP)	84
3.4.2 Resultados e Discussão - Experimento III	88
4 CONCLUSÕES	96
REFERÊNCIAS	98
APÊNDICES	106
ANEXOS	110

RESUMO

Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R

O uso das técnicas de análise multivariada está reservado aos grandes centros de pesquisa, às grandes empresas e ao ambiente acadêmico. Essas técnicas são muito interessantes porque utilizam simultaneamente todas as variáveis respostas na interpretação teórica do conjunto de dados, levando em conta as correlações existentes entre elas. Uma das principais barreiras para a utilização dessas técnicas é o seu desconhecimento pelos pesquisadores interessados na pesquisa quantitativa. A outra dificuldade é que a grande maioria de softwares que permitem esse tipo de análise (SAS, MINITAB, BMDP, STATISTICA, S-PLUS, SYSTAT, etc.) não são de domínio público. A disseminação do uso das técnicas multivariadas pode melhorar a qualidade das pesquisas, proporcionar uma economia relativa de tempo e de custo, e facilitar a interpretação das estruturas dos dados, diminuindo a perda de informação. Neste trabalho, foram confirmadas algumas vantagens das técnicas multivariadas sobre as univariadas na análise de dados de experimentos agropecuários. As análises foram realizadas com o auxílio do software R, um software aberto, “amigável” e gratuito, com inúmeros recursos disponíveis.

Palavras-chave: Análise multivariada; Correlação; Agropecuária; Análise de agrupamentos; Análise de variância multivariada (MANOVA); Análise de componentes principais; *Software R*

ABSTRACT

Application of multivariate analysis in agricultural experiments using R software

The use of the techniques of multivariate analysis is restricted to large centers of research, the higher companies and the academic environment. These techniques are very interesting because of the use of all answers variables simultaneously in theoretical interpretation of the data set, considering the correlations between them. One of the main obstacle to the usage of these techniques is that researchers interested in the quantitative research do not know them. The other difficulty is that most of the software that allow this type of analysis (SAS, MINITAB, BMDP, STATISTICA, S-PLUS, SYSTAT etc.) are not in public domain. Publishing the use of Multivariate techniques can improve the quality of the research, decrease the time spend and the cost, and make easy the interpretation of the structures of the data without cause damage of the information. In this report, were confirmed some advantages of the multivariate techniques in a univariate analysis for data of agricultural experiments. The analysis were taken with R software, a open software, “friendly” and free, with many statistical resources available.

Keywords: Multivariate statistical; Correlation; Agricultural; Cluster analysis; Multivariate analysis of variance (MANOVA); Principal components analysis; R software

1 INTRODUÇÃO

Muitos processos de experimentação são multivariados, pois envolvem a avaliação de diversas características, ou variáveis respostas, em todas as unidades experimentais.

A denominação “análise multivariada” corresponde a um grande número de métodos e técnicas que utilizam simultaneamente as informações de todas as variáveis respostas na interpretação do conjunto de dados, levando em conta as correlações existentes entre elas. A disseminação do uso das técnicas multivariadas pode melhorar a qualidade das pesquisas, proporcionar uma economia relativa de tempo e de custo, e facilitar a interpretação das estruturas dos dados, diminuindo a perda de informação.

As técnicas de análise multivariadas têm sido regularmente aplicadas em várias investigações científicas nas mais diversas áreas de pesquisa, com maior ou menor freqüência.

Na agronomia, o agrônomo pode estar interessado em prever o genótipo dos frutos, tendo como base informações sobre o número total de frutos por hectare, o peso médio dos frutos, o peso total médio por hectare, o número médio de frutos por planta, o índice de formato e o teor de açúcar. A aplicação de uma análise discriminante é indicada neste estudo (BARROSO; ARTES, 2003).

Em biologia, no melhoramento de plantas é necessário, após o final de uma geração, selecionar aquelas plantas que serão os genitores da geração seguinte. A seleção deve ser realizada de maneira que a próxima geração seja melhor que a anterior em relação à resposta média de uma série de características. Desta maneira, o objetivo do melhorista consiste em maximizar o ganho genético em um espaço mínimo de tempo. A análise fatorial, ou a análise de componentes principais, por exemplo, podem ser usadas para converter os valores de uma série de características em um índice, que é utilizado para a escolha dos futuros pais (LEDO; FERREIRA; RAMALHO, 2003).

Nas ciências médicas, como as reações de pacientes a um determinado tratamento são mensuradas por diversas variáveis respostas e possuem difícil diagnóstico, as técnicas multivariadas podem construir uma medida de resposta simples ao tratamento, que preserva a maior parte da informação das múltiplas variáveis respostas. Em outras situações as técnicas multivariadas podem ser usadas também quando a classificação de um paciente, baseada em diversos sintomas medidos, é difícil de ser realizada. Neste caso, uma técnica multivariada de classificação, em que se cria uma função que pode ser usada para separar as pessoas doentes das não doentes, pode ser implementada.

Na economia, a concessão de crédito a consumidores é uma prática realizada por bancos, supermercados, lojas de varejo e outras organizações. Essa decisão é usualmente fundamentada em técnicas de classificação que consistem em rotular o cliente como “bom” ou “ruim”. As variáveis observadas e nas quais são baseadas as decisões, são variáveis sócio-econômicas como: estado civil, nível educacional, sexo, se o cliente é proprietário da casa em que reside, etc. Como o volume de informações é grande, para auxiliar na tomada de decisão sobre um novo cliente, pode-se comparar suas características com as de clientes que já provaram ter pago suas contas em dia e as de outros, que ficaram inadimplentes. Baseado nas semelhanças e/ou diferenças das características com esses dois grupos, pode-se fazer uma previsão de comportamento de novos clientes (BARROSO; ARTES, 2003, e JOHNSON; WICHERN, 2002).

Na engenharia, no caso da engenharia civil, pode-se fazer a avaliação de imóveis levando em consideração inúmeros fatores como: sua localização, características de infra-estrutura do imóvel, idade do imóvel, categorização da acessibilidade, padrão de acabamento, estado de conservação etc. As variáveis importantes na formação do valor de uma determinada categoria ou subconjunto de imóveis podem não ser as mesmas para outro subconjunto, inclusive dentro da mesma região. A análise fatorial e a análise de agrupamentos podem ser utilizadas nestes casos (TRIVELLONI; HOCHHEIM, 1998), identificando-se as correspondências entre as diferentes variáveis, os fatores comuns principais, a combinação destas variáveis e também as possíveis classes ou tipos de imóveis.

Já na zootecnia, pode-se estudar o relacionamento entre as características de manejo e de criação de gado, ou de frangos de corte, com as características da carne após o abate, utilizando a técnica de análise de correlação canônica. Se os animais estiverem separados em grupos de acordo com algum tratamento, pode ser necessário utilizar a análise de variância multivariada para comparar as médias desses grupos em um determinado conjunto de variáveis respostas. Na produção de carne, por exemplo, observam-se várias características para que seja avaliada a qualidade da carne, como: perda de peso por gotejamento, perda de peso por cozimento, gordura intramuscular, maciez, etc. Muitas dessas variáveis são redundantes, o que acaba dificultando o trabalho. Pela análise de componentes principais pode-se reduzir o número de variáveis observadas, sem que haja perda demasiada de informação simplificando a avaliação da qualidade da carne.

Contudo, os dados resultantes de pesquisas com temas agropecuários, como nutrição e alimentação animal, fisiologia animal, melhoramento genético animal ou vegetal, produção e tecnologia de alimentos, produção e qualidade de plantas forrageiras, dentre outras, são analisados

por diversas técnicas estatísticas, sendo mais freqüente o uso da análise de variância univariada e da análise de regressão linear, múltipla ou não-linear. Como parece natural supor a existência de correlação entre essas variáveis, as informações providas por diversas análises univariadas podem resultar incompletas ou pouco informativas.

O uso de técnicas de análise multivariada, geralmente, fica reservado aos centros de pesquisa, às grandes empresas e ao ambiente acadêmico. Uma das principais barreiras para a utilização dessas técnicas é o seu desconhecimento pelos pesquisadores interessados na pesquisa quantitativa. Outra dificuldade é que a grande maioria de softwares que permitem esse tipo de análise (SAS, MINITAB, BMDP, STATISTICA, S-PLUS, SYSTAT etc.) não são de domínio público.

Este trabalho tem por objetivo utilizar técnicas multivariadas para re-analisar dados de experimentos agropecuários já publicados e analisados por técnicas univariadas, no intuito de confrontar os resultados obtidos, proporcionando uma nova visão, ou seja, uma releitura dos resultados. Com isso, pretende-se mostrar que as técnicas de análise multivariada são muito informativas e podem facilitar a discussão dos resultados. As análises serão implementadas utilizando o software gratuito e aberto, R (2007).

2 REVISÃO DE LITERATURA

O intuito desta revisão é apresentar aplicações das técnicas multivariadas, dando uma breve idéia do que se pode conseguir com o uso destas ferramentas estatísticas, estimulando seu uso no meio agropecuário. A revisão bibliográfica foi realizada em trabalhos brasileiros e evidencia o fato de que no Brasil estas técnicas ainda são pouco utilizadas neste meio.

Um levantamento sobre as publicações em três revistas brasileiras bem conceituadas, durante os anos de 2000 a 2007 revelou que:

- A Revista *Bragantia* publicou 437 artigos, dos quais apenas 7 utilizaram alguma técnica multivariada, o que representa 1,6% de sua publicação;
- A Revista *Scientia Agricola*, neste mesmo período, publicou 859 artigos e apenas 13 (1,5%) destes utilizaram alguma das técnicas multivariadas;
- Já na Revista da Sociedade Brasileira de Zootecnia, somente 32 (1,5%) artigos dos 2177 publicados usaram alguma abordagem multivariada.

Logo, as três revistas durante estes últimos oito anos publicaram 3473 artigos, dos quais apenas 52 (1,5%) usaram alguma técnica multivariada. Destes 52 artigos: 19 (36,5%) são referentes a área de Melhoramento Genético Animal e Vegetal; 13 (25%), à Produção Animal; 7 (13,4%), a Solos e Nutrição de Plantas; 4 (7,7%), a Forragicultura; 4 (7,7%), a Sistemas de Produção Agronegócio e Economia Aplicada; 2 (3,8%), a Fisiologia Vegetal; 1 (1,9%), a Agrometeorologia; 1 (1,9%), a Aqüicultura e 1 (2,1%), a Estatística Aplicada.

Na seção 2.1 é apresentada uma breve revisão histórica da estatística multivariada, evidenciando as dificuldades de implementação e as suas particularidades. Na seção 2.2 são apresentados os conceitos básicos multivariados e algumas preocupações importantes sobre o uso de qualquer técnica multivariada. A visão de autores sobre as técnicas multivariadas é mencionada na seção 2.3, como também trabalhos científicos da área agropecuária, e em cada subseção deste item (2.3) é feita uma introdução histórica de algumas técnicas relevantes de análise multivariadas.

2.1 A Estatística Multivariada

A Estatística Multivariada é definida como um conjunto de métodos estatísticos utilizados em situações nas quais várias variáveis são medidas simultaneamente em cada unidade

experimental, ou seja, é o ramo da estatística que tem por objetivo o resumo, a representação, a análise e a interpretação de dados amostrados de populações nas quais para cada unidade experimental são avaliadas *diversas variáveis respostas*, contínuas ou não. Os métodos de análise de dados multivariados permitem um estudo global dessas variáveis, colocando em evidência as ligações, semelhanças ou diferenças entre elas, perdendo o mínimo de informação.

As primeiras idéias sobre a análise de dados utilizando múltiplas respostas, surgem das contribuições de Pearson (1901), Fisher (1928), Hotelling (1931), Wilks (1932) e Bartlett (1937) (Figura 1), que iniciam o desenvolvimento de procedimentos analíticos para tratar dessas situações.

Figura 1 - Da esquerda para direita: Karl Pearson (Inglaterra - 1857★ ; 1936†), Ronald Aylmer Fisher (Inglaterra/Austrália - 1890★ ; 1962†), Harold Hotelling (EUA - 1895★ ; 1973†), Samuel Stanley Wilks (EUA - 1906★ ; 1964†) e Maurice Stevenson Bartlett (Inglaterra - 1910★ ; 2002†)

O fato de que nenhuma variável consegue individualmente caracterizar de maneira adequada a unidade experimental, ou consegue discriminar indivíduos com relação a qualquer critério que seja empregado, é a principal justificativa do experimentador para medir diversas variáveis em cada unidade. A necessidade de compreensão das relações entre as diversas variáveis faz com que as análises multivariadas sejam complexas ou até mesmo difíceis, embora sejam de grande utilidade aos pesquisadores. Elas podem ser muito informativas e podem facilitar a discussão dos resultados.

Em geral, as variáveis medidas nas mesmas unidades experimentais estão relacionadas entre si e quanto maior o número de variáveis, mais complexa torna-se a análise por métodos comuns de estatística univariada. Além disso, desde que as diversas medidas feitas em cada uma das unidades experimentais são correlacionadas, é inapropriado aplicar análises univariadas separadamente, para cada uma das variáveis.

Hair Jr. et al. (2006) reforçam a idéia que qualquer pesquisador que examine apenas

relações entre duas variáveis e evite o uso de análise multivariada estará ignorando poderosas ferramentas que podem lhe dar informações potencialmente úteis. Já diziam os estatísticos Hardyck e Petrinovich¹ em 1976 que, os métodos de análise multivariada predominariam no futuro e resultariam em drásticas mudanças na maneira como os profissionais de pesquisa pensariam seus problemas e planejavam suas pesquisas.

Devido à grande quantidade de cálculos necessários para a realização das análises, sua expansão só foi possível graças ao avanço da tecnologia computacional e ao grande número de softwares estatísticos com módulos de análise multivariada implementados, como por exemplo: SAS, MINITAB, BMDP, STATISTICA, S-PLUS e SYSTAT. O software R, em particular, se torna um grande aliado, porque é um software “amigável”, de domínio público (gratuito) e aberto, com inúmeros recursos disponíveis.

2.2 Algumas Considerações e Conceitos Importantes

Nesta seção são apresentados alguns conceitos imprescindíveis na aplicação das técnicas de análise multivariada.

2.2.1 Os Dados Multivariados

Definição 2.1 : Define-se uma *observação multivariada* como sendo uma coleção de medidas de p variáveis feitas num mesmo indivíduo.

A representação desses dados é feita com a notação y_{ij} para indicar um valor particular da i -ésima unidade amostral ou experimental e da j -ésima variável mensurada. Conseqüentemente, as n medidas de p variáveis podem ser arranjadas numa matriz retangular $\mathbf{Y}$, com n linhas e p colunas:

$$\mathbf{Y}_{(n \times p)} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{y}'_1 \\ \mathbf{y}'_2 \\ \vdots \\ \mathbf{y}'_n \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_p \end{bmatrix},$$

¹HARDYCK, C.D., PETRINOVICH, L.F. **Introduction to Statistics for the Behavioral Sciences.** 2^a. ed., Philadelphia: Saunders, 1976.

onde $\mathbf{y}_i$ é um vetor $(p \times 1)$ de observações das p variáveis no indivíduo i ($i = 1, 2, \dots, n$); e $\mathbf{y}_j$ é um vetor $(n \times 1)$ de observações da variável j nos n indivíduos.

Desta forma, cada indivíduo i é representado no espaço p -dimensional por um ponto, onde suas coordenadas são dadas por $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ip})'$, para $i = 1, 2, \dots, n$.

Vale lembrar que uma adequada organização e avaliação dos dados são essenciais para que a análise multivariada seja aplicada de forma correta. Para tanto, alguns fatores devem ser levados em consideração, como por exemplo: deve-se investigar a forma como os dados foram gerados, as medidas utilizadas e a confiabilidade destes dados.

O tipo dos dados em estatística multivariada também é de grande importância, pois podem ajudar a definir a técnica a ser aplicada. De acordo com Hair Jr. et al. (2006), os dados podem ser de dois tipos:

- i) **Dados métricos:** Também chamados de dados quantitativos, dados intervalares ou dados proporcionais, essas medidas identificam ou descrevem indivíduos (ou objetos) não apenas na posse de um atributo, mas também pela quantia ou grau em que o indivíduo pode ser caracterizado pelo atributo. Por exemplo, a idade ou o peso de um bovino de corte.
- ii) **Dados não-métricos:** Também chamados de dados qualitativos, são atributos, características ou propriedades categóricas que identificam ou descrevem um indivíduo ou objeto. Diferem dos dados métricos no sentido de indicarem a presença de um atributo, mas não a quantia. São também conhecidos como dados nominais ou dados ordinais. Exemplos: cor de pelagem de equinos (tordilho, alazão, castanho, malhado, baio etc.), raça de bovinos de corte (Hereford, Charolesa, Limonsin etc.), etc.

Como nem sempre é possível que todas as variáveis respostas métricas estejam definidas na mesma unidade de medida e assim, evitar que ordens de grandezas (escalas) diferentes entre as variáveis prejudiquem as análises (ou seja, nos casos em que exista uma grande diferença entre as variâncias das variáveis originais), sugere-se que a análise escolhida seja realizada com os dados das variáveis padronizadas (ver Anexo A).

2.2.2 O Tamanho da Amostra

Segundo Hair Jr. et al. (2006), o tamanho da amostra afeta todos os resultados de uma análise. Para amostras pequenas, a sofisticação e complexidade da técnica multivariada

podem facilmente resultar em um baixo poder estatístico para o teste identificar realisticamente resultados significantes ou, um “ajuste” muito fácil dos dados, de modo que os resultados são artificialmente bons, mas sem poder de generalização. Um impacto semelhante também ocorre para amostras muito grandes, as quais podem tornar os testes estatísticos muito sensíveis. Sempre que tamanhos de amostras excederem 200 ou 400 indivíduos o pesquisador deverá examinar todos os resultados significantes, garantindo que estes tenham significado prático devido ao poder estatístico aumentado pelo tamanho da amostra.

No geral, em estatística multivariada, como a maioria dos resultados é assintótico, tem-se melhores resultados se o experimento a ser analisado tiver um grande número de elementos observados (n). Os autores como Johnson e Wichern (2002), Mingoti (2005), dentre outros, consideram “ n grande” como sendo $n > 50$. Uma outra condição que precisa ser satisfeita é de $(n - p) > 50$, onde p é o número de variáveis respostas avaliadas. Esta última condição evita o mau condicionamento (*degeneração*) das matrizes de dados a serem utilizadas. Se $(n - p) < 50$ as técnicas multivariadas devem ser utilizadas com cautela.

Já Hair Jr. et al. (2006) colocam como regra geral que, o mínimo é ter pelo menos cinco vezes mais observações do que o número de variáveis a serem analisadas, e o tamanho mais aceitável teria uma proporção de dez observações para uma variável. Além do tamanho da amostra geral, estes autores completam ainda que, em técnicas onde compara-se grupos, como a MANOVA e a Análise Discriminante, é recomendado que o pesquisador também considere o tamanho da amostra de cada grupo. Como uma orientação prática, cada grupo deve ter no mínimo 20 observações. Mas mesmo que todos os grupos excedam a 20 observações, o pesquisador também deve considerar os tamanhos relativos dos grupos. Se os grupos variam muito em tamanho, isso pode influenciar os resultados. Logo, um pesquisador ou usuário de técnicas multivariadas sempre deve avaliar os resultados à luz do tamanho da amostra usada na análise.

2.2.3 Observações Perdidas ou Incompletas

A maioria das técnicas multivariadas utiliza somente os indivíduos com informações completas, ou seja, se para um indivíduo o valor de alguma variável tiver sido perdido, este indivíduo é eliminado do processo de análise. Sendo assim, Hair Jr. et al. (2006) lembram que em muitas análises multivariadas, os dados perdidos podem eliminar tantas observações que uma amostra que era adequada, fica reduzida a uma amostra imprópria.

Como há poucas orientações sobre o reparo de dados perdidos, os pesquisadores mesmo reunindo observações adicionais, preferem a solução mais prática, que consiste em eliminar os indivíduos incompletos. Assim sendo, deve-se evitar procedimentos de coleta de dados que, de algum modo, contribuam para que a perda de dados ocorra com maior frequência (MINGOTI, 2005).

2.2.4 A Distribuição Normal Multivariada

Muitos dos métodos estatísticos multivariados baseiam-se no pressuposto de que os dados são retirados de uma população com distribuição normal multivariada, que é uma generalização da distribuição normal univariada para $p \geq 2$. Sabe-se também que, mesmo quando os dados não seguem uma distribuição exatamente normal, é quase sempre possível aproximar a distribuição real à normal.

Segundo Barroso e Artes (2003), um vetor aleatório p -dimensional, $\mathbf{y}$, segue uma distribuição normal multivariada com vetor média $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$, positiva definida (Anexo B) - denota-se: $\mathbf{y} \sim N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, se sua função densidade de probabilidade for dada por:

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right] .$$

O produto $(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ é conhecido como Distância Generalizada de Mahalanobis (mais detalhes no Anexo C).

A distribuição normal multivariada de probabilidade é gerada no software R por meio do pacote *mvtnorm* (mais detalhes em GENZ; BRETZ; HOTHORN, 2006).

A análise das distribuições marginais univariadas e bivariadas auxiliam na verificação da suposição de multinormalidade. O fato de se garantir que todas as distribuições univariadas e bivariadas são normais, não implica necessariamente que o vetor aleatório $\mathbf{y}_i$ tenha distribuição normal multivariada (ANDERSON, 2003 apud MINGOTI, 2005). Porém, se o vetor aleatório $\mathbf{y}_i$ tem distribuição normal multivariada, então todas as distribuições univariadas e bivariadas são normais. Na prática, quando as distribuições uni e bivariadas são normais, a chance de se ter um vetor normal p -variado é muito grande. Estas suposições de normalidade's podem ser verificadas por:

- i) **A normalidades univariada das p -variáveis:** pode ser avaliada por gráficos de probabilidade normal (*normal plot*), por histogramas ou *boxplot's*, ou em testes de aderência, como o

de Shapiro-Wilk² (1965). Transformar os dados originais é uma alternativa utilizada quando os dados não provêm de uma distribuição normal univariada. Uma classe de transformações muito utilizada é a de Box e Cox³ (1964) e detalhes podem ser encontrados em Johnson e Wichern (2002). No entanto, nem sempre é possível obter-se uma transformação que cumpra a tarefa adequadamente.

No R, o comando *pairs* faz somente os gráficos de dispersão das variáveis. Com maiores recursos, o comando *scatterplot.matrix* do pacote *car* faz o mesmo, porém unindo histogramas, *boxplot's* ou *normal plot's* aos gráficos de dispersão. Em resumo, o pacote *car* é o mais indicado para a verificação desta suposição. Maiores detalhes sobre seu uso são encontrados em Fox (2007).

- ii) **A normalidades bivariada das p -variáveis:** pode ser avaliada através da construção de gráficos de dispersão do tipo Y_j versus $Y_{j'}$, $j \neq j'$ e $j, j' = 1, 2, \dots, p$. De acordo com a teoria, todos os pares de variáveis $(Y_j, Y_{j'})$ devem ter uma distribuição normal bivariada e, portanto, os gráficos de dispersão devem indicar a forma de uma elipse. Sugere-se também o uso de gráficos do tipo *boxplot bivariado* (ver Anexo D).
- iii) **A normalidade multivariada:** é verificada pelo gráfico de probabilidade *Q-Q plot* (Anexo E). Quando a normalidade p -variada se ajusta bem aos dados amostrais, este gráfico deve resultar em uma nuvem de pontos próximos a uma reta. Curvas diferentes da reta revelam ausência de normalidade.

Um outro procedimento usado para a detecção da normalidade multivariada, que não será utilizado neste trabalho, está relacionado aos testes de hipóteses propostos por Mardia⁴ (1970), que são fundamentados nos coeficientes de assimetria e curtose da distribuição normal multivariada (mais detalhes também em FERREIRA, 1996).

²SHAPIRO, S.S.; WILK, M.B. An analysis of variance test for normality. **Biometrika**, Cambridge, v. 52, p. 591-611, 1965.

³BOX, G.E.P.; COX, D.R. An Analysis of Transformations. **Journal of the Royal Statistical Society**. Londres, v. 26, p. 211-243, 1964.

⁴MARDIA, K.V. Measures of skewness and kurtosis with applications. **Biometrika**, Cambridge, v. 57, p. 519-530, 1970.

2.2.5 Outliers Multivariados

Como no caso univariado, antes de aplicar algum método multivariado deve-se investigar a existência de valores discrepantes (*outliers*), que podem afetar os resultados finais da análise estatística. Logo, é fundamental que seja feita uma análise exploratória dos dados na tentativa de identificar pontos desse tipo.

Em dados multidimensionais, uma observação é considerada *outlier* se está muito distante das restantes no espaço p -dimensional definido pelas variáveis (FIGUEIRA, 1998), ou seja, deve ser uma observação não representativa da população, devendo, portanto, apresentar valores extremos em diversas variáveis e não apenas em uma ou outra. Cabe salientar que é preciso ter muito cuidado com estes *outliers multivariados*, pois é possível que uma observação seja considerada um ponto discrepante em termos multivariados e não o seja em termos univariados.

As observações atípicas podem ser identificadas sob uma perspectiva univariada, bivariada ou multivariada. A perspectiva univariada é aquela usual, para o caso de uma única variável (RENCER; SCHAALJE, 2008). A bivariada se refere aos gráficos de dispersão bidimensionais, aliados a elipses de confiança (maiores detalhes em JOHNSON; WICHERN, 2002; e EVERITT, 2005) - *boxplot bivariado* (Anexo D). Já na perspectiva multivariada, gráficos de dispersão tridimensionais auxiliam na identificação de *outliers*, juntamente com a Distância de Mahalanobis, D_i^2 e gráficos do tipo *Q-Q plots* (Anexo E) - MINGOTI (2005).

O uso da distância de Mahalanobis (Anexo C) é sugerido por muitos textos como um método para detectar *outliers* em dados multivariados. Para indicar valores críticos de *outliers*, baseados em D_i^2 , é sugerido a estatística de teste $[p(n-1)/(n-p)]F_{(p,n-p,\alpha)}$, isto é, valores de D_i^2 maiores que o valor crítico desta estatística são considerados *outliers*. Esta aproximação F é considerada mais adequada do que a distribuição $\chi_{(p,\alpha)}^2$, especialmente quando se lida com pequeno número de indivíduos. No entanto, Penny (1996) afirma que, na prática esta distribuição F é inadequada para testar *outliers* multivariados em pequenas amostras.

Identificadas como observações atípicas, por meio destes métodos, o pesquisador deve selecionar as que mostram verdadeira peculiaridade em comparação com o restante da população.

O comando *mahalanobis* no R, calcula a Distância de Mahalanobis. O pacote *rgl* do software R pode ser muito útil na confecção de gráficos de dispersão das variáveis, como também os pacotes *car*, *rggobi* e *scatterplot3d*, dentre outros. Detalhes do uso destes pacotes são encontrados em Fox (2007), Adler e Murdoch (2007), Lang e Swayne (2007) e Ligges (2007), respectivamente.

Algumas técnicas multivariadas são também grandes aliadas na detecção de *outliers*,

como:

- **Análise de Agrupamentos Hierárquicos:** depois de realizado o agrupamento, pode-se identificar grupos formados por apenas um elemento. Cada um destes elementos pode ser classificado como possível *outlier*, pois nenhuma outra observação foi considerada similar para ser colocada no mesmo grupo destas observações suspeitas.
- **Análise de Componentes Principais:** utilizam-se os escores das últimas componentes principais para a confecção de gráficos de dispersão, bi e tridimensional e *Q-Q plots*. Este método se justifica pelo fato de que a magnitude dos últimos componentes principais determina quão bem os primeiros se ajustam as observações. Na prática, as observações suspeitas serão aquelas que, no gráfico de dispersão dessas últimas componentes, se encontrarem distantes da nuvem de pontos (mais detalhes em JOHNSON; WICHERN, 2002).

2.2.6 Estatísticas Descritivas

Grandes conjuntos de dados dificultam a obtenção de informações pertinentes ao fenômeno estudado. Muitas informações contidas nos dados podem ser resumidas nos valores de estatísticas descritivas. As estatísticas descritivas que mensuram posição, variação e associação linear, calculadas em n observações de p variáveis, são organizadas em vetores e matrizes e são descritas a seguir:

a) **Vetor de Médias Amostrais:** é um vetor que contém a média de cada uma das p variáveis e obtido da seguinte forma:

$$\bar{\mathbf{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \end{bmatrix} = \frac{1}{n} \begin{bmatrix} y_{11} & y_{21} & \cdots & y_{n1} \\ y_{12} & y_{22} & \cdots & y_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1p} & y_{2p} & \cdots & y_{np} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n} \mathbf{Y}' \mathbf{1}$$

onde $\mathbf{1}$ é um vetor ($n \times 1$) de “uns”. Geometricamente, considere o vetor de “uns”, que forma ângulos iguais com cada um dos n eixos coordenados. A projeção ortogonal do vetor

$\mathbf{y}'_j = [y_{1j} \ y_{2j} \ \dots \ y_{nj}]$, $j = 1, 2, \dots, p$, no vetor de comprimento unitário $(1/\sqrt{n}) \mathbf{1}$ é igual a:

$$\mathbf{y}'_j \left(\frac{1}{\sqrt{n}} \mathbf{1} \right) \frac{1}{\sqrt{n}} \mathbf{1} = [y_{1j} \ y_{2j} \ \dots \ y_{nj}] \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \left(\frac{1}{n} \right) \mathbf{1} = \bar{y}_j \mathbf{1}$$

isto é, a média $\bar{y}_j$ corresponde ao múltiplo de $\mathbf{1}$ necessário para dar a projeção de $\mathbf{y}_j$ sobre a linha determinada por $\mathbf{1}$ (JOHNSON; WICHERN, 2002).

No software R o comando `?colMeans`⁵ (para matrizes), ou `mean` (para lista de dados - data.frame) calcula o vetor de médias.

b) Matriz de Variâncias e Covariâncias Amostras: Com uma única variável, a variância da amostra é usada para descrever a variabilidade nas mensurações desta variável em relação à média. Quando p variáveis são observadas em cada unidade da amostra ou do experimento, as variâncias e as covariâncias são descritas pela *matriz de variâncias e covariâncias amostrais* $\mathbf{S}$, sendo obtida por:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix} = \frac{1}{n-1} \mathbf{Y}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{Y}$$

onde $\mathbf{I}$ é uma matriz identidade de dimensão p ; $s_{jj} = \text{var}(Y_j)$ é a *variância da j -ésima variável* e $s_{jj'} = \text{cov}(Y_j, Y_{j'})$ é a *covariância amostral* entre as variáveis j e j' , com $j, j' = 1, 2, \dots, p$ e $j \neq j'$. Além disso, $s_{jj'} = s_{j'j}$, para todo j e j' . Desta maneira, a matriz $\mathbf{S}$ contém p variâncias e $\frac{1}{2}p(p-1)$ covariâncias, potencialmente diferentes. A interpretação de $s_{jj'}$ é realizada observando o seguinte:

- i) Se grandes (pequenos) valores de uma variável são observados em conjunto com grandes (pequenos) valores da outra variável, $s_{jj'}$ será positiva;
- ii) Se grandes (pequenos) valores de uma variável ocorrem com pequenos (grandes) valores da outra, $s_{jj'}$ será negativa;

⁵O caracter “?” antecedendo qualquer comando do software R solicita o menu de ajuda, mostrando o que faz o comando, como usá-lo e suas opções.

- iii) Se os dados estão dispersos sem indicar uma direção, não há associação entre os valores das duas variáveis, $s_{jj'}$ será próximo de zero.

No R, a matriz de variâncias e covariâncias de um certo conjunto de dados é obtida pelo comando *cov*.

- c) **Desvio Padrão Amostral:** é uma medida de variação expressa na mesma unidade de medida das observações e é calculada como a raiz quadrada da variância amostral ($\sqrt{s_{jj}}$). Geometricamente, o comprimento (ou norma) do vetor de desvios (ou residual) é proporcional ao desvio padrão:

$$\mathbf{d}_j = \mathbf{y}_j - \bar{y}_j \cdot \mathbf{1}, \text{ para } j = 1, 2, \dots, p.,$$

onde $\mathbf{d}_j$ é o vetor ($n \times 1$) de desvios das observações da variável j em relação à sua média $\bar{y}_j$.

- d) **Matriz de Correlações Amostrais:** O coeficiente de correlação linear de Pearson amostral entre as variáveis j e j' é

$$r_{jj'} = \text{corr}(Y_j, Y_{j'}) = \frac{s_{jj'}}{\sqrt{s_{jj}}\sqrt{s_{j'j'}}}.$$

A matriz de correlações amostrais $\mathbf{R}$ pode ser obtida por:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} = \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2}$$

onde $\mathbf{D} = \text{diag}(\mathbf{S})$ é a matriz formada pelos elementos da diagonal da matriz $\mathbf{S}$. A matriz $\mathbf{S}$ também pode ser obtida a partir da matriz $\mathbf{R}$ fazendo:

$$\mathbf{S} = \mathbf{D}^{1/2} \mathbf{R} \mathbf{D}^{1/2}.$$

O coeficiente de correlação possui as seguintes propriedades:

- i) $-1 \leq r_{jj'} \leq 1$;
- ii) $r_{jj'} = 0$, implica em inexistência de associação linear entre as variáveis;
- iii) O sinal de $r_{jj'}$ indica a direção da associação: se $r_{jj'} < 0$, há uma tendência de um dos valores do par ser maior (menor) que sua média, quando o outro for menor (maior) do que

a sua média, e $r_{jj'} > 0$ indica que quando um valor do par for grande (pequeno) o outro também o será;

- iv) Os valores de $r_{jj'}$ não se alteram com a alteração da escala de uma das variáveis, pois é uma medida de associação linear entre duas variáveis que não depende das suas unidades de mensuração.

Geometricamente a correlação amostral $r_{jj'}$, corresponde ao co-seno do ângulo formado pelos vetores de desvios $\mathbf{d}_j$ e $\mathbf{d}_{j'}$. Quando $\mathbf{d}_j$ e $\mathbf{d}_{j'}$ têm orientações muito parecidas e o ângulo entre eles for muito pequeno, a correlação amostral entre as variáveis j e j' será próxima de 1. Se $\mathbf{d}_j$ e $\mathbf{d}_{j'}$ têm orientações opostas e o ângulo entre eles é próximo a 180° , a correlação amostral entre as variáveis j e j' será próxima de -1 (JOHNSON; WICHERN, 2002).

No software R, a matriz de correlações amostrais é obtida pelo comando *cor*.

- e) **Variância Amostral Generalizada:** é usada quando deseja-se expressar a variação de todas as variáveis por um único valor numérico. A *variância amostral generalizada* (VG) é definida como:

$$VG = |\mathbf{S}|,$$

sendo $|\cdot|$ o determinante da matriz $\mathbf{S}$, o qual se reduz à variância amostral para o caso de uma única variável ($p = 1$). Contudo, esta medida não é muito informativa, pois não consegue caracterizar estruturas de covariâncias diferentes. Por exemplo, duas matrizes completamente diferentes podem ter o mesmo valor VG. Entretanto, em qualquer análise estatística o resultado $|\mathbf{S}| = 0$ indica a existência de variáveis redundantes entre as p -variáveis. A questão de quais variáveis redundantes devem ser removidas no caso de degenerescência não é fácil de responder. Quando há possibilidade de escolha, o pesquisador deve reter as medidas de uma variável presumidamente causal ao invés de uma com característica secundária. No entanto, quando isso não acontece, pode-se utilizar a técnica de Análise de Componentes Principais para resolver esta situação.

Geometricamente a VG é proporcional ao quadrado do volume da elipsóide gerado pelos p vetores de desvios $\mathbf{d}_j$, $j = 1, 2, \dots, p$ (JOHNSON; WICHERN, 2002).

f) **Variância Amostral Total:** é uma medida capaz de sintetizar a informação sobre a variância das variáveis respostas. A *variância amostral total* (VT) é definida pela soma dos elementos da diagonal da matriz de variâncias e covariâncias $\mathbf{S}$:

$$\mathbf{VT} = tr(\mathbf{S}) = s_{11} + s_{22} + \dots + s_{pp} = \sum_{j=1}^p s_{jj} ,$$

onde $tr(\cdot)$ indica o traço da matriz. Essa medida é utilizada na Análise de Componentes Principais.

Geometricamente a VT representa a soma dos quadrados dos comprimentos dos vetores residuais $\mathbf{d}_j$, $j = 1, 2, \dots, p$. Ela não considera as orientações dos vetores residuais, sendo portanto limitada para ser utilizada com variáveis padronizadas, pois seu valor será sempre o mesmo para conjuntos de dados distintos desde que o número de variáveis destes seja igual (JOHNSON; WICHERN, 2002).

Todas essas estatísticas são muito sensíveis a observações discrepantes (*outliers*). Ferreira (1996) lembra que, as estatísticas $s_{jj'}$ e $r_{jj'}$ não refletem todo o conhecimento da associação entre as duas variáveis, Y_j e $Y_{j'}$, pois podem existir associações não lineares entre elas.

O desenvolvimento das técnicas multivariadas apresentados na literatura, pressupõe o conhecimento do vetor de médias populacional $\boldsymbol{\mu}$ e da matriz de variâncias e covariâncias populacional $\boldsymbol{\Sigma}$. Como na prática não se tem esse conhecimento, sugere-se substituir essas quantidades por seus estimadores usuais $\bar{\mathbf{y}}$ e $\mathbf{S}$, respectivamente. Assim, os resultados passarão a ser aproximados e terão um melhor desempenho para grandes amostras (BARROSO; ARTES, 2003).

As técnicas multivariadas só devem ser aplicadas com a presença significativa de covariância (ou correlação) entre as variáveis respostas. Caso contrário, essas técnicas não apresentarão grandes vantagens sobre as técnicas univariadas. Segundo Hair Jr. et al. (2006), o teste mais amplamente usado para esse fim é o teste de esfericidade de Bartlett (ver Anexo F). Ele examina as correlações entre todas as variáveis dependentes e avalia se existe, coletivamente intercorrelação significativa.

2.2.7 Modelo Linear Multivariado

O modelo linear multivariado é, na essência, uma generalização do caso univariado. Considere o problema de modelar o relacionamento entre p variáveis respostas $Y_1, Y_2, \dots, Y_p$ e um

único conjunto de variáveis preditoras $X_1, X_2, \dots, X_r$. Em que:

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}x_{11} + \dots + \beta_{r1}x_{1r} + \varepsilon_1 \\ &\vdots \\ Y_p &= \beta_{0p} + \beta_{1p}x_{n1} + \dots + \beta_{rp}x_{nr} + \varepsilon_p \end{aligned}$$

Desta maneira, segundo Johnson e Wichern (2002), e Cuadras (2006), o *modelo linear multivariado* é expresso matricialmente por:

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{E},$$

onde

$\mathbf{Y}$ é a matriz de variáveis respostas ($n \times p$) - contém n observações multivariadas sobre p variáveis dependentes;

$\mathbf{X}$ é a matriz de delineamento ($n \times (r + 1)$), de zeros e uns;

$\boldsymbol{\beta}$ é matriz de parâmetros desconhecidos ($n \times (r + 1)$); e

$\mathbf{E}$ é a matriz de erros aleatórios ($n \times p$); cada linha de $\mathbf{E}$ é um vetor normal p -variado com vetor de média $\mathbf{0}$ e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$ positiva definida (Anexo B), isto é, $E(\varepsilon_i) = 0$ e $Cov(\varepsilon_j, \varepsilon_{j'}) = s_{jj'}$, para $j, j' = 1, 2, \dots, p$.

Sendo:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = [\mathbf{y}_1 \vdots \dots \vdots \mathbf{y}_p],$$

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1r} \\ x_{20} & x_{21} & \cdots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nr} \end{bmatrix} = [\mathbf{x}_0 \vdots \mathbf{x}_1 \vdots \dots \vdots \mathbf{x}_r],$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rp} \end{bmatrix} = [\boldsymbol{\beta}_1 \vdots \dots \vdots \boldsymbol{\beta}_p], \text{ e}$$

$$\mathbf{E} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{np} \end{bmatrix} = [\boldsymbol{\varepsilon}_1 \vdots \dots \vdots \boldsymbol{\varepsilon}_p].$$

A matriz $\boldsymbol{\beta}$ de parâmetros pode ser estimada pelo método usual de mínimos quadrados, da mesma forma como é feito para o caso univariado.

No software R, para ajustar um modelo linear multivariado utiliza-se o comando *manova*.

2.3 As Técnicas Multivariadas

As técnicas multivariadas podem ser classificadas de diversas maneiras. De acordo com Mingoti (2005), a estatística multivariada pode ser dividida em:

- a) Técnicas Exploratórias: promovem a simplificação da estrutura de variabilidade dos dados. Esses métodos têm um apelo prático muito interessante, pois em sua grande maioria independem do conhecimento da forma matemática da distribuição de probabilidades geradora dos dados amostrais. Quando esta distribuição é conhecida pode-se fazer inferências sobre os resultados obtidos. Algumas das técnicas exploratórias são: Análise de Componentes Principais (ACP); Análise Fatorial (AF); Análise de Correlação Canônica (ACC); Análise de Agrupamentos (AA); Análise Discriminante (AD) e Análise de Correspondência (AC).
- b) Técnicas de Inferência Estatística: permitem que conclusões sejam tiradas acerca da população usando informações de uma amostra multivariada. Dentre elas estão a Análise de Variância Multivariada (MANOVA) e a Análise de Regressão Multivariada.

Já para Hair Jr. et al. (2006), é possível dividir as técnicas multivariadas em:

- a) Técnicas de Dependência: são aquelas em que uma variável ou conjunto de variáveis é identificado como a variável dependente⁶ a ser predita ou explicada por outras variáveis conhecidas como variáveis independentes⁷. Por exemplo: MANOVA e AD.

⁶Variável dependente: efeito presumido, ou resposta, a uma mudança na(s) variável(eis) independente(s).

⁷Variável independente: causa presumida de qualquer mudança na variável dependente.

- b) Técnicas de Interdependência: são aquelas em que nenhuma variável ou grupo de variáveis é definida(o) como independente ou dependente. Neste caso, o procedimento envolve a análise simultânea de todas as variáveis no conjunto. Exemplos: ACP e AF.

Resumindo, independente da divisão adotada, existem vários métodos de análise multivariada com finalidades bem diversas entre si. Portanto, é preciso saber que conhecimento se pretende gerar, ou melhor, o que se pretende afirmar a respeito dos dados. Reis (1997) relata que o truque na estatística multivariada, se existe, não está nos cálculos, fácil e rapidamente feitos num computador com software adequado. O truque consiste em escolher o método apropriado ao tipo de dados, usá-lo corretamente, saber interpretar os resultados e retirar deles as conclusões corretas.

Na agropecuária ainda existem poucos trabalhos de pesquisa que se utilizam das técnicas multivariadas. Nesses, as técnicas mais utilizadas são: a MANOVA, a ACP, a AD e a AA.

2.3.1 Análise de Componentes Principais (*Principal Components Analysis*)

Esta técnica teve origem em 1901, com Karl Pearson⁸ sendo consolidada somente em 1933 com Hottelling⁹. Segundo Morrison (1976), a Análise de Componentes Principais (ACP) surge da necessidade de se conhecer as estruturas de dependência das variáveis e a *priori* não é encontrado nenhum padrão de causalidade. Mingoti (2005) afirma que, seu objetivo principal é o de explicar a estrutura de variâncias e covariâncias de um vetor aleatório composto de p -variáveis aleatórias iniciais, podendo-se resumir sua informação.

A ACP requer que os dados das p variáveis avaliadas sejam métricos. A técnica consiste basicamente em transformar um conjunto original de variáveis $(Y_1, Y_2, \dots, Y_p)$ em outro conjunto de dimensão equivalente $(C_1, C_2, \dots, C_p)$, tal que:

$$C_j = e_{1j}Y_1 + e_{2j}Y_2 + \dots + e_{pj}Y_p$$

onde $e_{jj'}$ são os coeficientes calculados pela técnica, $j, j' = 1, 2, \dots, p$. Ela pode ser considerada uma técnica exata, pois em sua composição não se tem a presença do erro, sendo sua estrutura basicamente matemática. O novo conjunto de variáveis possui propriedades importantes e de grande interesse.

⁸PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, Philadelphia. Series 6, n. 2, p. 559-572, 1901.

⁹HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, Washington, v. 24, p. 417-441, 498-520, 1933.

Esta técnica busca imprimir um tratamento estatístico a um número relativamente alto de variáveis heterogêneas, que possuam um grau considerável de aspectos comuns, isto é, com um elevado grau de correlação entre si. Desta forma, o que se busca é condensar o conjunto inicial de muitas variáveis ($Y_j, j = 1, 2, \dots, p$) em um número bem menor de novas variáveis ($C_k, k = 1, 2, \dots, q$, sendo $q < p$) chamadas *componentes principais* e conseguir uma pequena perda de informações.

De acordo com Cruz (1990), as componentes principais (CP's) são independentes entre si e são estimadas com o propósito de reter, em ordem de estimação, o máximo da informação, em termos de variação total contida nos dados iniciais. Cruz e Regazzi (1997, apud BARBOSA et al., 2006) acrescentam que é por este motivo que é possível avaliar a importância de cada característica estudada sobre a variação total, possibilitando o descarte das variáveis redundantes (menos discriminantes), por estarem correlacionadas a outras variáveis, pela sua invariância ou por serem uma combinação linear de outras características.

Em muitas situações, os pesquisadores geram um considerável acréscimo de trabalho ao avaliarem um grande número de características (variáveis respostas). Como consequência têm um aumento no trabalho de caracterização sem melhoria na precisão, tornando a análise mais trabalhosa e dificultando a interpretação dos dados. Nestes casos, pode-se minimizar o problema fazendo-se uso da técnica de ACP, eliminando aquelas variáveis respostas que menos contribuem ao estudo (LIBERATO; VALE; CRUZ, 1999; e BARBOSA et al., 2006).

Uma vez determinadas as CP's, os seus valores numéricos, denominados de *escores*, podem ser calculados para cada elemento amostral. Deste modo, os valores de cada componente podem ser analisados, utilizando-se técnicas estatísticas usuais como análise de variância e análise de regressão, dentre outras.

Daher, Moraes e Cruz (1997) utilizam a ACP em estudos sobre divergência genética entre amostras genéticas de capim-elefante e observaram que, de um total de 22 caracteres avaliados em três anos, apenas oito deles (36,4%) foram selecionados como os mais importantes para a determinação da divergência genética.

Barbosa et al. (2005a) trabalham com a ACP em um estudo envolvendo 367 suínos, onde avaliaram 33 características de carcaça. Das 33 variáveis obtidas, 17 (51,5%) foram passíveis de descarte. Em outro trabalho, Barbosa et al. (2005b) selecionam variáveis de desempenho de 435 suínos pela ACP. De 11 características de desempenho, seis foram passíveis de descarte, porque contribuíam pouco para a variação total, isto é, foram consideradas redundantes.

Barbosa et al. (2006) também aplicam a ACP para avaliar características de qualidade da carne suína que, em geral, podem ser geneticamente melhoradas pela seleção. Foram analisadas 10 características de 326 animais e as três primeiras CP's explicaram 60, 65% da variação total dos dados. Pelo critério de Jolliffe, as componentes com autovalores (Anexo I) menores que 0,7 podem ser descartadas, e com base nos resultados obtidos, pode-se afirmar que 40% das variáveis analisadas foram consideradas redundantes, podendo ser descartadas em experimentos futuros.

Resultado semelhante é encontrado por Destefanis, Barge e Brugiapaglia (2000) que trabalharam com esta mesma técnica em análises químicas, físicas e sensoriais da carne de bovinos jovens e verificaram que as três primeiras CP's explicaram aproximadamente 63% da variação total. Os autores concluíram que a técnica de ACP é um procedimento muito efetivo para resumir o julgamento da qualidade da carne.

2.3.2 Análise de Correspondência (*Correspondence Analysis*)

A técnica de Análise Correspondência (AC) pode ser considerada um caso especial da técnica de ACP, porém dirigida a dados categóricos organizados em tabelas de contingência e não a dados contínuos.

Segundo Jobson¹⁰ (1996, apud MINGOTI, 2005), testes como qui-quadrado, por exemplo, são utilizados para avaliar se a informação contida nas linhas da tabela são independentes ou não da informação contida nas colunas. Uma outra abordagem para se avaliar a relação das linhas e colunas é a AC, que é um método de estatística multivariada. Mingoti (2005) lembra que a AC tem uma relação direta com a estatística qui-quadrado e é uma técnica muito útil para a análise de dados quantitativos.

Segundo Mingoti (2005), os primeiros artigos que introduzem noções de AC datam de 1933. No período de 1933 a 1960 são publicados vários trabalhos, notadamente em ecologia e psicologia. Porém, uma melhor formalização da metodologia é proposta na França, por Benzécri (1960), para o estudo de tabelas de contingência em lingüística. A partir de 1975 a técnica vem sendo utilizada em diversas áreas do conhecimento.

Esta técnica é especialmente indicada para descrever matrizes numéricas com grande volume de dados discretos e sem uma estrutura claramente definida *a priori*. Czermainski (2004) diz que, que a AC é uma técnica de análise exploratória de dados adequada para analisar tabelas

¹⁰JOBSON, J.D. **Applied multivariate data analysis**. v. I e II. New York: Springer Verlag, 731p., 1996.

de duas ou de múltiplas entradas, levando em conta algumas medidas de correspondência entre linhas e colunas. Basicamente, a técnica converte uma matriz de dados não negativos em um tipo particular de representação gráfica, em que as linhas e colunas da matriz são simultaneamente representadas em dimensão reduzida.

A representação gráfica obtida através da AC possibilita visualizar a distribuição das variáveis na sua relação com todas as outras, ou seja, permite a visualização das relações mais importantes de um grande conjunto de variáveis. Cada categoria de cada variável é representada por um ponto, e as distâncias entre os pontos representam as relações entre as variáveis que se deseja analisar. Por conseguinte, é essencial a compreensão do modelo utilizado para a determinação destas distâncias (Greenacre (1981), Lebart et al. (1977) e Lebart et al. (1984), apud CARVALHO e STRUCHINER, 1992).

Contudo, Souza et al. (2002) ressaltam uma desvantagem da AC nestes casos: como variáveis quantitativas podem ser transformadas em qualitativas, de acordo com a categorização dessas variáveis, pode-se ter perda de informações. E acrescentam ainda que somente é possível apresentar gráficos de associações se as variáveis testadas tiverem três ou mais categorias.

A AC é empregada por Souza et al. (2002) para avaliar associações e similaridades de variáveis categóricas avaliadas em 87 rebanhos bovinos leiteiros classificados em grupos de: ordenha manual (26), ordenha mecânica balde ao pé (28) e ordenha mecânica canalizada (33). A representação gráfica da AC mostrou ser uma forma alternativa e objetiva de evidenciar tendências de associações existentes entre variáveis categóricas, mesmo quando não é possível identificar essas associações por meio de testes tradicionais como o teste do qui-quadrado. Além de fornecer informações sobre associações entre variáveis, foi possível também identificar grupos de propriedades agropecuárias que possuem características comuns ou similaridades. A aplicação da AC na medicina veterinária preventiva e epidemiologia fornece, por meio de representação gráfica, evidências de quais fatores de risco estão associados com a prevalência ou incidência de enfermidades, bem como identifica grupos ou populações que possuem os mesmos fatores de risco.

Na área agropecuária esta técnica é pouco utilizada.

2.3.3 Análise de Correlação Canônica (*Canonical Correlation Analysis*)

A Análise de Correlação Canônica (ACC) é inicialmente proposta por Hotelling (1935¹¹, 1936¹²) e tem como objetivo principal a identificação e quantificação das relações lineares existentes entre dois conjuntos de múltiplas variáveis $\{Y_1, Y_2, \dots, Y_p\}$ e $\{X_1, X_2, \dots, X_q\}$, sendo elas métricas ou não-métricas (JOHNSON; WICHERN, 2002).

Vessoni (1998) diz que esta técnica pode ser muito útil em problemas que possuam mais de uma variável métrica dependente. O uso da ACC pode simplificar o problema e determinar quais variáveis são mais importantes na análise. Desta forma, pode-se realizar a análise em duas etapas, primeiro determinando os fatores relevantes, e posteriormente realizando regressões simples entre os mesmos.

Esta técnica pode ser de grande utilidade no estudo de dependências multivariadas, mas é uma das técnicas de análise multivariada menos explorada pelos usuários mais comuns, em geral, por falta de conhecimento adequado sobre o assunto e de sua potencialidade (MINGOTI, 2005).

Hardoon, Szedmak e Shawe-Taylor (2003) comparam a ACC com o problema de encontrar bases vetoriais para dois conjuntos de variáveis, tal que a correlação entre as projeções das variáveis nestas bases vetoriais sejam mutuamente maximizadas.

Para uma melhor compreensão da técnica, considerem $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_p]$ e $\mathbf{Y} = [Y_1 \ Y_2 \ \dots \ Y_q]$ dois conjuntos de variáveis avaliadas em n indivíduos. A técnica procura estabelecer a forma e a dimensão do relacionamento entre esses conjuntos. Assim, obtém-se a matriz de covariância S particionada:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$$

sendo $\mathbf{S}_{11}$ ($p \times p$) e $\mathbf{S}_{22}$ ($q \times q$) as matrizes de variâncias e covariâncias amostrais entre as variáveis do conjunto $\mathbf{X}$ e do conjunto $\mathbf{Y}$, respectivamente. As covariâncias entre as variáveis de diferentes conjuntos, uma variável de $\mathbf{X}$ e outra de $\mathbf{Y}$, estarão contidas na matriz $\mathbf{S}_{12}$ ($p \times q$), ou na $\mathbf{S}_{21}$ ($q \times p$). Analisar essas covariâncias pode ser extremamente trabalhoso, ainda mais se p e q forem grandes. Por conta deste problema, Vessoni (1998) afirma que o principal objetivo da correlação canônica é resumir as associações entre $\mathbf{X}$ e $\mathbf{Y}$ em função de algumas poucas correlações escolhidas, ao invés das $p \times q$ correlações.

¹¹HOTELLING, H. The most predictable criterion. **Journal of Educational Psychology**, Washington, v.26, p.139-142, 1935.

¹²HOTELLING, H. Relations between two sets of variates. **Biometrika**, Cambridge, v.28, p.321-377, 1936.

A idéia básica é resumir a informação de cada conjunto de variáveis respostas em combinações lineares (pares canônicos) dessas variáveis, sendo que a escolha dos coeficientes dessas combinações é feita tendo-se como critério a maximização da correlação entre os conjuntos de variáveis respostas. Na ACC não existe a distinção entre variável independente e dependente, existem somente dois conjuntos de variáveis e se busca a máxima correlação entre ambos.

A ACC pode acomodar qualquer variável métrica sem a suposição estrita de normalidade. No sentido estrito, permite também que dados não-métricos transformados (na forma de variáveis dicotômicas) também sejam usados, se a forma da distribuição não diminuir a correlação com outras variáveis. No entanto, a normalidade multivariada é exigida para o teste da significância de cada função canônica. Deste modo, apesar da normalidade não ser estritamente exigida, é altamente recomendada que todas as variáveis sejam avaliadas quanto a normalidade e transformadas se necessário (HAIR JR. et al., 2006).

Hair Jr. et al. (2006) relatam que a ACC apresenta o menor número de restrições sobre os tipos de dados nos quais ela opera. Como as outras técnicas impõem restrições mais rígidas, em geral, crê-se que a informação obtida a partir delas seja de melhor qualidade e pode ser apresentada de uma maneira melhor para a interpretação. Por essa razão, segundo estes autores, muitos pesquisadores consideram a ACC como uma última alternativa, a ser usada quando todas as outras técnicas mais exigentes foram descartadas. Mas em situações com múltiplas variáveis dependentes e independentes, a ACC é a técnica multivariada mais adequada e poderosa. Os autores ainda completam que a ACC se limita a identificar relações lineares.

A ACC pode ser vista como uma extensão da regressão linear múltipla (VESSONI, 1998; MINGOTI, 2005, dentre outros) e pode-se demonstrar que a Análise de Variância Univariada (ANOVA) e a Análise Discriminante (AD) são casos particulares da ACC (JOHNSON; WICHERN, 2002).

As técnicas de ACP e ACC são utilizadas por Vainionpää et al. (2000) em um estudo envolvendo cinco diferentes cultivares de batata. Foram avaliadas 34 variáveis, que inicialmente foram divididas em seis grupos para facilitar a utilização da ACP, e desta forma selecionar os grupos de variáveis que seriam usadas na ACC. As duas primeiras CP's foram responsáveis por 83,5% do total da variância dos dados, e nove variáveis puderam ser descartadas. Para a ACC utilizaram as 25 variáveis restantes, que se dividiram em dois grupos: 12 variáveis preditoras e 13 variáveis dependentes. Os quatro primeiros pares canônicos foram significativos (as correlações canônicas obtidas foram: 0,88; 0,85; 0,81 e 0,68, respectivamente), indicando dependência entre

os dois conjuntos considerados. Porém, para facilitar a interpretação, somente os três primeiros pares canônicos foram considerados. A ACC foi sensível o suficiente para detectar as mudanças nos fatores da cadeia de produção e os resultados deram uma visão da diversidade de variáveis relacionadas à qualidade da batata.

Trugilho, Lima e Mori (2003) usam a ACC para analisar sete clones de *Eucalyptus grandis* e três de *Eucalyptus saligna*. No total foram avaliadas 13 variáveis respostas, separadas em três grupos: I) características químicas da madeira; II) características físicas da madeira; e III) características dimensionais das fibras da madeira. Avaliou-se a Correlação Canônica (CC) existente entre o grupo I com o grupo II e entre o grupo I com o grupo III. Observou-se que as CC's foram elevadas e que nos dois casos o primeiro e segundo pares canônicos (0,77 e 0,65 para o primeiro caso; e 0,94 e 0,88 para o segundo caso) foram significativos. A ACC permitiu observar que os grupos de variáveis considerados não são independentes.

A associação desfavorável entre características geralmente causa atraso no progresso genético e nesse contexto a ACC pode ser aplicada possibilitando o estudo da associação entre dois conjuntos de variáveis. BARBOSA et al. (2005c) estudam a associação entre características de desempenho e de qualidade da carcaça de 844 suínos. Observaram que os dois grupos de características não são independentes, pois os dois primeiros pares canônicos foram estatisticamente significativos neste estudo.

2.3.4 Análise de Variância Multivariada (*Multivariate Analysis of Variance*)

A Análise de Variância Multivariada (MANOVA) é introduzida pela formulação original de Wilks¹³ (1932). No entanto, ela se torna uma ferramenta prática para pesquisadores somente depois do desenvolvimento de estatísticas de teste apropriadas com distribuições tabeladas e da ampla disponibilidade de programas de computador para processar essas estatísticas (HAIR JR. et al., 2006).

Segundo Reis (1997), a MANOVA é uma extensão da análise de variância simples (ANOVA) e a principal diferença entre as duas reside no fato da ANOVA avaliar as diferenças entre as médias de grupos apenas para uma variável resposta, enquanto que na MANOVA se procede à comparação entre as médias de grupos para diversas variáveis respostas simultaneamente.

Hair Jr. et al. (2006) relatam que a MANOVA é uma técnica de dependência que

¹³WILKS, S.S. Certain Generalizations in the Analysis of Variance. **Biometrika**, Cambridge, v. 24, p. 471-494, 1932.

mede as diferenças para duas ou mais variáveis dependentes métricas, com base em um conjunto de variáveis categóricas (não métricas) que atuam como variáveis independentes.

Demétrio (1985) acrescenta que, de maneira geral, as informações fornecidas pela modelagem univariada são contempladas pela extensão multivariada, sem, no entanto, levar em consideração um nível de significância conjunto dos testes e o aproveitamento das correlações existentes entre as variáveis. Isto é, uma MANOVA substitui p ANOVA's considerando a correlação existente entre as p variáveis respostas observadas e o nível de significância conjunto.

Rao¹⁴ (1952, apud DEMÉTRIO, 1985) afirma que a MANOVA consiste em analisar as variâncias e as covariâncias de variáveis correlacionadas por meio da comparação de matrizes de estimativas de variâncias e covariâncias.

Laforge (1981) assemelha a MANOVA ao estudo da dispersão dos centróides de aglomeração ou das nuvens de pontos de um espaço multidimensional (o centro de gravidade ou centróide da nuvem de pontos tem por coordenadas o vetor de médias dos indivíduos da amostra em questão).

Ao discutir a hipótese de homogeneidade de matrizes de variâncias e covariâncias, o autor comenta também que a MANOVA resiste bem a uma pequena heterogeneidade de variâncias e covariâncias. Mardia (1971, apud DEMÉTRIO, 1985) dando ênfase ao estudo do efeito da não-normalidade dos dados, no caso de experimentos com um fator, chega a conclusão que a MANOVA é robusta à não-normalidade, enquanto que os testes de igualdade de matrizes de covariâncias não o são. Ito e Schull (1964, apud DEMÉTRIO, 1985) e Korin (1972, apud DEMÉTRIO, 1985), estudam o efeito da não-normalidade, todos mostram que os resultados dos testes não são grandemente afetados por heterogeneidade das matrizes de variâncias e covariâncias, quando o número de repetições é grande e o mesmo para todos os tratamentos.

Uma diferença básica em relação ao caso univariado, diz respeito aos testes de hipóteses e intervalos de confiança da MANOVA, porque não existe um procedimento único ou melhor para a sua realização. Essas dificuldades aliadas à maior complexidade dos cálculos e da interpretação dos resultados, comparando-se ao caso univariado, torna a MANOVA uma ferramenta útil, mas que deve ser utilizada com cautela.

São várias as razões para se preferir um teste multivariado a vários testes univariados. Segundo Reis (1997), dentre as mais importantes, incluem-se as seguintes:

¹⁴RAO, C.R. **Advanced statistical methods in biometric research**. New York: John Wiley & Sons, 1952. 390 p.

- i) A utilização de testes estatísticos separados para cada variável provoca um erro tipo I global demasiadamente elevado, isto é, a probabilidade de rejeitar a hipótese nula quando ela é falsa toma valores que ultrapassam o aceitável. Por exemplo, considere que foram realizados 10 testes univariados com distribuição *t-Student* e em cada um utilizou-se um nível de significância de 0,05. Se admitirmos que os testes são independentes (porque, de fato, os testes não o são) a probabilidade de se rejeitar pelo menos uma hipótese nula é:
- $$1 - P[\text{não rejeitar nenhuma das hipóteses nulas}] = 1 - \underbrace{[0,95 \times \dots \times 0,95]}_{10 \text{ vezes}}$$
- $$= 1 - 0,598 \approx 0,401, \text{ que é um valor inaceitável.}$$
- ii) Embora as diferenças entre grupos possam não ser significativas para cada variável separadamente, quando analisadas em conjunto poderão surgir diferenças significativas. Pequenas diferenças para cada variável poderão combinar-se para produzir uma diferença global significativa. Nesses casos, os testes multivariados tornam-se mais poderosos, ou seja, é mais elevada a probabilidade de rejeitar a hipótese nula quando ela é falsa.
- iii) Os testes univariados ignoram muitas informações importantes contida nos dados, como as correlações entre as variáveis. Com o uso de testes multivariados, estas informações são incorporadas na análise através da matriz de variâncias e covariâncias e desta maneira refinam-se os resultados.
- iv) As decisões tomadas a partir dos resultados de testes univariados e de um teste multivariado poderão ser muito diferentes e mesmo contraditórias. Alguns autores sugerem como o melhor processo, proceder-se a um teste multivariado e, em caso de diferenças significativas, realizar testes univariados para identificar as variáveis que mais contribuem para essa decisão.

Como última recomendação, Finney (1956, apud DEMÉTRIO, 1985) sugere que a importância da análise multivariada na interpretação de dados experimentais deve ser cuidadosamente examinada do ponto de vista prático.

Para utilização da informação da MANOVA, pode-se utilizar uma ampla gama de metodologias multivariadas para complementar os resultados obtidos e propiciar subsídios para utilização prática das informações resultantes. Dentre as várias possibilidades, destacam-se as variáveis canônicas (VC), os componentes principais (CP) e a análise de fatores.

Viana et al. (2001) utilizam a MANOVA para avaliar o desempenho de quatro linhagens de matrizes de frangos de corte: duas da Universidade Federal de Viçosa (UFV) e duas

outras provenientes de marcas comerciais existentes no mercado. Avaliou-se quatro características de importância econômica em três períodos da vida produtiva das aves, observou-se efeito significativo de genótipo, com a aplicação de testes de comparações múltiplas os resultados mostraram diferenças no desempenho entre as marcas comerciais e os genótipos da UFV e diferença entre os genótipos da UFV nos períodos inicial e médio. No período total, as linhagens da UFV apresentaram desempenho igual ao das marcas comerciais.

Em cruzamentos dialéticos (melhoramento de plantas) os melhoristas necessitam avaliar vários caracteres para melhor inferir sobre a superioridade de populações. Ledo, Ferreira e Ramalho (2003) escrevem as expressões para a MANOVA do modelo de cruzamentos dialéticos de Gardner e Eberhart (1966), com o intuito de fornecer meios mais eficientes para a seleção de genótipos superiores. A técnica se mostrou eficiente e pode ser utilizada para estimar a heterose em várias características, simultaneamente.

Em um banco de dados é muito comum aplicar sequencialmente mais que uma técnica multivariada para complementar os resultados. Por exemplo, com as informações resultantes da MANOVA pode-se utilizar outras metodologias multivariadas para complementar os resultados obtidos. Dentre as várias possibilidades, destacam-se as Variáveis Canônicas (VC), as Componentes Principais (CP) e a análise de fatores.

2.3.5 Análise de Agrupamentos (*Cluster Analysis*)

Encontrar nos dados uma estrutura natural de agrupamento é uma importante técnica exploratória, e é exatamente este o propósito da Análise de Agrupamentos (AA). Segundo Barroso e Artes (2003), a AA é o nome dado a um conjunto de técnicas utilizadas na identificação de padrões de comportamento em bancos de dados através da formação de grupos homogêneos de casos. A AA também é conhecida como *análise de conglomerados*, *classificação* ou *cluster analysis*.

Para Cruz e Regazzi (1994, apud ABREU et al., 2002), a AA tem por finalidade reunir, por algum critério de classificação pré-determinado, as unidades amostrais em vários grupos, de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre os grupos. Segundo Hair Jr. et al. (2006) a idéia é maximizar a homogeneidade de objetos dentro de grupos, ao mesmo tempo em que se maximiza a heterogeneidade entre os grupos.

A AA permite uma avaliação da similaridade (ou dissimilaridade) entre objetos levando em consideração várias características concomitantemente. Para Hair Jr. et al. (2006) é

uma técnica que reúne indivíduos ou objetos em grupos tais que os objetos no mesmo grupo são mais parecidos uns com os outros do que com os objetos de outros grupos.

Hair Jr. et al. (2006) afirmam que observações atípicas distorcem a verdadeira estrutura e tornam os agrupamentos obtidos não representativos da verdadeira estrutura da população. Por essa razão, uma verificação preliminar de observações atípicas é sempre necessária. Os autores sugerem como o modo mais fácil de conduzir tal projeção seja preparar um diagrama de perfil gráfico. O diagrama de perfil lista as variáveis ao longo do eixo horizontal e os valores das variáveis ao longo do eixo vertical. Cada ponto do gráfico representa o valor da variável correspondente, e os pontos são conectados para facilitar a interpretação visual. Desta maneira, cada linha representa um objeto.

O processo de agrupamento envolve duas etapas: a primeira relaciona-se com a estimação de uma medida de similaridade (ou dissimilaridade) entre as unidades amostrais; e a segunda, com a adoção de uma técnica de agrupamento para a formação dos grupos (Cruz e Regazzi (1994), apud ABREU et al., 2002).

Segundo Freitas¹⁵ (1996, apud ABREU et al., 2002), dado um conjunto de n indivíduos e p variáveis, a estimação das medidas de similaridade (ou dissimilaridade) consiste na conversão da matriz de observações $n \times p$, em uma matriz quadrada e simétrica de ordem n , de similaridades (ou dissimilaridades) individuais, que são medidas da distância entre pares de indivíduos. Na posição (i, j) dessa matriz encontra-se a distância entre o i -ésimo e o j -ésimo indivíduo.

Barroso e Artes (2003) relatam que as distâncias são as medidas de dissimilaridade mais utilizadas no estudo de bancos de dados com variáveis quantitativas. De acordo como Lima (1969) e Barroso & Artes (2003), uma medida d_{ij} é definida como uma *distância* entre os pontos i e j se: a) $d_{ij} \geq 0$, para qualquer i e j ; b) $d_{ii} = 0$; c) $d_{ij} = d_{ji}$; d) $d_{ij} \leq d_{im} + d_{mj}$. As distâncias mais comuns são apresentadas no Apêndice A.

Em muitas situações a pesquisa envolve a análise de variáveis qualitativas. Nesses casos, há duas formas de tratamento: ou são transformadas em variáveis quantitativas e usa-se uma medida de dissimilaridade, ou então se trabalha com coeficientes de similaridade especialmente desenvolvidos para variáveis qualitativas (ver JOHNSON e WICHERN (2002), BARROSO e ARTES (2003), e REIS (1997), para mais detalhes).

Existe ainda a possibilidade determinar uma medida de parecença que envolva os

¹⁵FREITAS, R.T.F. **Estudo da divergência genética de suínos em cruzamentos, utilizando-se técnica de análise multivariada**. 1996. 152p. Tese (Doutorado em Genética e Melhoramento) - Universidade Federal de Viçosa, Viçosa, 1996.

dois tipo de variáveis: qualitativas e quantitativas. Esta medida pode ser obtida ponderando-se as partes referentes às variáveis quantitativas e as qualitativas (mais detalhes em BARROSO; ARTES, 2003).

Em relação às técnicas de agrupamento, elas admitem abordagens hierárquicas e abordagens não-hierárquicas (ou de partição). As primeiras consideram que de início cada indivíduo se encontra isolado e o processo segue aproximando estes indivíduos de acordo com suas similaridades até que se atinja uma estabilidade relativa, que variará em função dos objetivos do trabalho. As técnicas de partição operam em sentido contrário, particionando um agrupamento único inicial.

Apesar da utilidade da AA, Hair Jr. et al. (2006) destacam algumas advertências sobre esta técnica:

- a) A AA é usada como uma técnica exploratória ou descritiva, pois não possui base estatística sobre a qual possa-se realizar inferências.
- b) O agrupamento obtido pela técnica não é único, já que a pertinência a um agrupamento para qualquer número de soluções depende de muitos elementos do procedimento, e muitas soluções diferentes podem ser obtidas pela variação de um ou mais elementos;
- c) A AA sempre criará grupos, independentemente da verdadeira existência de qualquer estrutura nos dados;
- d) A solução de agrupamentos é totalmente dependente das variáveis usadas como base para a medida de similaridade. O acréscimo ou a eliminação de variáveis relevantes pode ter um impacto substancial sobre a solução resultante.

Assim, o pesquisador deve tomar muito cuidado ao avaliar o impacto de cada decisão envolvida na execução de uma análise de agrupamento.

A AA relaciona-se com outras técnicas multivariadas já conhecidas. É comum, por exemplo, quando se trabalha um grande número de variáveis, tentar reduzir a dimensão do conjunto de variáveis através da análise fatorial, canônica ou de componentes principais. Daí, os escores dos primeiros fatores, variáveis ou componentes são usados na AA. Além disso, para verificar a adequação da partição obtida com essa análise, quando já são conhecidos os grupos e os seus componentes, é comum a utilização da análise discriminante.

Trabalhando primeiramente com a MANOVA, Moura et al. (1999) estudam 15 variáveis respostas de 10 linhagens de pimentão, quanto à eficiência nutricional em relação ao

fósforo (P). Os tratamentos foram distribuídos em arranjo fatorial 10 x 5, constituídos de 10 linhagens, cinco doses de P e quatro repetições, em delineamento de blocos casualizados. Para cada dose de P aplicada ao solo foi realizada uma análise multivariada. Constatou-se a variabilidade genética entre as linhagens estudadas em todas as doses de P. Em seguida, a técnica de AA, baseada na distância generalizada de Mahalanobis (cujo princípio básico é manter a homogeneidade dentro e heterogeneidade entre os grupos formados) foi aplicada, possibilitando encontrar a dose de P mais adequada para estudos genéticos (250 mg de P/kg de solo). Por meio desta técnica, o nível de P mais adequado para estudos genéticos futuros é aquele que proporciona melhor discriminação entre as linhagens, ou seja, que apresenta maior número de grupos distintos de linhagens. Identificou-se também a produção de matéria seca da parte aérea como a característica que mais contribuiu para a divergência genética com esta dosagem.

Abreu et al. (2002) utilizam a AA para estudar a diversidade genética de híbridos resultantes do cruzamento entre linhagens de matrizes de frango de corte. O delineamento experimental foi em blocos casualizados, consistindo de três blocos, 82 cruzamentos e nove fêmeas por cruzamento. Foram medidas 11 características de produção. A AA foi realizada, adotando-se a distância euclidiana média como medida de dissimilaridade, sobre a qual empregou-se o método de agrupamento de otimização de Tocher. Formaram-se 29 grupos de cruzamentos, sendo que os cruzamentos foram distribuídos uniformemente dentro dos grupos, não existindo grupos com grandes concentrações de cruzamentos.

Rodrigues et al. (2002) caracterizam a variabilidade genética de parte do germoplasma existente em poder de produtores de feijão no Rio Grande do Sul (37 cultivares locais) e de cultivares produzidas pela pesquisa no Estado (14 cultivares), reunindo-as em grupos de similaridade genética, utilizando 40 variáveis respostas (descritores morfológicos). Primeiramente empregaram a ACP e em seguida a AA, adotando a distância euclidiana ao quadrado como medida de dissimilaridade e o método de Ward. O uso destas técnicas possibilitou identificar descritores ineficientes ou redundantes no estudo da variabilidade genética e reuniu as cultivares estudadas em quatro grupos distintos de similaridade genética. As cultivares locais revelaram variabilidade superior à encontrada nas cultivares oriundas da pesquisa, o que evidenciou a importância da sua inclusão em programas de melhoramento.

Ferreira et al. (2003) avaliam a divergência genética entre clones de palma forrageira, que substitui o milho na ração do gado leiteiro. O delineamento deste experimento foi em blocos casualizados, com três repetições, sendo 20 características avaliadas em 19 clones (tratamentos).

Pela MANOVA verificou-se diferença entre os vetores de médias de clones. Com a aplicação ACP foi possível reduzir a dimensionalidade original para duas dimensões (CP's), com explicação de 85% da variação total, sendo considerada como característica passível de descarte a porcentagem de infestação por cochonilha. Nove grupos foram discriminados pela AA. As características de maior discriminação foram as espessuras dos artículos primário, secundário e terciário, número de artículo primário e pesos médios de matéria verde por artículos secundário e terciário.

A definição de um tipo de clima leva em consideração vários elementos climáticos conjuntamente. Dessa forma, a análise de quanto um clima é diferente de outro deve ser feita levando em consideração várias características que o diferem dos demais. Rolim et al. (2007) usam a AA para avaliar a aplicabilidade das classificações em estudos agroclimáticos para o Estado de São Paulo, pela capacidade de separação dos climas por dois sistemas de classificações climáticas (Köppen e Thornthwaite). Os Sistemas de Classificações Climáticas são pouco utilizados no âmbito de estudos agrícolas pois, freqüentemente, considera-se sua escala de atuação muito abrangente. Utilizando cinco elementos climáticos, as separações por estes sistemas de classificação foram realizadas para cada elemento climático. Em seguida, a AA foi aplicada adotando como medida de dissimilaridade a distância euclidiana com método Ward, em função de todos os cinco elementos climáticos mensais conjuntamente. Esta análise permitiu verificar a eficiência na separação dos climas pelos dois sistemas de classificação climática.

2.3.6 Análise Fatorial (*Factor Analysis*)

A Análise Fatorial (AF) é uma das técnicas multivariadas mais conhecidas. Muitos dos seus exemplos iniciais aparecem nas áreas da psicologia e ciências sociais, na tentativa de identificar os fatores relacionados com a inteligência humana e ligá-los, de algum modo, à etnia. Esta técnica surgiu de estudos desenvolvidos por Charles Spearman¹⁶ (1904), que interpretou o fator “g” como um índice geral de inteligência. Apesar da sua criação datar do início do século XX, somente em 1940, com Lawley¹⁷, surge o primeiro trabalho com um maior rigor matemático, o que fez aumentar a sua aceitação (BARROSO; ARTES, 2003).

Johnson e Wichern (2002) completam que, basicamente, o modelo fatorial é moti-

¹⁶SPEARMAN, C. General Intelligence objectively determined and measured. **American Journal of Psychology**, Chicago, v. 15, p. 201-293, 1904.

¹⁷LAWLEY, D.N. The estimation of factor loadings by the method of maximum likelihood. **Proceedings of the Royal Statistical Society** of Edenburg, section A, 60, p. 64-82, 1940.

vado pelo seguinte argumento: suponha que as variáveis podem ser agrupadas por suas correlações, isto é, todas as variáveis dentro de um particular grupo sejam altamente correlacionadas entre si, mas tenham correlações relativamente pequenas com variáveis em grupos diferentes. Então, é admissível que cada grupo de variáveis represente um único fator, que é responsável pelas correlações observadas. É este tipo de estrutura que a análise fatorial pretende confirmar.

Segundo Fernandes e Lima (1991), a AF tem como princípio básico a redução do número original de variáveis respostas a um conjunto menor de “fatores” (ou variáveis latentes) independentes e não observados, que explicam de forma simples e reduzida, as variáveis originais. Na composição desses fatores, tem-se que:

- a) as variáveis mais correlacionadas se combinam dentro do mesmo fator, promovendo uma redução do número inicial de variáveis em um número menor de fatores;
- b) as variáveis que compõem um determinado fator são praticamente independentes das que constituem outro fator;
- c) a derivação dos fatores se processa visando maximizar a percentagem de variância total relativa a cada fator consecutivo;
- d) os fatores não são correlacionados entre si.

Uma das utilidades da AF, segundo Barroso e Artes (2003), é a identificação dos *constructos* existentes em um conjunto de dados. Constructo é a mensuração de um conceito abstrato, que muitas vezes não pode ser medido através da observação de uma única variável.

As variáveis respostas para a aplicação da AF geralmente são métricas. Porém, em alguns casos as variáveis dicotômicas¹⁸ (codificadas 0-1) podem ser empregadas. Se todas as variáveis são dicotômicas, então formas especializadas de AF, como a análise fatorial booleana, são mais adequadas (HAIR JR. et al., 2006).

De acordo com Toledo e Nicolella (2002), a AF demanda três etapas:

- i) Obtenção da matriz de correlação entre as variáveis;
- ii) Extração dos fatores comuns. Há na literatura vários métodos, os mais populares são o método das componentes principais e o método da máxima verossimilhança;

¹⁸Variável Dicotômica: variável não-métrica transformada em uma variável métrica designando-se 1 ou 0 a um objeto, dependendo se este possui ou não uma característica particular.

iii) Rotação dos eixos relativos aos fatores comuns. É comum obter uma matriz de difícil interpretação, na qual não é possível identificar quais variáveis são mais importantes para cada fator. Diante deste problema, processa-se uma rotação da matriz de cargas fatoriais, associando de maneira mais nítida um número de variáveis a cada fator, simplificando a solução e facilitando a interpretação.

Para uma situação com p variáveis, o modelo de AF ortogonal pode ser expresso da seguinte forma:

$$(Y_j - \bar{y}_j) = a_{j1}F_1 + \dots + a_{jm}F_m + e_j, \quad j = 1, \dots, p,$$

onde: Y_j são variáveis respostas originais; $\bar{y}_j$ são as médias das variáveis; F_i são os *fatores comuns* e explicam as correlações entre as variáveis, sendo $i = 1, 2, \dots, m$; a_{ji} são as *cargas fatoriais* (ou *factor loadings*), que refletem a importância do fator i na explicação da variável j ; e e_j é erro aleatório, que capta a *variação específica* (ou fator específico) da variável Y_j não explicada pela combinação linear das cargas fatoriais com os fatores comuns.

Diferentemente da ACP, na formação das p componentes, o erro aleatório está presente no modelo de AF, o que torna a técnica não-exata. Além disso, as suposições que são exigidas pela AF, acaba transformando a técnica em uma modelagem estatística. Contudo, Hair Jr. et al. (2006) dizem que, a normalidade multivariada na AF é necessária somente se um teste estatístico é aplicado para significância dos fatores, mas esses testes raramente são usados.

Johnson e Wichern (2002) afirmam que a AF pode ser entendida como uma extensão da ACP. Ambas podem ser vistas como tentativas de aproximar a matriz de variâncias e covariâncias, embora a aproximação baseada no modelo de AF seja bem mais elaborada e nem sempre possa ser aplicada. Uma questão primária na AF é se os dados são consistentes com uma estrutura prescrita.

A AF e a ACP têm como principal objetivo uma redução da dimensionalidade do espaço das variáveis. Por causa desta similaridade entre as duas técnicas, muitos autores consideram a ACP como uma forma alternativa da AF. Segundo Regazzi (2002), este tipo de consideração pode causar confusão, pois existem diferenças entre as duas técnicas:

- a) Na ACP a ênfase é explicar a variância total, em contraste com a AF que visa explicar as covariâncias entre as variáveis respostas;
- b) Na AF as variáveis originais são expressas como combinações lineares dos fatores, enquanto que os CP's são funções lineares das variáveis originais;

- c) Essencialmente, a ACP não requer pressuposições, enquanto que a AF requer várias pressuposições importantes, que são mais conceituais do que estatísticas, como: as variáveis devem apresentar correlações significativas para aplicação da técnica; cada fator comum deve ter variância igual a *um*, etc.;
- d) As CP's são únicas (se assumirmos que a matriz de variâncias e covariâncias possui autovalores distintos - Anexo I), enquanto que os fatores são passíveis de rotações. A possibilidade de se fazer uma rotação visando uma melhor interpretação dos fatores é uma das vantagens da AF sobre a ACP;
- e) Se o número de fatores for alterado, os fatores (estimados) também se alteram. Isto não ocorre na ACP.

Regazzi (2002) comenta também que, se o objetivo é encontrar e descrever alguns fatores de interesse, a AF pode ser mais útil se o modelo de fatores se ajusta bem aos dados e se a interpretação dos fatores rotacionados é de agrado. Por outro lado, se o objetivo for definir um menor número de variáveis a serem utilizadas em uma outra análise, iríamos ordinariamente preferir a ACP, embora em alguns casos este objetivo também possa ser alcançado com a AF.

Cooper (1983) compara a AF com a análise de regressão, onde cada variável resposta pode ser vista como uma variável dependente que é regredida sobre um conjunto de variáveis independentes não observáveis, os fatores comuns. Conseqüentemente, pode-se considerar também cada fator como uma variável dependente que é regredida sobre todas as variáveis originais, e neste caso, os coeficientes de regressão correspondem às cargas fatoriais que identificam a natureza dos fatores desconhecidos.

Vale lembrar que existem dois tipos diferentes de AF: a exploratória e a confirmatória. A primeira busca encontrar os fatores subjacentes (não observados) às variáveis originais amostradas. Na AF confirmatória, o usuário tem em mãos um modelo fatorial pré-especificado (modelo hipotético) e deseja verificar se é aplicável ou consistente com os dados amostrais de que dispõe (MINGOTI, 2005).

Fernandes e Lima (1991) usam a AF e a AA para identificar e caracterizar sistemas de produção de leite na Região da Zona da Mata e Campo das Vertentes de Minas Gerais. Foram identificados dois sistemas com estrutura de produção e características tecnológicas bastante diferenciadas. As técnicas de análise mostraram-se adequadas, levando em conta um conjunto de 16 variáveis respostas e não uma única variável de estratificação. Esta identificação dos sistemas de produção efetivamente usados pelos produtores é importante para subsidiar as instituições de

pesquisa agropecuária e de extensão rural na geração e transferência de tecnologias compatíveis com a realidade dos produtores.

Em um trabalho semelhante, Aleixo, Souza e Ferraudó (2007) estudam os diferentes grupos de produtores de leite da Cooperativa Nacional Agroindustrial (COONAI) para serem avaliados dentro de suas características e, posteriormente, verificar sua viabilidade e seus entraves econômicos. A COONAI é uma grande cooperativa de pequenos produtores e necessita de elaboração de estratégias de desenvolvimento para as classes produtivas diferenciadamente. Foram considerados 72 produtores, selecionados conforme 27 variáveis relacionadas a fatores produtivos. Aplicaram o método da AF onde analisaram os três primeiros fatores, com 52,76% da variância explicada acumulada. Em seguida, empregaram o método de AA, encontrando quatro grupos. Pela caracterização destes grupos, juntamente com a análise dos fatores obtidos, é possível viabilizar intervenções técnicas diferenciadas, o que permite a consolidação de condições de sustentabilidade a partir das reais necessidades de incorporação tecnológica dos produtores, resultando no êxito da atividade.

Já Carrer (2000) utiliza a AF juntamente com a AA para caracterizar e diferenciar a pecuária regional de corte no Brasil. Esta análise permitiu a caracterização de quatro diferentes sistemas produtivos e possibilitou concluir que esta atividade sofreu, desde sua gênese histórica, um processo de diferenciação que estabelece neste fim de século, uma situação de grandes diferenças inter-regionais com relação ao estágio de desenvolvimento desta atividade no país.

Toledo e Nicolella (2002) aplicam a AF para avaliar a qualidade da água em uma microbacia de Guaíra/SP, sob diferentes usos, agrícola e urbano. Foram medidas 10 variáveis respostas. Os três primeiros fatores explicaram 71% da variância dos dados. O primeiro fator explicou 47% da variância dos dados e foi utilizado na construção do índice de qualidade de água (IQA). O oxigênio dissolvido, fósforo total, amônia e condutividade elétrica foram as variáveis que mais contribuíram na determinação do IQA. Os resultados obtidos indicaram uma diferença entre os valores de IQA para as três estações. O uso da técnica de AF permitiu também avaliar as condições de deterioração da qualidade de água para a região de Guaíra, com vistas à obtenção de indicadores específicos, os quais poderão ser monitorados ao longo do tempo.

Pinto et al. (2005a) realizam uma avaliação morfológica de potros e potras da raça Mangalarga Marchador, utilizando a AF. Neste experimento efetuou-se 19 mensurações lineares e 11 medidas angulares. A AF foi empregada para estudar a existência de relações importantes entre as medidas morfométricas. Nas medidas lineares foram identificados de seis a

sete fatores, ao passo que, nas medidas angulares, o número de fatores foi seis. Foram observadas correlações importantes entre as medidas lineares, as quais possibilitaram a identificação de fatores denominados sustentação, altura, rendimento, equilíbrio, dentre outros, enquanto, nas medidas angulares, foram identificados fatores que podem estar associados à comodidade, agilidade, rendimento, impulsão e força para o andamento.

2.3.7 Análise Discriminante (*Discriminant Analysis*)

Um problema comum em muitas áreas de pesquisa ocorre quando existem dois ou mais grupos de objetos, para os quais um grande número de características foi medido, e deseja-se classificar novos objetos baseado no mesmo conjunto de características. Para a solução deste tipo de problema, Ferguson (1994, apud SANTOS, MATO e CLENELL, 2003) recomenda a técnica de Análise Discriminate (AD).

O problema da discriminação entre dois ou mais grupos visando uma posterior classificação foi inicialmente abordado por Fisher¹⁹ em 1936, com a obtenção de uma combinação linear das características observadas que apresentava o maior poder de discriminação entre os grupos.

Segundo Züge e Chaves Neto (1999), trata-se de uma técnica que tem o objetivo de confirmar ou verificar uma classificação feita *a priori*. Haddad (1989, apud CHINELATTO NETO; CASTRO; LIMA, 2005) completa que a técnica de AD testa a significância dessa classificação prévia e determina quais variáveis têm o poder de distinguir em que grupo devem estar as unidades do estudo. Já para Hair Jr. et al. (2006), a AD é a técnica apropriada para testar a hipótese de que as médias de um conjunto de variáveis independentes para dois ou mais grupos são iguais.

O objetivo da AD para Reis (1997) e Mingoti (2005) é construir uma regra de classificação, ou seja, encontrar uma combinação linear das variáveis independentes, que minimize a probabilidade de classificação incorreta dos indivíduos (erro de classificação). Além disso, é importante construir uma regra que minimize o custo de classificação incorreta.

Para construir esta regra, as variáveis discriminantes devem ser identificadas. É chamada de *Variável Discriminante* aquela variável resposta (ou o conjunto de variáveis respostas) que possui o maior poder de discriminação entre os grupos. Depois de encontrada, é possível estimar um conjunto de funções, chamadas *funções discriminantes*, que permitirão a classificação de novos casos, cujo agrupamento seja inicialmente desconhecido (REIS, 1997). As funções discriminantes

¹⁹FISHER, R.A. The use of multiple measurement in taxonomic problems. **Annals of Eugenics**, Londres, v. 7, p. 179-188, 1936.

são ortogonais e permitem que as observações sejam classificadas otimizadaamente.

A função discriminante permite conhecer o valor do escore discriminante, dos dados analisados. Este escore corresponde ao valor encontrado após a aplicação da função discriminante. O ponto de corte é o determinante para classificar uma nova observação, sendo obtido pela média das médias dos escores discriminantes de cada grupo. Desta forma, Reis (1997) afirma que, a AD pode ser compreendida como um sistema de pontuações que, a cada indivíduo, faz corresponder uma pontuação resultante de uma média ponderada dos valores que, para ele, assumem as variáveis independentes. Uma vez determinada essa pontuação, ela pode ser transformada numa probabilidade *a posteriori* desse indivíduo pertencer a cada um dos grupos.

As hipóteses básicas do método de AD são:

- a) O grupo de variáveis discriminantes tem distribuição normal multivariada;
- b) As matrizes de covariâncias dos grupos são iguais; e
- c) Os grupos diferem quanto as médias.

Para Hair Jr. et al. (2006) o propósito básico da AD é estimar a relação entre uma variável dependente não-métrica (categórica) e um conjunto de variáveis independentes métricas (variáveis discriminantes). A estimação é conseguida estabelecendo-se os pesos da variável estatística²⁰ para cada variável, maximizando a variância entre grupos relativa à variância dentro dos grupos.

Quando as variáveis discriminantes são variáveis mistas (quantitativas, nominais e ordinais), os objetivos da AD são atingidos mais comumente utilizando a regressão logística, as árvores de classificação ou as redes neurais artificiais (BARROSO; ARTES, 2003).

Everitt (2005) acrescenta que pode-se também estar interessado em classificar observações em diversos grupos. Isto é, para dois grupos, uma única dimensão (função discriminante) é necessária para que se possa classificar um indivíduo nestes grupos. Entretanto, quando há mais de dois grupos, a função terá mais de uma solução, refletindo o fato que mais de um sentido é preciso para descrever as diferenças entre os grupos. Assim, com g grupos e p variáveis haverá $d = \min(p, g - 1)$ soluções, ou seja, d funções discriminantes. Estas melhores dimensões de separação são conhecidas também como *Variáveis Canônicas* (VC).

Em outras palavras, as variáveis canônicas, ou funções discriminantes, podem ser vistas como delimitantes dos grupos no espaço p -dimensional.

²⁰Combinação linear que representa a soma ponderada de duas ou mais variáveis independentes (variáveis discriminantes) que formam a função discriminante.

A primeira função discriminante, ou variável canônica, é também conhecida como Função Discriminante Linear de Fisher (FDLF). Ela produz o maior valor possível para o teste F usado na comparação das médias; tem a propriedade de minimizar as probabilidades de má classificação, quando as populações são normalmente distribuídas com parâmetros μ_k e Σ_k conhecidos, $k = 1, 2, \dots, g$; e também capta a maior quantidade de informação contida nas variáveis (REGAZZI, 2002). Analogamente, a segunda função discriminante irá captar a maior informação contida nas variáveis, não captada pela primeira função discriminante, e assim por diante, até que seja esgotada toda informação.

A AD possui algumas semelhanças e diferenças com algumas outras técnicas:

- i) Ela difere dos métodos de AA, pois segundo Mingoti (2005), para a sua aplicação, é necessário que os grupos nos quais cada novo elemento amostral pode ser classificado, sejam conhecidos *a priori*, considerando-se suas características gerais.
- ii) Possui algumas semelhanças com a Regressão Múltipla (RM), porém tem filosofias e objetivos muito diferentes. A análise de RM é pressuposto que a variável dependente tem distribuição normal enquanto as variáveis independentes são determinísticas. A regressão utiliza um modelo matemático que, com base em certos pressupostos, permite encontrar estimativas dos parâmetros satisfazendo certas propriedades estatísticas desejáveis. Já na AD, o pressuposto reverte-se: pressupõe-se que as variáveis independentes tem distribuição conjunta normal multivariada enquanto a dependente é fixa e de tipo nominal. A AD utiliza uma estratégia para encontrar processos apurados de classificação dos indivíduos (REIS, 1997).
- iii) A MANOVA e AD são “imagens espelhadas”: as variáveis dependentes métricas em MANOVA são as variáveis independentes em AD, e a variável dependente não-métrica da AD se torna a(s) variável(eis) independente(s) em MANOVA. Além disso, ambas usam os mesmos métodos na formação de variáveis estatísticas (combinações lineares das variáveis) e na avaliação da significância estatística entre grupos. As diferenças, entretanto, se concentram em torno dos objetivos das análises e do papel da(s) variável(eis) não-métrica (HAIR JR. et al., 2006).
- iv) A AD também é muito semelhante a Regressão Logística²¹ em relação ao objetivo de identificar o grupo ao qual um objeto pertence. Porém esta última é limitada em sua forma básica, a dois grupos, apesar de formulações alternativas possibilitarem lidar com mais de dois grupos (detalhes em HAIR JR. et al., 2006).

²¹Forma especial de regressão na qual a variável dependente é não-métrica (dicotômica, binária).

Para testar a validade (acurácia) das funções discriminantes é possível classificar o conjunto de casos originais e comparar os agrupamentos obtidos com os grupos pré-definidos e assim estimar a percentagem de casos corretamente classificados a partir das variáveis utilizadas (REIS, 1997).

Daoyu e Lawes (2000) utilizam a MANOVA e a técnica de AD para selecionar pais de uma nova geração de kiwi. O experimento foi realizado no esquema fatorial 3x2, ou seja, analisaram-se seis populações da fruta kiwi em dois anos distintos, onde foram medidas 24 características fenotípicas da fruta. A MANOVA indicou que as populações diferiram entre si. A AD mostrou as características mais importantes na distinção entre as populações. As duas primeiras funções discriminantes contabilizaram 84% da potência total discriminante e foram estas que asseguraram a eficiência da análise.

Fonseca et al. (2000) avaliam o desempenho e divergência genética entre três raças suínas. Seis características reprodutivas foram mensuradas. Inicialmente a MANOVA foi aplicada, e verificou-se que a média de pelo menos uma das raças consideradas difere de outra pelo teste de Roy, e desta maneira calculou-se a diferença mínima significativa (dms) para testar contrastes entre médias. Como análise complementar obteve-se a primeira VC, que explicou 64,67% da variação observada. Para explicar no mínimo 80% da variação observada, também considerou-se a segunda VC. As raças Landrace e Large White apresentaram melhor desempenho nas características reprodutivas e foram geneticamente mais semelhantes quando comparadas à raça Duroc.

Com o mesmo raciocínio, Torres Filho et al. (2005) estudam a divergência genética entre duas linhagens de suínos da raça Large White. Observaram 8 variáveis, sendo 3 de desempenho e 5 reprodutivas. As duas linhagens avaliadas apresentaram divergência genética tanto para características de desempenho como reprodutivas, sugerindo que bons resultados de heterose e de complementariedade podem ser obtidos na exploração de matrizes F1 obtidas do cruzamento entre elas.

Rodrigues e Ando (2003) avaliam 65 variedades de arroz-de-sequeiro quanto à sensibilidade à radiação gama para classificá-los em dois grupos: Índica e Japônica. Essa classificação tem auxiliado na seleção dos progenitores e ampliação da base genética dos programas de melhoramento. As sementes foram submetidas a sete doses de radiação gama em um delineamento em blocos casualizados com três repetições. Inicialmente, todos os caracteres foram submetidos à ANOVA para detectar a existência de diferenças entre os grupos. Em seguida, realizou-se a MANOVA considerando três variáveis: percentagens de plântulas emergentes, sobreviventes e a

altura das plântulas, visando detectar diferenças entre os grupos de arroz em cada dose, incluindo o controle não irradiado e puderam constatar que os grupos apresentaram diferenças significativas. A visualização da discriminação dos grupos em espaço bidimensional foi feita pela construção de gráficos com as duas primeiras VC de cada dosagem. Os resultados mostraram que a sensibilidade à radiação gama foi eficiente para discriminar os grupos Índica e Japônica nas dosagens de 300 e 360 Gy; e as variedades de Japônica foram mais sensíveis à radiação gama do que as do grupo Índica.

Pinto et al. (2005b) utilizam a AD para estudar as diferenças morfológicas entre potros e potras da raça Mangalarga Marchador em diferentes idades, ou seja, em função do sexo e das idades dentro de cada sexo. Foram efetuadas 25 mensurações lineares e 11 medidas angulares no corpo dos animais. Na discriminação das idades foram utilizados animais ao nascimento, aos 2, 4, 6, 8, 10 e 12 meses de idade, enquanto que na discriminação dos sexos, foram utilizados animais recém nascidos, aos 6 e 12 meses de idade. Os machos foram, em geral, maiores que as fêmeas nas medidas lineares que variam em função do sexo, nas diferentes idades avaliadas, exceto a medida de perímetro torácico, que apresentou valor médio maior nas potras. As medidas lineares foram mais interessantes na identificação de animais sub ou superdesenvolvidos, pois permitem maiores percentuais de acerto na classificação dos animais nas faixas etárias avaliadas até os 12 meses de idade. Os erros de classificação quanto à idade foram observados, principalmente, após os quatro meses de idade podendo indicar os efeitos ambientais no período pós-desmame.

A quantidade e a qualidade do alimento fornecido às larvas em laboratório podem levar à formação de castas diferentes. Silva et al. (2005) avaliam a utilização das técnicas de AD e ACP, de modo a efetuar a discriminação das castas das abelhas adultas obtidas em laboratório em relação àquelas desenvolvidas naturalmente. Para determinar se as abelhas adultas obtidas em laboratório eram pertencentes à uma das três castas (operárias, rainhas e intercastas), foram comparadas com um controle constituído por operárias e rainhas da mesma origem das operárias desenvolvidas no laboratório. Utilizando oito caracteres morfológicos, discriminativos entre rainhas e operárias, aplicaram a AD. Todas as abelhas adultas obtidas em laboratório foram classificadas corretamente quanto a morfologia, como operárias. Em seguida, a ACP foi utilizada e as duas primeiras componentes explicaram cerca de 61,40% da variação total. Entre os caracteres avaliados, os que menos contribuíram para a determinação das castas foram o comprimento do mesoscuto e a largura da cabeça, sendo, portanto, dispensáveis em estudos futuros.

Em dados de pomares de maçãs, Maluche-Baretta, Amarante e Klauberg Filho

(2006) utilizam a AD e a ACC com o objetivo de identificar diferenças entre dois sistemas de produção, manejo convencional e orgânico, com base em atributos microbiológicos e químicos do solo. Oito variáveis discriminantes foram utilizadas e somente uma Função Discriminante (FD) foi necessária. A estatística de teste Lambda de Wilks indicou diferenças significativas entre as áreas estudadas. Segundo a AD, atributos relacionados ao carbono são mais sensíveis do que atributos relacionados ao nitrogênio na discriminação dos sistemas. Em seguida, a ACC foi realizada entre os atributos microbiológicos (composto de 4 variáveis) e químicos (composto de 11 variáveis) avaliados. Os dois primeiros pares canônicos foram significativos (sendo as correlações canônicas dadas por 0,90 e 0,75, respectivamente), contudo, somente o primeiro par foi considerado, podendo verificar a alta correlação canônica entre atributos biológicos e químicos do solo nas áreas estudadas.

Um trabalho interessante que mostra a aplicação sequencial das técnicas de análise multivariada é o de Chinellatto Neto, Castro e Lima (2005), em que se desejava identificar e caracterizar os produtores de leite de Minas Gerais. Inicialmente empregaram a AF em um conjunto de 12 variáveis, identificando três fatores que são capazes de diferenciar bem os produtores, explicando 76,1% da variância total dos dados. De posse dos escores fatoriais para cada um dos 294 produtores de leite, com relação a cada fator, realizou-se a AA, visando a identificação de grupos homogêneos de produtores, resultando na classificação de cinco grupos, onde dois foram desconsiderados porque eram formados somente por uma propriedade. Depois de agrupados os produtores em grupos homogêneos, utilizaram a AD para testar a classificação dos grupos, considerando quatro variáveis discriminantes. Duas funções discriminantes foram utilizadas (explicando 84,5 e 15,5%, respectivamente), confirmando-se 94,5% da classificação previamente feita pela AA. Em seguida, a ACC foi aplicada, associando variáveis que representaram os gastos com a sanidade do rebanho com a área destinada à alimentação constituída por pastagens e silagem, ambos os conjuntos formados por 4 variáveis. Encontraram um coeficiente de correlação de 0,818 entre o primeiro par canônico. Os resultados mostraram que os cuidados com a sanidade do rebanho e a alimentação por meio de pastagens são dois critérios de diferenciação entre os grupos de produtores.

Após a realização desta revisão, pode-se perceber que as ferramentas constituídas pelas técnicas de análise multivariada de dados, apresentam adequado potencial de uso para a elucidação dos problemas que este trabalho se propõe a investigar.

3 METODOLOGIA

Neste trabalho serão apresentadas aplicações de algumas técnicas multivariadas, mais especificamente a AA, a MANOVA e a ACP, que auxiliarão na busca dos objetivos aqui impostos. Dois conjuntos de dados são utilizados, um proveniente de experimento realizado junto a Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo (FZEA-USP), em Pirassununga/SP, e o outro desenvolvido na Faculdade de Ciências Agrárias e Veterinária da Universidade Estadual Paulista, em Jabotibacal/SP, e serão descritos nas seções 3.2, 3.3 e 3.4. Preferiu-se apresentar as seções Materiais e Métodos, Resultados e Discussões de cada aplicação separadamente para facilitar a compreensão.

3.1 Suporte Computacional

As técnicas multivariadas foram implementadas utilizando o ambiente R (versão 2.5.1). O R é uma linguagem de programação (DALGAARD, 2002) que permite manipular dados, fazer cálculos e construir gráficos estatísticos (MURRELL, 2006 e R, 2007). Caracteriza-se como um sistema completamente planejado e coerente e não apenas um conjunto ampliado de ferramentas muito específicas e inflexíveis, como são freqüentemente em outros programas de análise de dados, como o SAS, MINITAB etc. O R é um veículo para o desenvolvimento de novos métodos interativos de análise de dados.

O *software* pode ser obtido pelo site do CRAN (R, 2007). No CRAN é possível baixar não só o pacote principal do R, mas também os pacotes opcionais chamados de contribuídos (em inglês *Contributed Packages*) e uma série de manuais. O R dispõe de um grupo de discussão na internet, *R-Stat*, onde muitas dúvidas do funcionamento do *software* podem ser sanadas, além de possibilitar a troca de experiências pelos participantes.

Com o uso deste software, soluciona-se o problema de indisponibilidade de pacotes estatísticos específicos para aplicação das técnicas multivariadas, pois além de gratuito é amigável, aberto e com inúmeros recursos disponíveis. Contudo, a bibliografia sobre a implementação das técnicas de análise multivariada no R ainda é pequena e os materiais disponíveis estão fragmentados em poucos livros e na web.

Nos Anexos K, L e M são listados os comandos básicos para implementar as técnicas multivariadas que foram utilizadas no presente trabalho.

3.2 Material e Métodos - Experimento I

Quando se realiza um experimento em blocos casualizados, supõe-se inicialmente que a área experimental ou o material utilizado (indivíduos) seja heterogêneo e se conheça a causa dessa heterogeneidade. É importante salientar que à medida que se aumenta o controle local diminui-se o número de graus de liberdade do resíduo de uma análise de variância (univariada ou multivariada). Como este número é um indicador de precisão da análise, só se deve realizar o controle local quando realmente for necessário, ou seja, controle local desnecessário somente irá causar perda de sensibilidade na análise de variância (BARBIN, 2003).

Na experimentação com animais é comum o uso do delineamento em blocos casualizados quando o material experimental disponível é muito heterogêneo. Neste caso o pesquisador precisa juntar os animais em blocos, que sejam internamente muito homogêneos, a fim de que sejam oferecidas as mesmas condições experimentais a todos os tratamentos.

Na montagem dos blocos os pesquisadores utilizam informações de características disponíveis que julgam diferenciar o material experimental. Contudo, pela dificuldade de se avaliar diversas características simultaneamente, os pesquisadores acabam utilizando uma ou outra característica que julgam mais importantes para decidir sobre a composição dos blocos. Com isso o processo acaba sendo realizado de forma mais intuitiva e sem utilizar adequadamente todas as informações que foram inicialmente eleitas como discriminadoras dos blocos.

Para auxiliar o pesquisador nessa tarefa de montar os blocos, propõe-se o uso da técnica multivariada de Análise de Agrupamento (*cluster analysis*), que utiliza as informações de todas as características simultaneamente. Esta técnica de análise exploratória tem o propósito de encontrar nos dados uma estrutura de agrupamento natural.

O objetivo deste estudo é utilizar dados de um experimento delineado em blocos casualizados e montar novos blocos com o auxílio da Análise de Agrupamentos (AA), de maneira que as variâncias internas dos blocos sejam as menores possíveis.

Os dados utilizados são provenientes de um experimento com suínos pertencentes ao trabalho de Fagundes (1999), desenvolvido no Setor de Suinocultura do Campus de Jabotibacal, da Universidade Estadual Paulista. No experimento utilizou-se um delineamento em blocos casualizados, com 4 tratamentos e 2 blocos de 20 animais, que eram colocados em baias separadas. Na construção dos blocos, o pesquisador utilizou as informações sobre idade inicial (em dias) e peso inicial (em kg) dos 40 animais. As estatísticas descritivas dos dados se encontram no Apêndice B.

Previamente avaliou-se a presença de valores discrepantes (*outliers*) multivariados

nos dados utilizando gráficos de perfis, *Q-Q plot*, gráfico de dispersão das últimas componentes principais e *boxplot's*. Este procedimento é fundamental porque, segundo Mingoti (2003), alguns métodos de classificação são sensíveis à presença de observações discrepantes, podendo formar agrupamentos indesejáveis.

Os métodos de agrupamento exigem que todas as variáveis tenham a mesma escala de medida e caso isso não ocorra, elas devem ser padronizadas, de tal forma tenham média igual a zero e variância igual a um (Anexo A). A finalidade deste procedimento é equalizar a importância estatística de todas as variáveis utilizadas (MOITA NETO, 2004).

A técnica foi aplicada como descrita em 3.2.1 e os novos agrupamentos obtidos com o uso da técnica foram comparados com o agrupamento montado pelo pesquisador. A comparação dos resultados baseou-se nas variâncias internas dos blocos, na variância total e no número de indivíduos em cada grupo.

3.2.1 A Análise de Agrupamentos (AA)

Esta técnica consiste de diversas técnicas e algoritmos que têm o objetivo separar (ou agrupar) objetos em grupos similares, tendo como base alguma medida de parença, tal que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características (MINGOTI, 2005 e POWER e CAMPBELL, 1992).

No R estão disponíveis funções para a realização da AA, contudo, recomenda-se o pacote *cluster*, por apresentar mais recursos (mais detalhes em MAECHLER et al., 2007).

As medidas de parença têm um papel central nos algoritmos de agrupamentos. Considere que para cada elemento amostral têm-se informações de p -variáveis armazenadas em um vetor, as medidas de parença (que são métricas matemáticas) são definidas para avaliar se dois pontos quaisquer estão próximos e se podem fazer parte de um mesmo grupo, ou não. As medidas de parença podem ser definidas para variáveis quantitativas e qualitativas (JOHNSON; WICHERN, 2002). Existem dois tipos de medidas de parença:

- i) **Medidas de Similaridade:** quanto *maior* o valor observado, mais *parecidos* são os indivíduos (ou objetos).
- ii) **Medidas de Dissimilaridade:** quanto *maior* o seu valor, mais *diferentes* são os indivíduos (ou objetos). Geralmente, os algoritmos utilizados na análise de agrupamento estão baseados

em medidas de dissimilaridade ou distâncias (ver Apêndice A).

Várias métricas diferentes existem e cada uma delas produz um determinado tipo de agrupamento. Para que se possa proceder ao agrupamento de elementos, é necessário que se decida *a priori* a medida de parença que será utilizada. E para isso, o tipo de variável (qualitativa ou quantitativa) também deve ser levando em consideração. No presente experimento, como as variáveis são quantitativas utilizar-se-á a distância euclidiana como medida de dissimilaridade, que é calculada por:

$$d_{ij} = \sqrt{\sum_{k=1}^p (y_{ik} - y_{jk})^2},$$

onde d_{ij} é a distância do elemento i ao j , com $i, j = 1, 2, \dots, n$; y_{ik} e y_{jk} são os valores observados da variável k , $k = 1, 2, \dots, p$, para os indivíduos i e j .

Escolhida a métrica, as distâncias d_{ij} são calculadas para todos os elementos e armazenadas numa matriz $\mathbf{D}_{(n \times n)}$, chamada de *matriz de distâncias*, que é simétrica e com zeros na diagonal principal.

3.2.1.1 Métodos de Agrupamento

Como no caso das medidas de parença, existem vários métodos de agrupamento e a escolha exige o conhecimento de suas propriedades e dos objetivos da pesquisa. Contudo, Barroso e Artes (2003) recomendam que, na medida do possível, deve-se utilizar mais de um método sobre um mesmo conjunto de dados e através da comparação dos grupos formados, pode-se adotar a solução que representar melhor a situação em estudo (a que oferecer melhor diferenciação entre as unidades analisadas).

A maioria dos algoritmos utilizados na formação dos agrupamentos pode ser classificada em duas grandes famílias de métodos: os Hierárquicos e os Não-Hierárquicos.

3.2.1.1.1 Métodos Hierárquicos

Nos métodos Hierárquicos os indivíduos são classificados em grupos em diferentes etapas, de modo ordenado (hierárquico), produzindo uma árvore de classificação chamada de *Dendrograma* (ou *Dendograma*, que é um gráfico bidimensional que combina a ocorrência da fusão com a estimativa de distância das unidades agrupadas (Figura 2). Estes agrupamentos podem

ser utilizados tanto para agrupar indivíduos como para agrupar variáveis. Quando o dendrograma construído é das variáveis, a similaridade entre duas variáveis aponta forte correlação entre elas. Os dendrogramas de indivíduos são mais comuns do que os de variáveis (MOITA NETO, 2004).

Figura 2 - Dendrograma de um exemplo com cinco indivíduos, onde se utilizou a distância euclidiana e o método hierárquico do vizinho mais distante

Os métodos hierárquicos podem ser subdivididos em dois grupos: métodos aglomerativos e métodos divisivos.

Os **Métodos Aglomerativos** (ou *linkage methods*) são os métodos mais comuns, nos quais classificam-se as unidades em grupos, em sucessivas fusões, baseando-se em uma medida de parença escolhida a priori (distância euclidiana, no presente caso), reduzindo a um único grupo ao final. Um algoritmo geral para os agrupamentos hierárquicos aglomerativos com n objetos (itens ou variáveis) é o seguinte:

- 1) Inicia-se com n grupos, cada um com um único elemento, e com uma matriz simétrica de ordem n de distâncias $\mathbf{D} = (d_{ij})$.
- 2) Busca-se na matriz $\mathbf{D}$ o par de grupos mais similar (menor distância)
- 3) Fundir os grupos u e v e nomeá-lo por (uv) . Recalcular e rearranjar as distâncias na matriz $\mathbf{D}$ utilizando a medida de parença e o método aglomerativo escolhido.
 - 3.a) Eliminar as linhas e colunas correspondentes aos grupos u e v ;
 - 3.b) Acrescentar uma nova linha e uma nova coluna com as distâncias entre o grupo (uv) e os demais grupos.
- 4) Repetir os passos 2 e 3 num total de $(n - 1)$ vezes, até que todos os objetos estejam agrupados em único grupo. Anotar a identidade dos grupos que vão sendo fundidos e as respectivas distâncias nas quais isto ocorre.

Facilmente observa-se no algoritmo a propriedade de hierarquia, isto é, em cada estágio do algoritmo, cada novo conglomerado formado é um agrupamento de conglomerados formados nos estágios anteriores, implicando que, uma vez unidos dois elementos num mesmo *cluster*, estes não poderão ser separados.

Segundo Barroso e Artes (2003), o que diferencia os métodos aglomerativos é a regra de redefinição da matriz de parença a cada união de pares de objetos.

Para apresentar os métodos hierárquicos aglomerativos, considere G_1 e G_2 dois grupos de objetos, com $g_1 \geq 1$ e $g_2 \geq 1$ objetos, respectivamente e $d[G_1, G_2]$, a distância entre eles.

a) Método do vizinho mais próximo (*ligação simples, mínima distância* ou *single linkage*): baseia-se na menor distância entre um objeto de G_1 e um objeto de G_2 , ou seja:

$$d[G_1, G_2] = \min_{i \in G_1, j \in G_2} (d_{ij})$$

b) Método do vizinho mais distante (*ligação completa, máxima distância* ou *complete linkage*): baseia-se na maior distância entre um objeto de G_1 e um objeto de G_2 , ou seja:

$$d[G_1, G_2] = \max_{i \in G_1, j \in G_2} (d_{ij})$$

c) Método das médias das distâncias (*ligação média* ou *average linkage*): baseia-se na média das distâncias entre todos os objetos de G_1 e os de G_2 .

$$d[G_1, G_2] = \sum_{i \in G_1} \sum_{j \in G_2} \frac{d_{ij}}{g_1 g_2}$$

d) Método do Centróide (ou *centroid method*): baseia-se na distância entre os centróides dos grupos G_1 e G_2 , que são definidos como a média das coordenadas de todos os objetos de um mesmo grupo.

e) Método de Ward: baseia-se na noção de que os grupos de observações multivariadas devem ser agrupados, aproximadamente, numa elipse. A alocação de um elemento a um grupo é feita maximizando a homogeneidade dentro dos grupos, ou minimizando o total das somas de quadrados dentro de grupos, também conhecida como *soma de quadrados de erros* (ESS), que é calculado por:

$$ESS = \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})' (\mathbf{y}_i - \bar{\mathbf{y}})$$

onde $\mathbf{y}_i$ é o vetor multivariado de medidas associado com o i -ésimo objeto e $\bar{\mathbf{y}}$ é a média de todos os itens. Mais detalhes em Ward²² (1963).

Os métodos de ligação simples, completa e média podem ser utilizados tanto para variáveis quantitativas, quanto qualitativas, ao contrário dos métodos de centróide e de Ward que são apropriados apenas para variáveis quantitativas, já que têm como base a comparação de vetores de médias (BARROSO; ARTES, 2003).

A escolha do número final de grupos g em que o conjunto de dados deve ser repartido é subjetiva. Na realidade, o propósito é encontrar o número g que esteja associado à “partição natural” dos elementos que estão sendo comparados e agrupados (MINGOTI, 2005), e para isso a opinião do pesquisador é essencial.

Existem algumas estatísticas (que não serão abordadas neste trabalho) que podem ser utilizadas para determinar o número de grupos no conjunto de dados. Esses valores são colocados em um gráfico semelhante ao *scree plot*, devendo-se procurar por um cotovelo. Porém, estas estatísticas são basicamente heurísticas (suas distribuições não são conhecidas), servindo apenas como um conjunto de métodos que podem auxiliar na solução deste problema. Além da análise do comportamento dessas estatísticas em função do número de grupos, deve-se avaliar também o dendrograma e os itens que compõem cada um dos grupos.

Os **Métodos Divisivos** partem de um único grupo com os n elementos e por divisões sucessivas vão sendo divididos em 2, 3, ..., etc. grupos, de tal modo que os indivíduos em um subgrupo estão longe dos indivíduos do outro. Esses indivíduos são novamente divididos em subgrupos e o processo continua até que cada indivíduo forme um grupo. Estes métodos não serão abordados neste trabalho.

3.2.1.1.2 Métodos Não-Hierárquicos

Os métodos não-hierárquicos de agrupamento foram desenvolvidos para agrupar objetos (ou indivíduos) ao invés de variáveis, em k grupos que podem ser definidos antecipadamente, ou determinados durante a execução do procedimento. Esses métodos exigem a pré-fixação de critérios que produzam medidas sobre a qualidade da partição produzida. Basicamente, esses métodos seguem os seguintes passos:

²²WARD, J.H. Jr. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, Alexandria, v. 58, p. 236-244, 1963.

- i) Selecionar k centróides de grupos ou sementes iniciais, onde k é o número de grupos desejados.
- ii) Designar cada observação ao grupo mais próximo;
- iii) Realocar cada observação a um dos k grupos de acordo com uma regra de parada pré-determinada;
- iv) Parar o processo se não existe mais nenhuma realocação de pontos ou se a realocação satisfizer o critério estipulado na regra de parada. Caso contrário, voltar para o passo ii).

Muitos dos algoritmos não-hierárquicos diferem com respeito ao método usado para obter os centróides ou os pontos-sementes ou ainda quanto à regra usada para realocar os itens (JOHNSON; WICHERN, 2002). O método mais popular é o *método das k-médias* ou *k-means*. MacQueen em 1967, sugeriu o termo *k-means* para descrever um algoritmo que designa cada objeto ao grupo que tem o centróide mais próximo. Esse método é mais sensível à presença de *outliers*, fazendo com que sua prévia identificação seja necessária. Na sua forma mais simplificada, o processo é composto de três passos:

- a) Agrupar os objetos em k grupos iniciais arbitrariamente;
- b) Percorrer a lista de objetos e calcular as distâncias de cada um deles ao centróide dos grupos. Fazer a re-alocação do objeto ao grupo em que ele apresentar distância mínima, obviamente se não for o grupo ao qual este já pertença. Recalcular os centróides dos grupos que ganharam ou perderam algum objeto.
- c) Repetir o passo b) até que nenhuma alteração seja feita.

Barroso e Artes (2003) alertam que a aplicação de métodos hierárquicos a grande massas de dados pode ser proibitiva, tanto em termos computacionais, como, muitas vezes, na análise dos resultados obtidos. Nessas circunstâncias o método das k-médias parece ser mais indicado.

3.2.1.2 Validação e Interpretação dos Resultados

Após o uso de diferentes métodos de agrupamento, deve-se certificar de que os grupos construídos realmente diferem entre si. Nesta etapa da análise podem ser empregados vários testes estatísticos univariados ou multivariados. Dentre os multivariados pode-se citar a MANOVA, usada

para verificar se há diferença estatisticamente significativa entre os vetores de médias de cada grupo, e a análise discriminante, que é usada quando já são conhecidos os grupos e os seus componentes.

Existe ainda uma medida de validação conhecida como *correlação cofenética*, que é mais utilizada para validar os agrupamentos hierárquicos. No software R a correlação cofenética pode ser encontrada no pacote *vegan*, pelo comando *cophenetic* (mais detalhes em OKSANEN et al., 2007).

Definição 3.1 : Define-se a **correlação cofenética** como sendo a correlação entre as distâncias previstas e as efetivamente observadas. Quanto mais próxima de *um* for a correlação cofenética, melhor será a qualidade do agrupamento.

Certificando-se de que o agrupamento escolhido é um bom agrupamento, deve-se interpretar os resultados e caracterizar os grupos, evidenciando as suas diferenças e semelhanças. Para isso, é necessário lançar mão de técnicas descritivas e eventualmente utilizar os resultados dos testes de validação como balisa da interpretação. Para facilitar a identificação dos grupos, pode-se utilizar representações gráficas multivariadas das médias observadas para as variáveis em cada grupo. Vale lembrar que a presença do pesquisador nesta etapa é fundamental, pois ele poderá caracterizar os grupos mais facilmente.

Quando o número de variáveis é muito grande, fica difícil interpretar uma tabela. Nesse contexto, a utilização de gráficos de representação de casos facilita a observação de semelhanças e diferenças entre os grupos. O gráfico de perfil e o gráfico radar são exemplos de gráficos de representação (mais detalhes em BARROSO e ARTES, 2003).

3.2.2 Resultados e Discussão - Experimento I

Após análise gráfica, nenhum *outlier* multivariado foi encontrado. Os gráficos *Q-Q plot*, dispersão das duas CP's e perfis podem ser avaliados nas Figuras 3 e 4. Porém, a padronização das variáveis foi necessária (Figura 5), por conta das diferentes escalas de medida das variáveis idade e peso. A normalidade dos dados não foi verificada, pois apenas se deseja obter novos agrupamentos.

A partir da matriz de distâncias, **D**, foram aplicados os seguintes métodos hierárquicos aglomerativos: vizinho mais próximo, vizinho mais distante, média das distâncias, centróide e Ward. Numa análise posterior foi também avaliado o método não hierárquico das k-médias.

Figura 3 - *Q-Q plot* das variáveis peso e idade para identificar outliers e gráfico de dispersão dos escores das duas CP's

Figura 4 - Perfis individuais dos pesos e idades iniciais dos 40 suínos

Avaliando os dendrogramas dos métodos hierárquicos apresentados nas Figuras 6 e 7, e considerando a presença de apenas dois grupos em cada figura, os métodos do vizinho mais próximo e centróide indicaram a presença de um grupo unitário (animal 28). Se um destes dois

Figura 5 - *Boxplot's* dos dados originais (a) e padronizados (b) das variáveis

métodos fosse escolhido, haveria a necessidade de considerar três grupos (blocos) e desconsiderar o grupo unitário para obter-se uma solução mais satisfatória, porque nenhum outro animal foi considerado similar ao animal 28. O dendrograma do método do vizinho mais distante foi muito semelhante ao do método das médias das distâncias, sendo poucos os indivíduos classificados em blocos diferentes; porém, o número de indivíduos incluídos em cada um dos blocos foi bem diferente. O método de Ward apresentou a solução mais apropriada ao problema, sugerindo a construção de dois blocos com aproximadamente o mesmo número de indivíduos.

Baseando-se nas correlações cofenéticas dos agrupamentos hierárquicos apresentadas na Tabela 1 pode-se avaliar que o método das médias das distâncias proporcionou o melhor agrupamento de animais, apresentando a maior correlação cofenética (0,7583). Este resultado concorda com Barroso e Artes (2003), que afirmam que o método das médias das distâncias produz melhores partições que os métodos de ligação simples e completa.

Tabela 1 - Correlações cofenéticas dos agrupamentos hierárquicos

Método Hierárquico	Correlação Cofenética
Centróide	0,749
Vizinho mais próximo	0,748
Vizinho mais distante	0,694
Média das distâncias	0,758
Ward	0,671

Figura 6 - Dendrogramas resultantes dos métodos do vizinho mais próximo (a), vizinho mais distante (b) e média das distâncias (c)

Figura 7 - Dendrogramas resultantes dos métodos Centróide (d) e Ward (e)

Uma análise comparativa final foi feita entre os métodos de médias das distâncias, de Ward, k-médias e a solução original utilizada pelo pesquisador. Na Tabela 2 que apresenta a discriminação dos indivíduos em cada bloco, para cada um dos métodos, percebe-se que as soluções propostas são bastante distintas: nenhum animal é designado ao mesmo bloco nas quatro soluções, exceto o animal 8.

Na Tabela 3 que apresenta o tamanho, a média e a variância interna dos blocos para cada uma das variáveis, pode-se perceber que as menores variâncias internas estão, na sua maioria, associadas a soluções resultantes dos métodos de análise de agrupamento.

A solução do método da média das distâncias apresenta menores variâncias internas

para ambas as variáveis que a solução original, confirmando o comentário de Barroso e Artes (2003). Observa-se ainda que as variâncias internas dos dois blocos são bem parecidas entre si, mas os blocos têm números de animais muito distintos.

O método de Ward teve um bom desempenho apresentando variâncias internas menores que as da solução original, com exceção da variância interna do bloco 2 para a variável idade, cujo valor é quase o dobro da variância interna do bloco original. Uma característica interessante dessa solução é que os blocos têm números de animais muito parecidos. Tal resultado reiterou o comentário de Barroso e Artes (2003), de que o método de Ward tende a produzir grupos com aproximadamente o mesmo número de elementos.

O método das k-médias apresentou variâncias internas dos blocos relativamente pequenas, com exceção da variável idade no bloco 1. Os blocos resultantes têm números de elementos muito distintos, o que não aconteceu com a solução do método de Ward, porém, o método das k-médias se apresentou melhor que o método da média das distâncias.

Avaliando as variâncias totais dos blocos, apresentadas na Tabela 4, percebem-se que as soluções resultantes da aplicação dos diferentes métodos de Análise de Agrupamentos foram melhores (variâncias totais menores) que a solução original. O método da média das distâncias apresentou o menor valor da variância total.

Tabela 2 - Distribuição dos animais nos blocos (1 ou 2) do trabalho original (M4) e das soluções obtidas pelos métodos das médias das distâncias (M1), Ward (M2) e k-médias (M3)

Animal	Método				Animal	Método			
	M1	M2	M3	M4		M1	M2	M3	M4
1	1	1	2	2	21	2	2	1	2
2	1	1	2	2	22	1	1	1	2
3	2	2	1	2	23	1	1	2	1
4	2	2	1	2	24	1	1	2	2
5	1	1	1	2	25	2	2	1	2
6	1	2	2	1	26	1	1	2	1
7	1	1	2	2	27	2	2	1	1
8	1	1	1	1	28	1	2	2	1
9	1	1	2	1	29	2	2	1	2
10	2	2	1	2	30	1	1	2	1
11	1	1	2	1	31	1	1	2	2
12	1	2	2	2	32	1	1	2	2
13	1	2	2	1	33	1	2	2	2
14	1	2	2	2	34	1	1	2	1
15	1	2	2	1	35	1	1	2	1
16	1	2	2	1	36	1	1	2	1
17	1	1	2	1	37	1	1	2	2
18	1	1	1	2	38	1	1	2	2
19	2	2	1	1	39	1	1	2	1
20	1	1	2	1	40	2	2	1	1

Tabela 3 - Números de animais (n), médias e variâncias internas (var) dos blocos (B) formados pelos métodos de análise de agrupamentos e do trabalho original

Variável	B	Método											
		Original			Média das dist.			Ward			K-médias		
		n	média	var	n	média	var	n	média	var	n	média	var
Peso	1	20	25,10	1,41	31	24,81	1,06	23	24,30	0,38	27	25,13	1,24
	2	20	24,95	0,87	9	25,78	0,63	17	26,00	0,44	13	24,81	0,86
Idade	1	20	73,15	2,34	31	72,87	1,38	23	73,13	1,30	27	73,63	3,47
	2	20	74,05	3,73	9	76,11	1,11	17	74,24	5,19	13	73,54	2,77

Para cada método também foi calculada a distância euclidiana entre os centróides de cada bloco (Tabela 4). Os métodos das médias das distâncias e Ward apresentaram as maiores distâncias entre os centróides, sugerindo que os blocos obtidos por essas técnicas, sejam mais heterogêneos entre si, em relação aos outros métodos avaliados.

Tabela 4 - Variância Total (VT) dos blocos, segundo o método e distância euclidiana dos centróides dos blocos 1 e 2

Bloco	Método			
	Original	Média das distâncias	Ward	K-médias
1	3,7552	1,7430	1,6808	3,6250
2	4,6000	2,4441	5,6293	4,7151
Total	8.3552	4.1871	7.3101	8.3401
Distância euclidiana	0,91	3,38	2,0239	0,3345

Como o interesse do pesquisador está em montar dois blocos de 20 animais, tais que a variância interna de cada bloco seja a menor possível, o método de Ward forneceu a solução mais interessante ao problema, porque apresentou blocos com números de indivíduos mais aproximados (Tabela 3).

Entretanto, uma solução para o problema de construir dois blocos de suínos com 20 indivíduos por bloco, que denominaremos de Ward*: baseia-se na re-alocação dos indivíduos excedentes do bloco 1 (B1) para o bloco 2 (B2), tendo como critério de escolha destes indivíduos o seguinte algoritmo:

- i) Calcula-se a distância euclidiana de todos os indivíduos de B1 ao centróide de B2;
- ii) O indivíduo de B1 “mais próximo” do centróide de B2 é transferido para B2;
- iii) Recalcula-se o centróide de B2 e retorne ao passo i), até que se tenha 20 animais em cada bloco.

Três indivíduos foram re-allocados do bloco 1 para o bloco 2. Os resultados do método Ward* (Tabela 5) mostraram que a variância interna dos blocos é ainda pequena, e percebe-se uma redução da variância interna do bloco 2, em relação a variável idade. A variância total deste procedimento (Tabela 6) foi inferior àquela obtida pelo método de Ward, que já era satisfatória. Porém, a distância euclidiana dos centróides foi menor do que aquela obtida do método de Ward, mas ainda maior do que a do trabalho original.

Tabela 5 - Números de animais, médias e variâncias internas dos blocos formados pelo método de Ward, Ward* e do trabalho original

Variável	Bloco	Método								
		Original			Ward			Ward*		
		<i>n</i>	média	variância	<i>n</i>	média	variância	<i>n</i>	média	variância
Peso	1	20	25,10	1,4105	23	24,30	0,3804	20	24,20	0,3526
	2	20	24,95	0,8658	17	26,00	0,4375	20	25,85	0,5026
Idade	1	20	73,15	2,3447	23	73,13	1,3004	20	73,15	1,5026
	2	20	74,05	3,7342	17	74,24	5,1918	20	74,06	4,5763

Tabela 6 - Variância Total (VT) dos blocos, segundo o método e distância euclidiana dos centróides dos blocos 1 e 2

Bloco	Método		
	Original	Ward	Ward*
1	3,7552	1,6808	1,8552
2	4,6000	5,6293	5,0789
Total	8,3552	7,3101	6,9341
Distância euclidiana	0,9124	2,0239	1,8794

A solução alternativa, proposta em Ward* mostrou-se eficiente na montagem de blocos homogêneos e com os mesmos números de indivíduos.

Logo, sugere-se ao pesquisador utilizar o agrupamento obtido pelo método das médias das distâncias, se o interesse estiver na variância interna, variância total e na distância entre os centróides; ou o método de Ward, se o interesse estiver em criar blocos de mesmo tamanho.

Sempre que o pesquisador tiver interesse em criar blocos de tamanhos distintos, sugere-se que o mesmo inicie o processo de agrupamento com um número maior de indivíduos e elimine os indivíduos em excesso de cada bloco, somente após a aplicação da técnica.

3.3 Material e Métodos - Experimento II

Os dados utilizados nesta seção, foram obtidos da dissertação de Caetano (2005). Trata-se de um estudo para avaliar o aparelho reprodutor do molusco *Achatina fulica*, *escargot* comestível, em idade potencial de reprodução.

Segundo Caetano (2005) e Pacheco (2004 apud CAETANO, 2005), este molusco foi introduzido no Brasil, primeiramente no Paraná no final da década de 80 e a segunda introdução ocorreu pelo porto de Santos entre os anos de 1996 a 1998. Distribuiu-se por todo litoral de São Paulo e interior, através da promoção de cursos rápidos para a formação de criadores com finalidades comerciais. As características climáticas do litoral, solo arenoso e a inexistência de predadores e de inúmeros abrigos em áreas urbanas, contribuiu para que ocorressem verdadeiras explosões populacionais destes animais. Entretanto, no estado o fenômeno acontece em menor quantidade, ocorrendo animais somente no período de verão úmido. Verifica-se a inexistência de animais na área rural, mesmo em áreas de diferentes culturas.

Devido ao desconhecimento da biologia desse animal e sua proliferação desordenada em nosso país, faz-se necessário desenvolver pesquisas e estudos acerca de sua biologia e principalmente de seu aparelho reprodutor, hábitos e comportamentos reprodutivos, para que se possa contribuir para o estabelecimento de programas de controle e erradicação da espécie em questão. Sabe-se apenas que os animais são hermafroditas e que o processo de reprodução destes é extremamente sensível ao efeito ambiental. Desta maneira, focou-se nas diferenças entre animais de cativeiro e asselvajados.

O delineamento adotado no estudo foi o inteiramente ao acaso, com quatro tratamentos (locais de coleta ou situação): o primeiro grupo originário de criação em cativeiro (LAB) no Helicário Experimental Professora Doutora Lor Cury - FMVZ/USP, e os três grupos restantes, de animais asselvajados provenientes da área urbana de Pirassununga/SP: quintal com horta (DPE), terreno baldio com pomar (MBA) e terreno baldio murado contendo entulho (CJA).

Foram abatidos, pelo método a frio, 30 animais por região de coleta, o que totalizou cerca de 120 animais. A avaliação só teve início após a morte dos animais por congelamento. Os moluscos foram submetidos a avaliações, externa e interna, e foram mensuradas 15 variáveis.

Os dados foram submetidos à MANOVA, com o intuito de comparar os quatro tratamentos simultaneamente para todas as variáveis e confrontar os seus resultados com os obtidos por Caetano (2005). As estatísticas descritivas dos dados se encontram no Apêndice C.

Para uso na MANOVA foram consideradas somente 8 variáveis: Y_1 : peso do animal (g); Y_2 : comprimento da concha (mm); Y_3 : largura da concha (mm); Y_4 : razão corporal ($= \frac{Y_2}{Y_3}$); Y_5 : peso da concha (g); Y_6 : peso da carne (g); Y_7 : peso glândula prostática (g) e Y_8 : peso da glândula albumina (g). Algumas variáveis foram excluídas da análise por diversos motivos: variável discreta (número de ovos); variável contínua com muitos zeros (peso da massa de ovos), variáveis que

dificultaram a verificação das pressuposições da análise (percentuais de pesos das frações do animal em relação ao peso corporal).

A glândula prostática funciona como o útero do animal. A glândula albumina/ovotestis, ou glândula de proteína, tem uma grande flutuação no tamanho e conteúdo durante os diferentes estágios do ciclo reprodutivo (TOMIYAMA, 1993 apud CAETANO, 2005), isto é, ela fica maior antes da postura de ovos e imediatamente após a postura, este órgão definha-se, tornando ao que pode ser seu tamanho original. O peso desta glândula serve como um bom índice de preparação para a oviposição.

Gráficos de dispersão das últimas componentes principais, *Q-Q plot's*, *boxplot's* uni e bivariados e a distância de Mahalanobis, como descrita em 2.2.5, foram utilizados na verificação de *ouliers*.

3.3.1 Análise de Variância Multivariada (MANOVA)

A análise de variância multivariada (MANOVA) é usada para investigar se os vetores de médias dos g grupos (ou tratamentos) são iguais, ou seja, faz a comparação entre as médias das diferentes variáveis simultaneamente. A MANOVA é uma extensão da Análise de Variância (ANOVA) e ambas utilizam dois passos seqüenciais:

- a) Testa-se a hipótese global de igualdade de médias entre os grupos;
- b) Se o resultado do passo anterior for significativo, utilizam-se testes adicionais no sentido de explicar as diferenças entre os grupos (comparações múltiplas).

A MANOVA, no entanto, tem vantagens sobre a realização de sucessivas ANOVAS para diferentes variáveis, pois na primeira considera-se o nível de significância conjunto dos testes e aproveita-se as informações conjuntas das variáveis envolvidas.

De acordo com Hair Jr. et al. (2006), para os procedimentos de teste multivariado de MANOVA serem válidos, quatro suposições devem ser atendidas, que podem ser entendidas como generalizações dos pressupostos da ANOVA:

- i) Modelo aditivo para efeitos de grupos, blocos (se houver) e erro;
- ii) As observações devem ser não correlacionadas;
- iii) Homocedasticidade - as matrizes de variâncias e covariâncias devem ser iguais para todos os grupos (Teste de Box - Anexo G), isto é, $\Sigma_1 = \Sigma_2 = \dots = \Sigma_g$. Contudo, segundo Hair Jr.

et al. (2006, p. 287), uma violação dessa suposição tem impacto mínimo se os grupos têm aproximadamente o mesmo tamanho (ou seja, se o tamanho do maior grupo dividido pelo tamanho do menor for menor do que 1,5).

- iv) O conjunto de p variáveis dependentes deve seguir uma distribuição normal multivariada, ou seja, os erros devem ter distribuição normal multivariada, com matriz de variâncias e covariâncias Σ ($e_i \sim N_p(\mathbf{0}, \Sigma)$, sendo $\mathbf{0}$ o vetor nulo). Essa condição tem relevância diminuída quando as amostras são de grande dimensão.

A análise da variância multivariada (MANOVA) pode ser utilizada para qualquer delineamento experimental, sem apresentar dificuldades adicionais. Mas é preciso trabalhar com todas as variáveis simultaneamente, obtendo para elas matrizes de somas de quadrados e somas de produtos cruzados (SQPC).

Considere, por exemplo, um delineamento inteiramente casualizado (DIC) com g grupos (tratamentos), onde avaliou-se p variáveis de n observações, tal que $n = n_1 + n_2, \dots, n_g$, sendo n_k o número de indivíduos do grupo k , $k = 1, 2, \dots, g$. Assim, o modelo linear de análise de variância multivariada para este caso é dado matricialmente por:

$$\mathbf{Y}_{(n \times p)} = \mathbf{X}_{(n \times (g+1))} \boldsymbol{\beta}_{((g+1) \times p)} + \mathbf{E}_{(n \times p)},$$

sendo:

$$\mathbf{Y} = \begin{bmatrix} y_{111} & y_{121} & \cdots & y_{1p1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{11n_1} & y_{12n_1} & \cdots & y_{1pn_1} \\ y_{211} & y_{221} & \cdots & y_{2p1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{21n_2} & y_{22n_2} & \cdots & y_{2pn_2} \\ & \vdots & & \\ y_{g11} & y_{g21} & \cdots & y_{gp1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{g1n_g} & y_{g2n_g} & \cdots & y_{gpn_g} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ & \vdots & & & \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{g1} & \beta_{g2} & \dots & \beta_{gp} \end{bmatrix} \quad \text{e} \quad \mathbf{E} = \begin{bmatrix} \varepsilon_{111} & \varepsilon_{121} & \dots & \varepsilon_{1p1} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{11n_1} & \varepsilon_{12n_1} & \dots & \varepsilon_{1pn_1} \\ \varepsilon_{211} & \varepsilon_{221} & \dots & \varepsilon_{2p1} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{21n_2} & \varepsilon_{22n_2} & \dots & \varepsilon_{2pn_2} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{g11} & \varepsilon_{g21} & \dots & \varepsilon_{gp1} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{g1n_g} & \varepsilon_{g2n_g} & \dots & \varepsilon_{gpn_g} \end{bmatrix},$$

onde os índices kjm são: k = grupo $(1, \dots, g)$, j = variável $(1, \dots, p)$ e m = repetição $(1, \dots, n_k)$. O modelo linear de análise de variância multivariada também pode ser escrito como:

$$\mathbf{y}_{km} = \boldsymbol{\mu} + \boldsymbol{\tau}_k + \boldsymbol{\varepsilon}_{km},$$

$k = 1, 2, \dots, g$ e $m = 1, 2, \dots, n_k$, em que: $\mathbf{y}_{km}$ é o vetor $(p \times 1)$ de observações do k -ésimo grupo, do m -ésimo indivíduo; $\boldsymbol{\mu}$ é o vetor de constantes $(p \times 1)$ comuns a todos os grupos; $\boldsymbol{\tau}_k$ representa o vetor $(p \times 1)$ de efeitos do k -ésimo grupo; $\boldsymbol{\varepsilon}_{km}$ é o vetor de erros associado a $\mathbf{y}_{km}$ e tem distribuição $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ para todo k e m .

Para estimar os parâmetros de um modelo com um único fator (one way) pode-se estabelecer a restrição $\sum_{k=1}^g n_k \boldsymbol{\tau}_k = \mathbf{0}$. Observe que cada componente do modelo é um vetor de p componentes e os erros associados ao vetor $\mathbf{y}_{km}$ são correlacionados.

Antes de calcular as estatísticas de teste para diferenças de médias dos grupos, o pesquisador deve primeiramente determinar se as medidas dependentes são significativamente correlacionadas - teste de esfericidade de Bartlett (Anexo F). Se forem, a hipótese a ser testada é:

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_g$$

$$H_1: \boldsymbol{\mu}_l \neq \boldsymbol{\mu}_k, \text{ com } l \neq k, l, k = 1, 2, \dots, g.$$

H_0 equivale a testar que os vetores de médias dos g grupos são todos iguais, e H_1 , que ao menos um par de grupos é diferente em pelo menos uma variável.

Para proceder a MANOVA, calcula-se:

- A matriz $\mathbf{T}$ $(p \times p)$ de soma dos quadrados e produtos cruzados total (SQPCTotal):

$$\mathbf{T} = \sum_{k=1}^g \sum_{m=1}^{n_k} (\mathbf{y}_{km} - \bar{\mathbf{y}})(\mathbf{y}_{km} - \bar{\mathbf{y}})';$$

- A matriz $\mathbf{H}$ ($p \times p$) de soma dos quadrados e produtos cruzados *entre* os grupos ou matriz de soma dos quadrados e produtos cruzados da hipótese (SQPCHip):

$$\mathbf{H} = \sum_{k=1}^g n_k (\bar{\mathbf{y}}_k - \bar{\mathbf{y}})(\bar{\mathbf{y}}_k - \bar{\mathbf{y}})' ;$$

- E a matriz $\mathbf{W}$ ($p \times p$) de soma dos quadrados e produtos cruzados *dentro* dos grupos ou matriz de soma de quadrados e produtos cruzados do resíduo (SQPCRes);

$$\mathbf{W} = \sum_{k=1}^g \sum_{m=1}^{n_k} (\mathbf{y}_{km} - \bar{\mathbf{y}}_k)(\mathbf{y}_{km} - \bar{\mathbf{y}}_k)' = \sum_{k=1}^g (n_k - 1) \mathbf{S}_k$$

em que $\mathbf{S}_k$ é a matriz de variâncias e covariâncias amostrais do k -ésimo grupo; $\bar{\mathbf{y}}$ é o vetor ($p \times 1$) de constantes amostrais comuns a todos os grupos; e $\bar{\mathbf{y}}_k$ é o vetor ($p \times 1$) de médias amostrais do k -ésimo grupo.

Vale lembrar que, as matrizes $\mathbf{T}$, $\mathbf{H}$, e $\mathbf{W}$ são simétricas e que demonstra-se facilmente que: $\mathbf{T} = \mathbf{H} + \mathbf{W}$. Logo, para se obter a matriz de SQPCRes basta calcular: $\mathbf{W} = \mathbf{T} - \mathbf{H}$. Com estas matrizes o quadro da MANOVA é apresentado na Tabela 7.

Tabela 7 - Quadro da MANOVA para comparar vetores de médias de k populações

Fonte de Variação (FV)	graus de liberdade (gl)	Matriz de SQPC
Grupo	$g - 1$	$\mathbf{H}$
Resíduo	$v = \sum_{k=1}^g n_k - g$	$\mathbf{W}$
Total Corrigido	$\sum_{k=1}^g n_k - 1$	$\mathbf{T} = \mathbf{H} + \mathbf{W}$

Os gl.'s correspondem ao caso univariado e as distribuições multivariadas teóricas envolvem a densidade de Wishart (Anexo H.1).

Existem vários critérios para testes de hipóteses multivariados, contudo, segundo Johnson e Wichern (2002) e Reis (1997), os quatro mais utilizados são:

- Critério de Wilks** (ou *Lambda de Wilks*): baseado na razão de verossimilhança, é dado por:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{H} + \mathbf{W}|} = \prod_{j=1}^p \frac{1}{1 + \lambda_j} ;$$

ii) **Traço de Pillai:**

$$V = \text{tr} \left[\frac{\mathbf{H}}{\mathbf{H} + \mathbf{W}} \right] = \sum_{j=1}^p \frac{\lambda_j}{1 + \lambda_j} ;$$

iii) **Traço de Hotelling-Lawley:**

$$U = \text{tr} \left(\frac{\mathbf{H}}{\mathbf{W}} \right) = \sum_{j=1}^p \lambda_j ;$$

iv) **Raíz máxima de Roy:** alguns consideram

$$\Theta = \lambda_1 \quad \text{ou} \quad \Theta = \sum_{j=1}^p \frac{\lambda_j}{1 + \lambda_j} ,$$

onde λ_j são os autovalores (Anexo I) obtidos da solução da equação característica dada por:

$$(\mathbf{H} - \lambda_j \mathbf{W}) \mathbf{e}_j = 0 ,$$

com $j = 1, 2, \dots, p$, e $\mathbf{e}_j$ são os autovetores associados aos autovalores λ_j . Esses critérios podem ser aproximados pela distribuição F , tais aproximações podem ser encontradas em Reis (1997), Khattree e Naik (1999), Regazzi (2002) dentre outros e a aproximação é tão melhor quanto maior for o número de indivíduos ($n = n_1 + \dots + n_g$).

Resultados discordantes envolvendo as diferentes estatísticas poderão ser obtidos. Cabe ressaltar que esses quatro testes são competidores e não há teste melhor para todas as situações. Contudo, a decisão sobre a rejeição de H_0 deve ser a do maior número de resultados concordantes em um nível nominal de significância previamente adotado. Quando há empate de resultados entre esses testes, deve-se levar em consideração a seguinte hierarquia: $V \geq \Lambda \geq U \geq \Theta$, sendo os melhores o Traço de Pillai e o Critério de Wilks. Esta hierarquia é obtida de forma geral, combinando a robustez e a potência do teste (REIS, 1997).

No software R o quadro da MANOVA é obtido pelo comando *summary.manova* .

3.3.1.1 Comparações Múltiplas

Da mesma forma que na análise da variância univariada, a rejeição da hipótese da nulidade (H_0) não indica quais grupos ou combinações de grupos são diferentes entre si. A rejeição de H_0 pode ocorrer devido a diferenças entre grupos considerando-se uma variável, mas também, devido à existência de diferenças entre grupos ligadas a certas combinações lineares de variáveis.

Dentre os métodos propostos para comparações múltiplas, utilizou-se os intervalos simultâneos de Bonferroni.

Quando a hipótese H_0 (médias iguais para grupos) é rejeitada, aquelas médias que levaram à rejeição são de interesse e pode-se identificar qual ou quais componentes dos vetores de médias diferem significativamente dos demais. A aproximação de Bonferroni pode ser usada para construir intervalos de confiança simultâneos para a diferença de médias de grupos $\mu_k - \mu_l$ (ou $\tau_k - \tau_l$). Esses intervalos são mais curtos que os obtidos para todos os contrastes e requerem apenas valores críticos da estatística univariada t (FERREIRA, 1996).

Sejam $n = \sum_{k=1}^g n_k$ e μ_{kj} o j -ésimo componente (variável j) de μ_k de médias associadas ao grupo k . Como μ_k é estimado por $\bar{y}_k - \bar{y}$, tem-se que $\hat{\mu}_{kj} = \bar{y}_{kj} - \bar{y}_j$ e $(\hat{\mu}_{kj} - \hat{\mu}_{lj}) = \bar{y}_{kj} - \bar{y}_{lj}$, que é a diferença entre duas médias amostrais independentes.

Desde que p variáveis são consideradas e $g(g-1)/2$ comparações duas a duas serão realizadas, então para cada duas amostras será usado no intervalo- t um valor crítico $t_{n-g}(\alpha/2q)$, onde $q = pg \frac{(g-1)}{2}$ é o número de intervalos simultâneos de confiança declarados. Logo, para o modelo de MANOVA descrito, o intervalo de confiança de no mínimo $(1 - \alpha)$ protegido por Bonferroni para diferenças de médias de grupo $(\mu_{kj} - \mu_{lj})$ é dado por:

$$(\bar{y}_{kj} - \bar{y}_{lj}) \pm t_{(n-g)} \left(\frac{\alpha}{pg(g-1)} \right) \sqrt{\frac{w_{jj}}{n-g} \left(\frac{1}{n_k} + \frac{1}{n_l} \right)}$$

para todas as componentes $j = 1, 2, \dots, p$ e todos os grupos $l < k = 1, 2, \dots, g$. Aqui w_{jj} é o j -ésimo elemento da diagonal de $\mathbf{W}$ (JOHNSON; WICHERN, 2002). Se os extremos do intervalo apresentarem sinais diferentes, aceita-se a hipótese de igualdade entre as componentes.

3.3.2 Resultados e Discussão - Experimento II

Na análise dos dados foram encontrados 12 *outliers*: indivíduos 5, 8, 9, 19, 63, 70, 82, 92, 103, 104, 109 e 117. Estas indicações podem ser confirmadas pelos gráficos de perfis (Figura 8), gráfico das últimas componentes principais (Figura 9) e *Q-Q plot's* (Figura 10 (a) e (b)).

Após a retirada dos *outliers* foram calculados os vetores de média geral e de médias dos grupos (Tabela 8). Como os *outliers* das análises univariada e multivariada são diferentes, as médias dos grupos diferiram um pouco daquelas apresentadas na Tabela 13. A matriz de correlações amostrais, $\mathbf{R}$, também foi obtida.

Figura 8 - Perfis individuais dos dados iniciais (a) e dos dados sem os *outliers* multivariados (b)

Figura 9 - Gráficos das últimas componentes principais dos dados iniciais (a) e dos dados sem os *outliers* (b)

Tabela 8 - Médias das variáveis respostas geral ($\bar{y}$) e por grupo

variável	$\bar{y}$	Grupos			
		MAB	CJA	DPE	LAB
Y_1	34,60	28,99	31,07	29,86	49,80
Y_2	67,36	62,23	68,17	63,91	75,44
Y_3	37,97	36,00	37,13	36,15	43,00
Y_4	1,77	1,73	1,83	1,76	1,76
Y_5	9,03	5,93	11,47	6,15	12,42
Y_6	7,12	6,08	6,56	6,42	9,65
Y_7	0,82	0,88	0,45	0,73	1,29
Y_8	0,84	0,79	0,20	0,76	1,75
n	108	26	30	27	25

Figura 10 - *Q-Q plot's* dos dados iniciais (a), dados sem os *outliers* (b) e dados transformados (c)

$$R = \begin{bmatrix} 1,0000 & 0,8715 & 0,8483 & 0,3316 & 0,8855 & 0,6216 & 0,4312 & 0,2180 \\ 0,8715 & 1,0000 & 0,8788 & 0,5430 & 0,8128 & 0,6320 & 0,3902 & 0,1723 \\ 0,8483 & 0,8788 & 1,0000 & 0,0800 & 0,7941 & 0,6018 & 0,3840 & 0,1715 \\ 0,3316 & 0,5430 & 0,0800 & 1,0000 & 0,3023 & 0,2560 & 0,1582 & 0,0609 \\ 0,8855 & 0,8128 & 0,7941 & 0,3023 & 1,0000 & 0,5048 & 0,3907 & 0,2672 \\ 0,6216 & 0,6320 & 0,6018 & 0,2560 & 0,5048 & 1,0000 & 0,2331 & 0,2255 \\ 0,4312 & 0,3902 & 0,3840 & 0,1582 & 0,3907 & 0,2331 & 1,0000 & 0,6232 \\ 0,2180 & 0,1723 & 0,1715 & 0,0609 & 0,2672 & 0,2255 & 0,6232 & 1,0000 \end{bmatrix}$$

Tabela 9 - P-valores dos testes de Shapiro-Wilk das variáveis iniciais (SW1) e das variáveis transformadas (SW2) segundo indicação da família de transformações Box-Cox

Variável	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
SW1	0,0002	0,0039	0,1381	0,2837	$1,827 \times 10^{-06}$	0,0079	0,0493	$2,118 \times 10^{-11}$
Box-Cox	-0,67	-1,20	-1,45	0,13	-0,50	0,19	0,67	0,14
SW2	0,1917	0,3683	0,2298	0,2814	0,0672	0,5860	0,5357	0,0761

Histogramas, *boxplot's* (Figura 11) e *normal plot's* dos dados das oito variáveis respostas mostraram falta de normalidade univariada e os testes de Shapiro-Wilk confirmaram as suspeitas (Tabela 9). Optou-se em transformar todas as variáveis, segundo a família de transformações Box-Cox. Em seguida, os testes de Shapiro-Wilk foram novamente realizados, indicando agora a normalidade univariada das mesmas (Tabela 9). A Figura 12 mostra a normalidade univariada satisfeita para todas as variáveis envolvidas na análise. A normalidade bivariada das variáveis transformadas foi verificada por gráficos *boxplot bivariado* (Figura 13), que também foram satisfatórios. Assim, a normalidade multivariada pode então ser garantida (Figura 10 (c)).

Figura 11 - *Boxplot's* e gráficos de dispersão das variáveis iniciais

A hipótese H_0 do teste de esfericidade de Bartlett (Anexo F) foi rejeitada ao nível de significância de 5% ($\chi_{calc} = 1081,40 > \chi_{(28; 5\%)} = 41,34$), concluindo-se que as variáveis respostas não são independentes, tornando vantajoso o uso da MANOVA.

O teste de Box, para igualdade de matrizes de variâncias e covariâncias entre os grupos (Anexo G), rejeitou a hipótese H_0 ao nível de significância a 1% ($F_{calc} = 2,069 > F_{(108; 23277; 1\%)} = 1,3446$), concluindo-se que as matrizes de variâncias e covariâncias dos grupos são diferentes, mesmo com os dados transformados. Felizmente, segundo Hair Jr. et al. (2006), uma violação dessa suposição tem impacto mínimo se os grupos têm aproximadamente o mesmo tamanho. Para este ensaio tem-se que: $\frac{n_{maior}}{n_{menor}} = \frac{30}{25} = 1,2 < 1,5$, satisfazendo a condição apresentada pelo autor (descrita em 3.3.1).

Satisfeitas todas as suposições, aplicou-se a técnica multivariada. O quadro da MANOVA e as estatísticas de teste podem ser conferidas nas Tabelas 10 e 11.

As quatro estatísticas de testes rejeitaram a hipótese H_0 , de que os vetores de médias

Figura 12 - Gráficos *normal plot's* e gráficos de dispersão das variáveis transformadas

dos quatro grupos são todos iguais, pois foram extremamente significantes ao nível de de 1% (ou 5%), indicando que existe pelo menos um grupo com o vetor de médias diferente.

Para identificar qual (ou quais) componente dos vetores de médias que diferem significativamente dos demais calculou-se os intervalos simultâneos protegidos por Bonferroni ao nível $\alpha = 5\%$ (Tabela 12). Os resultados dessas comparações encontram-se na Tabela 13.

Figura 13 - *Boxplot's* bivariados das variáveis transformadas

Tabela 10 - Quadro da MANOVA para comparar os vetores de médias dos quatro grupos

FV	gl	Matriz de SQPC							
Grupo	3	$\begin{bmatrix} 0,0179 & 0,0007 & 0,0007 & 0,0004 & 0,0578 & -0,0713 & -0,2340 & -0,0959 \\ 0,0007 & 0,0000 & 0,0000 & 0,0000 & 0,0032 & -0,0029 & -0,0061 & -0,0018 \\ 0,0007 & 0,0000 & 0,0000 & 0,0000 & 0,0023 & -0,0028 & -0,0087 & -0,0035 \\ 0,0004 & 0,0000 & 0,0000 & 0,0010 & -0,0128 & -0,0012 & -0,0562 & -0,0337 \\ 0,0578 & 0,0032 & 0,0023 & -0,0128 & 0,4207 & -0,2226 & -0,0360 & 0,1575 \\ -0,0713 & -0,0029 & -0,0028 & -0,0012 & -0,2226 & 0,2885 & 0,9143 & 0,3788 \\ -0,2340 & -0,0061 & -0,0087 & -0,0562 & -0,0360 & 0,9143 & 5,5770 & 2,8255 \\ -0,0959 & -0,0018 & -0,0035 & -0,0337 & 0,1575 & 0,3788 & 2,8255 & 1,5064 \end{bmatrix}$							
Res	104	$\begin{bmatrix} 0,0335 & 0,0014 & 0,0011 & -0,0054 & 0,1011 & -0,1031 & -0,2250 & -0,0787 \\ 0,0014 & 0,0001 & 0,0001 & -0,0004 & 0,0046 & -0,0047 & -0,0090 & -0,0029 \\ 0,0011 & 0,0001 & 0,0000 & 0,0000 & 0,0035 & -0,0036 & -0,0066 & -0,0024 \\ -0,0054 & -0,0004 & 0,0000 & 0,0056 & -0,0192 & 0,0170 & 0,0346 & 0,0079 \\ 0,1011 & 0,0046 & 0,0035 & -0,0192 & 0,4422 & -0,3305 & -0,6075 & -0,2488 \\ -0,1031 & -0,0047 & -0,0036 & 0,0170 & -0,3305 & 0,7439 & 0,5582 & 0,2181 \\ -0,2250 & -0,0090 & -0,0066 & 0,0346 & -0,6075 & 0,5582 & 7,2848 & 2,7499 \\ -0,0787 & -0,0029 & -0,0024 & 0,0079 & -0,2488 & 0,2181 & 2,7499 & 1,8786 \end{bmatrix}$							
Total	107	$\begin{bmatrix} 0,0514 & 0,0021 & 0,0018 & -0,0050 & 0,1589 & -0,1744 & -0,4589 & -0,1746 \\ 0,0021 & 0,0001 & 0,0001 & -0,0004 & 0,0078 & -0,0076 & -0,0151 & -0,0048 \\ 0,0018 & 0,0001 & 0,0001 & 0,0000 & 0,0058 & -0,0064 & -0,0152 & -0,0058 \\ -0,0050 & -0,0004 & 0,0000 & 0,0066 & -0,0320 & 0,0158 & -0,0215 & -0,0258 \\ 0,1589 & 0,0078 & 0,0058 & -0,0320 & 0,8630 & -0,5531 & -0,6434 & -0,0913 \\ -0,1744 & -0,0076 & -0,0064 & 0,0158 & -0,5531 & 1,0324 & 1,4726 & 0,5970 \\ -0,4589 & -0,0151 & -0,0152 & -0,0215 & -0,6434 & 1,4726 & 12,8618 & 5,5754 \\ -0,1746 & -0,0048 & -0,0058 & -0,0258 & -0,0913 & 0,5970 & 5,5754 & 3,3850 \end{bmatrix}$							

FV: Fonte de Variação; gl: graus de liberdade; Res: Resíduo.

Tabela 11 - Valores das estatísticas para testar H_0 (médias iguais para grupos) da MANOVA

Estatísticas	valor	v_1	v_2	aprox.F	$\Pr(>F)$
Wilks	0,13	24	281,93	11,75	$p < 0,001$
Traço de Pillai	1,30	24	297	9,47	$p < 0,001$
Traço de Hotelling-Lawley	3,51	24	287	13,97	$p < 0,001$
Raíz máxima de Roy	2,46	8	99	30,47	$p < 0,001$

v_1 e v_2 são os graus de liberdade da distribuição F .

Tabela 12 - Limites Inferiores (LI) e Limites Superiores (LS) dos intervalos de confiança simultâneos para testar o vetor de média dos grupos (G): (1) MAB, (2) CJA, (3) DPE e (4) LAB, para todas as das variáveis (Y)

G	G	Y	LI	LS		G	G	Y	LI	LS	
1	2	1	-0,0127	0,0197		1	2	5	0,0549	0,1728	*
1	3	1	-0,0185	0,0147		1	3	5	-0,0789	0,0421	
1	4	1	0,0139	0,0478	*	1	4	5	0,0543	0,1775	*
2	3	1	-0,0215	0,0107		2	3	5	-0,1907	-0,0739	*
2	4	1	0,0109	0,0437	*	2	4	5	-0,0576	0,0616	
3	4	1	0,0159	0,0495	*	3	4	5	0,0732	0,1954	*
1	2	2	-0,0001	0,0014		1	2	6	-0,0983	0,0546	
1	3	2	-0,0006	0,0009		1	3	6	-0,0900	0,0668	
1	4	2	0,0007	0,0022	*	1	4	6	-0,2127	-0,0528	*
2	3	2	-0,0013	0,0002		2	3	6	-0,0655	0,0859	
2	4	2	0,0000	0,0015	*	2	4	6	-0,1882	-0,0337	*
3	4	2	0,0005	0,0021	*	3	4	6	-0,2004	-0,0420	*
1	2	3	-0,0004	0,0008		1	2	7	0,1211	0,5997	*
1	3	3	-0,0006	0,0006		1	3	7	-0,1039	0,3868	
1	4	3	0,0006	0,0019	*	1	4	7	-0,5138	-0,0136	*
2	3	3	-0,0008	0,0004		2	3	7	-0,4558	0,0180	
2	4	3	0,0004	0,0016	*	2	4	7	-0,8659	-0,3823	*
3	4	3	0,0006	0,0019	*	3	4	7	-0,6530	-0,1573	*
1	2	4	-0,0148	-0,0016	*	1	2	8	0,1048	0,3478	*
1	3	4	-0,0096	0,0040		1	3	8	-0,0703	0,1789	
1	4	4	-0,0092	0,0047		1	4	8	-0,2199	0,0341	
2	3	4	-0,0012	0,0120		2	3	8	-0,2923	-0,0517	*
2	4	4	-0,0008	0,0126		2	4	8	-0,4421	-0,1965	*
3	4	4	-0,0063	0,0074		3	4	8	-0,2731	-0,0214	*

* indica diferença significativa ao nível $\alpha = 5\%$.

As comparações múltiplas mostraram que existe diferença entre as médias dos grupos MAB e CJA para as variáveis razão corporal (Y_4), peso da concha (Y_5), peso da glândula prostática (Y_7) e peso da glândula albumina (Y_8). As médias dos grupos MAB e LAB diferiram para o peso do animal (Y_1), comprimento da concha (Y_2), largura da concha (Y_3), peso da concha (Y_5), peso da carne (Y_6) e peso da glândula prostática (Y_7). Para as variáveis peso da concha (Y_5) e peso da glândula albumina (Y_8) existe diferença entre as médias dos grupos CJA e DPE. Também existe diferença entre as médias dos grupos CJA e LAB para todas as variáveis, com exceção da razão corporal (Y_4) e do peso da concha (Y_5). Já as médias dos grupos DPE e LAB diferiram para todas as variáveis, com exceção da razão corporal (Y_4).

Tabela 13 - Resultados obtidos para as comparações múltiplas de médias por variável (MANOVA e ANOVA)

MANOVA								
Média	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
MAB	28,99 b	62,23 b	36,00 b	1,73 b	5,93 b	6,08 b	0,88 b	0,79 ab
CJA	31,07 b	68,17 b	37,13 b	1,83 a	11,47 a	6,56 b	0,45 c	0,20 c
DPE	29,86 b	63,91 b	36,15 b	1,76 ab	6,15 b	6,42 b	0,73 bc	0,76 b
LAB	49,80 a	75,44 a	43,00 a	1,76 ab	12,42 a	9,65 a	1,29 a	1,75 a
ANOVA								
Média	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
MAB	28,75 b	61,96 c	35,92 a	1,72 b	5,70 b	5,92 b	0,87 b	0,74 b
CJA	31,48 b	68,17 b	37,17 a	1,84 a	11,64 a	6,49 b	0,45 c	0,20 c
DPE	29,44 b	63,48 c	36,18 a	1,75 b	5,87 b	6,26 b	0,74 b	0,60 b
LAB	52,63 a	76,28 a	43,96 b	1,74 b	12,88 a	10,04 a	1,29 a	1,27 a

Os resultados obtidos com o uso da MANOVA concordam com os obtidos na ANOVA, entretanto, o número de diferenças significativas entre médias de grupos no caso multivariado é menor do que no caso univariado. Das 48 comparações possíveis entre as médias dos quatro grupos, 31 delas foram significativas no caso univariado e somente 25 indicaram diferenças entre grupos no caso multivariado. As médias dos grupos que diferiram no caso univariado e foram iguais no multivariado são: MAB e CJA na variável comprimento da concha (Y_2); MAB e LAB na variável peso da glândula albumina (Y_8); CJA e DPE nas variáveis comprimento da concha (Y_2), razão corporal (Y_4) e peso da glândula prostática (Y_7); e CJA e LAB na variável razão corporal (Y_4).

Segundo Demétrio (1985), essa discordância se justifica porque no caso multivariado o critério de rejeição de H_0 é mais rigoroso por levar em consideração um nível de significância conjunto, englobando todas as variáveis analisadas. Já para o caso univariado o nível de significância é somado isoladamente por análise, desconhecendo-se o nível conjunto de probabilidade para todas as variáveis, que será tanto maior quanto maior for o número de variáveis. Este fato acaba tornando impossível uma comparação exata dos resultados. Neste exemplo, o nível conjunto de probabilidade considerado pelos testes multivariados foi de 0,05; porém, na análise univariada o nível conjunto estará em torno de $1 - (0,95^8) = 0,34$, o que é inaceitável.

No geral, as médias das variáveis por grupos foram maiores em LAB, com exceção

da variável razão corporal. Segundo Caetano (2005), os maiores valores para peso, largura e comprimento já eram esperados para os animais criados em cativeiro (LAB), uma vez que a prática de criação destes animais obedece a um rigor técnico de criação, norteados pela pesquisa gerada no próprio setor. Acrescenta-se o fato de que estes animais não necessitam de muito esforço para se alimentarem e não sofrem os efeitos das mudanças climáticas bruscas, como ocorre com os animais asselvajados. Para Flauzino et al. (1997, apud Caetano, 2005), a taxa de postura estaria fortemente relacionada aos níveis de cálcio na dieta de reprodutores, sendo a exigência muito inferior no período reprodutivo do que no período de crescimento.

Já o grupo MAP apresentou as menores médias das variáveis, com exceção das variáveis peso glândula prostática (Y_7) e peso da glândula albumina (Y_8). O grupo CJA teve a maior média de razão corporal (Y_4) e as menores médias de peso da glândula prostática (Y_7) e peso da glândula albumina (Y_8). Estes resultados foram análogos aos da análise univariada. Porém, para a razão corporal (Y_4), na análise multivariada, o grupo CJA não diferiu dos animais em cativeiro (LAB), mas na univariada estes grupos são diferentes. Este resultado não chega a influenciar as demais conclusões. Caetano (2005) lembra que a glândula de albumina variou na sua forma, assim como a cor e tamanho, e essa variação depende do estado reprodutivo dos animais.

Observa-se também que o peso médio da concha (Y_5) para os grupos LAB e CJA não diferem; porém, o peso médio da carne (Y_6) difere para estes grupos. Caetano (2005) justifica este resultado pelo fato de que os animais da situação CJA (em sua maioria) encontravam-se magros, resultantes de estado de quase hibernação devido a escassez de alimentos, diferentemente da situação em cativeiro. Desta maneira, fica novamente confirmada as condições superiores de cativeiro, para o desenvolvimento e reprodução destes moluscos.

Em relação às variáveis relacionadas às glândulas reprodutivas e observando o número de ovos encontrados em cada grupo, os resultados evidenciam a superioridade reprodutiva dos animais da situação LAB e os baixos valores dos animais da situação asselvajados (CJA).

No geral, os resultados aqui encontrados não divergem dos obtidos pela análise univariada; contudo, existem algumas vantagens em aplicar a MANOVA ao invés de p ANOVAS. Deste modo, se o interesse do pesquisado estiver em conclusões a respeito de uma variável específica, a análise univariada é suficiente. Contudo, no caso do interesse do pesquisador estar numa conclusão conjunta, a análise multivariada leva a conclusões com um nível de significância conjunto controlado.

3.4 Material e Métodos - Experimento III

Para a aplicação da Análise de Componentes Principais (ACP), utilizou-se o conjunto de dados do experimento II, acrescido das variáveis referentes a percentagem das frações dos animais. Desta maneira, 13 variáveis foram consideradas: Y_1 : peso do animal (g); Y_2 : comprimento da concha (mm); Y_3 : largura da concha (mm); Y_4 : razão corporal ($= \frac{Y_2}{Y_3}$); Y_5 : peso da concha (g); Y_6 : percentual de concha (%); Y_7 : peso da carne (g); Y_8 : percentual de carne (%); Y_9 : peso glândula prostática (g); Y_{10} : percentual de glândula prostática (%); Y_{11} : peso da glândula albumina (g); Y_{12} : percentual de glândula albumina (%); e Y_{13} : percentual do aparelho reprodutor que é a soma da glândula prostática, glândula albumina e massa de ovos (%). As estatísticas descritivas dos dados se encontram no Apêndice C.

O objetivo do presente estudo é reduzir a dimensionalidade do conjunto original de variáveis com a menor perda de informação possível, eliminando as informações redundantes em decorrência da correlação entre variáveis, e descartar as que contribuem pouco para explicar a variação total, evitando que sejam utilizadas nas análises de experimentos futuros.

Como o conjunto de dados foi ampliado com outras variáveis, os *outliers* multivariados serão novamente identificados e retirados da análise. A normalidade multivariada das variáveis não será verificada, pois não serão feitas inferências com nenhuma das componentes obtidas. Logo, consideraram-se apenas os dados originais. Os dados iniciais foram corrigidos para o efeito dos grupos e assim, a ACP foi iniciada a partir deste ponto.

3.4.1 Análise de Componentes Principais (ACP)

O objetivo principal da ACP é o de explicar a estrutura de variâncias e covariâncias de um vetor aleatório composto de p -variáveis aleatórias iniciais, podendo-se resumir sua informação, ou reduzir a dimensão dos dados, eliminando as informações redundantes contida no complexo das variáveis originais, o que torna os resultados mais simples e de interpretação mais clara. O desenvolvimento desta análise tem as seguintes características:

- a) Não requer uma suposição de normalidade multivariada. Entretanto, as componentes principais (CP's) derivadas para populações com distribuição normal multivariada têm interpretações úteis em termos de elipsóides de confiança, além da possibilidade de se fazer algumas inferências sobre eles (alguns testes estão presentes em GROSSMAN; NICKERSON; FREEMAN, 1991 e MINGOTI, 2005).

- b) Não impõe qualquer modelo causal, mas também não permite detectar quaisquer relações de causa-efeito entre as variáveis iniciais mesmo se existirem.
- c) É uma técnica de análise intermediária (exploratória) em muitas investigações (como por exemplo, na regressão múltipla, análise de agrupamentos etc.) e, portanto não constitui um método final e conclusivo (FERREIRA, 1996).
- d) A técnica depende somente da estrutura de covariâncias $\mathbf{S}$ ou da matriz de correlações $\mathbf{R}$ do conjunto de variáveis observadas.

A técnica de ACP, realizada a partir da matriz de variâncias e covariâncias, consiste em transformar um conjunto de variáveis iniciais $Y_1, Y_2, \dots, Y_p$, correlacionadas entre si, em um novo conjunto de variáveis $C_1, C_2, \dots, C_p$, não correlacionadas (ortogonais); chamadas de *Componentes Principais*, arranjadas em ordem decrescente de variâncias (REGAZZI, 2002 e STEARNS et al., 2005). Isso é feito construindo combinações lineares das variáveis originais.

Teorema 3.1 : Seja $\mathbf{S}$ uma matriz de variâncias e covariâncias amostrais de ordem p , de um conjunto de variáveis $Y_1, Y_2, \dots, Y_p$, com pares de autovalor-autovetor²³ $(\lambda_j, \mathbf{e}_j)$, $j = 1, 2, \dots, p$, o *j-ésimo componente principal amostral* é dado por:

$$C_j = e_{1j}Y_1 + e_{2j}Y_2 + \dots + e_{pj}Y_p = \mathbf{Y} \mathbf{e}_j ,$$

onde $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$; $\mathbf{Y}$ é a matriz de observações das variáveis $Y_1, Y_2, \dots, Y_p$; $\mathbf{e}_j' \mathbf{e}_j = \sum_{j=1}^p \mathbf{e}_j^2 = 1$; e $\mathbf{e}_j' \mathbf{e}_{j'} = 0$, para $j \neq j'$, $j, j' = 1, 2, \dots, p$.

Dessa forma, pela decomposição espectral (Anexo J), facilmente prova-se as seguintes propriedades (demonstração em JOHNSON; WICHERN, 2002 e FERREIRA, 1996):

- i) $var(C_j) = var(\mathbf{Y} \mathbf{e}_j) = \lambda_j$
- ii) $cov(C_j ; C_{j'}) = cov(\mathbf{Y} \mathbf{e}_j ; \mathbf{Y} \mathbf{e}_{j'}) = 0$, significando que C_j e $C_{j'}$ são ortogonais para $j, j' = 1, 2, \dots, p$.

Geometricamente, essas combinações lineares particulares representam a seleção de novos sistemas de coordenadas obtidos pela rotação do sistema original que tem $Y_1, Y_2, \dots, Y_p$ como eixos das coordenadas. Os novos eixos representam as direções com máxima variabilidade e

²³Na prática, o sinal negativo apresentado por alguns autovetores apenas indica que estes estão atuando em sentido contrário aos demais dentro de cada componente.

forneem uma descrição simples e parcimoniosa da estrutura de covariâncias (e a correspondente interdependência entre as variáveis).

Do **Teorema 3.1** e da propriedade *i*), segue que:

$$\text{var}(C_1) \geq \text{var}(C_2) \geq \dots \geq \text{var}(C_p) \geq 0 .$$

A *porcentagem da variação total explicada pela j -ésima CP* é dada por:

$$\%VarExp(C_j) = \frac{\lambda_j}{\sum_{j'=1}^p \lambda_{j'}} \times 100 , \quad j = 1, 2, \dots, p ,$$

e a *variabilidade total da amostra* é

$$VT = \text{tr}(\mathbf{S}) = \sum_{j=1}^p \text{var}(Y_j) = \lambda_1 + \lambda_2 + \dots + \lambda_p = \sum_{j=1}^p \text{var}(C_j)$$

O coeficiente de correlação entre a componente C_i e a variável Y_j é:

$$\rho_{[C_j, Y_{j'}]} = \frac{\text{cov}(C_j ; Y_{j'})}{\sqrt{\text{var}(C_j)\text{var}(Y_{j'})}} = \frac{\lambda_j e_{jj'}}{\sqrt{\lambda_j} \sqrt{s_{j'j'}}} = \frac{e_{jj'} \sqrt{\lambda_j}}{\sqrt{s_{j'j'}}$$

Reis (1997) ressalta uma séria desvantagem dessa técnica multivariada: em geral, as CP's não são invariantes com relação a transformações nas escalas. A tentativa de resolução desse problema é feita padronizando todas as variáveis (Anexo A), o que equivale a aplicar a ACP à matriz de correlações $\mathbf{R}$. Desta maneira, a soma dos autovalores passa a ser

$$\sum_{j=1}^p \lambda_j = \text{tr}(\mathbf{R}) = p ,$$

que é o número de variáveis. A proporção de variância explicada pela componente C_j passa então a ser λ_j/p , $j = 1, 2, \dots, p$. E a correlação entre C_j e $Z_{j'}$ é definida por:

$$\rho_{[C_j, Z_{j'}]} = \frac{\text{cov}(C_j ; Z_{j'})}{\sqrt{\text{var}(C_j)\text{var}(Z_{j'})}} = \frac{\lambda_j e_{jj'}}{\sqrt{\lambda_j} \sqrt{s_{j'j'}}} = e_{jj'} \sqrt{\lambda_j} .$$

A partir das p variáveis originais é possível obter no máximo p CP's. No entanto, em geral, deseja-se obter uma redução do número de variáveis, isto é, a informação contida nas p -variáveis originais é substituída pela informação contida em k ($k < p$) CP's não correlacionadas, sem perda de uma quantidade demasiada de informação. A idéia principal nesse procedimento é que poucas dentre as primeiras CP's incorporam a maior variabilidade dos dados originais, podendo-se racionalmente descartar as demais componentes, reduzindo o número de variáveis (BARBOSA et

al., 2005b). Essa aproximação depende do número de CP's mantidas no sistema e pode ser medida através da avaliação da proporção de variância total explicada por elas. Existem vários critérios práticos para determinar quantas componentes deve-se excluir da análise e os mais utilizados na literatura, segundo Reis (1997), são:

- i) **Scree-Plot** (gráfico de cotovelo): foi proposto por Cattell²⁴ em 1966, é um gráfico dos autovalores (Anexo I), λ_j , em função da ordem das CP's, representando graficamente a porcentagem de variância explicada por componente. Quando esta porcentagem se reduz e a curva passa a ser quase paralela ao eixo das abscissas, podemos excluir os componentes correspondentes. Na Figura 14 sugere-se excluir todas as CP's acima de C_2 .

Figura 14 - Scree-Plot ilustrativo de um exemplo com seis componentes principais

- ii) Incluir as r primeiras CP's suficientes para explicar 70% da variação total ou mais.
- iii) Critério de Kaiser (1958): Incluir apenas as CP's cujos autovalores são superiores ou iguais à média dos autovalores. Se a análise for feita a partir de uma matriz de correlações, reter as CP's com variâncias ≥ 1 .

Para cada uma das CP's escolhidas pode-se calcular os escores de cada elemento amostral. Esses escores podem ser analisados utilizando-se técnicas estatísticas usuais como análise de variância e análise de regressão, dentre outras. No caso da análise multivariada, a ACP serve para auxiliar na escolha das variáveis independentes.

Vale lembrar que, deve-se interpretar as CP's obtidas, pois elas constituem as novas variáveis respostas que serão utilizadas nas análises subseqüentes do estudo. A interpretação de

²⁴CATTELL, R.B. The scree test for the number of factors. **Multivariate Behaviour Research**, Mahwah, v.1, p.245-176, 1966.

cada CP é baseada nos valores dos coeficientes da combinação linear das variáveis originais (ou transformadas) de cada CP, isto é, as variáveis originais que mais contribuem para a CP, C_j , são aquelas que possuem os coeficientes $\geq 0,50$. Os coeficientes de correlações entre as CP's e as variáveis originais também podem ser utilizados na interpretação, porém devem servir apenas como auxiliares, pois são medidas univariadas. A presença do pesquisador nesta etapa é de extrema utilidade, pois ele conhece a situação em estudo, e facilmente pode interpretar as novas variáveis.

Os critérios propostos até aqui são empregados para decidir quantas CP's podem ser consideradas no lugar das variáveis originais. Quando o objetivo da análise é selecionar as variáveis originais que podem ser desconsideradas em ensaios futuro, por serem redundantes, o critério descrito por Mardia et al. (1997, apud BARBOSA et al., 2006) é o mais indicado.

Segundo o autor, as variáveis altamente correlacionadas com as CP's de menor variância representam variação praticamente insignificante, sendo estas passíveis de descarte. O critério do número de variáveis descartadas é, conforme recomendação de Jolliffe^{25,26} (1972, 1973), baseado em dados simulados e reais; quando a ACP foi feita a partir da matriz de correlação, estabelece que o número de variáveis descartadas deve ser igual ao número de componentes cuja variância (autovalor - Anexo I) é inferior a 0,7.

A ACP no *software* R pode ser implementada pelo comando *princomp*, ou então utilizando-se o pacote *amap* e o comando *acp* (mais detalhes em LUCAS, 2007).

3.4.2 Resultados e Discussão - Experimento III

Utilizando os mesmos procedimentos descritos em 2.2.5, 14 *outliers* multivariados foram identificados (indivíduos: 1, 5, 8, 19, 48, 73, 92, 95, 103, 104, 106, 108, 109 e 117) e retirados da análise. Removeu-se o efeito de grupos dos dados originais e em seguida a matriz de variâncias e covariâncias amostrais, **S**, foi calculada.

²⁵JOLLIFFE, I.T. Discarding variables in a principal component analysis. I. Artificial data. **Applied Statistics**, Londres, v.21, p.160-173, 1972.

²⁶JOLLIFFE, I.T. Discarding variables in a principal component analysis. II. Real data. **Applied Statistics**, Londres, v.22, p.21-31, 1973.

$$\mathbf{S} = \begin{bmatrix} 138,62 & & & & & & & & & & & & & \\ 91,40 & 71,96 & & & & & & & & & & & & \\ 36,14 & 26,90 & 13,25 & & & & & & & & & & & \\ 0,66 & 0,59 & 0,08 & 0,01 & & & & & & & & & & \\ 53,85 & 36,68 & 13,99 & 0,29 & 24,98 & & & & & & & & & \\ 35,18 & 26,86 & 9,75 & 0,23 & 22,01 & 29,48 & & & & & & & & \\ 19,61 & 14,00 & 5,39 & 0,11 & 7,27 & 4,40 & 5,27 & & & & & & & \\ -19,26 & -10,17 & -4,48 & -0,05 & -7,78 & -5,06 & 4,61 & 26,74 & & & & & & \\ 1,92 & 1,20 & 0,62 & 0,00 & 0,64 & 0,23 & 0,22 & -0,41 & 0,12 & & & & & \\ 2,84 & 1,70 & 0,84 & 0,01 & 0,82 & -0,02 & 0,28 & -0,77 & 0,27 & 0,66 & & & & \\ 3,59 & 1,77 & 1,09 & -0,00 & 1,34 & 0,00 & 0,55 & -0,42 & 0,23 & 0,50 & 1,05 & & & \\ 2,44 & 0,73 & 0,94 & -0,02 & 0,79 & 0,53 & 0,53 & -0,01 & 0,44 & 1,05 & 2,05 & 4,66 & & \\ -3,23 & -4,08 & -0,51 & -0,07 & -5,57 & -10,68 & -2,49 & -4,90 & 0,69 & 1,85 & 0,52 & 1,59 & 32,34 & \end{bmatrix}$$

Avaliando a matriz de variâncias e covariâncias, verifica-se que as variâncias (diagonal da matriz $\mathbf{S}$) são bem diferentes (maior variância = 138,62 e a menor variância = 0,01) entre si, como também as unidades de mensuração de cada variável (g, mm e %). Assim, optou-se por padronizar as variáveis (como no Anexo A) para evitar efeito de escala; as variáveis padronizadas foram denotadas por Z_j , $j = 1, 2, \dots, 13$. Conseqüentemente, a matriz de correlações $\mathbf{R}$, apresentada a seguir, foi utilizada para a aplicação da técnica.

$$\mathbf{R} = \begin{bmatrix} 1,00 & & & & & & & & & & & & & \\ 0,92 & 1,00 & & & & & & & & & & & & \\ 0,84 & 0,87 & 1,00 & & & & & & & & & & & \\ 0,52 & 0,65 & 0,19 & 1,00 & & & & & & & & & & \\ 0,92 & 0,87 & 0,77 & 0,53 & 1,00 & & & & & & & & & \\ 0,55 & 0,58 & 0,49 & 0,40 & 0,81 & 1,00 & & & & & & & & \\ 0,73 & 0,72 & 0,64 & 0,43 & 0,63 & 0,35 & 1,00 & & & & & & & \\ -0,32 & -0,23 & -0,24 & -0,10 & -0,30 & -0,18 & 0,39 & 1,00 & & & & & & \\ 0,47 & 0,41 & 0,50 & 0,08 & 0,37 & 0,12 & 0,28 & -0,23 & 1,00 & & & & & \\ 0,30 & 0,25 & 0,29 & 0,08 & 0,20 & -0,01 & 0,15 & -0,18 & 0,96 & 1,00 & & & & \\ 0,30 & 0,20 & 0,29 & -0,03 & 0,26 & 0,18 & 0,24 & -0,08 & 0,66 & 0,60 & 1,00 & & & \\ 0,10 & 0,04 & 0,12 & -0,09 & 0,07 & 0,05 & 0,11 & -0,00 & 0,59 & 0,60 & 0,93 & 1,00 & & \\ -0,05 & -0,08 & -0,02 & -0,11 & -0,20 & -0,35 & -0,19 & -0,17 & 0,35 & 0,40 & 0,09 & 0,13 & 1,00 & \end{bmatrix}$$

Esta matriz $\mathbf{R}$ apresenta os coeficientes de correlação simples entre as 13 características dos animais em estudo. Os coeficientes de correlação que foram superiores a 0,50 são todos positivos, indicando que estas variáveis são diretamente proporcionais. Em destaque na matriz $\mathbf{R}$ estão as correlações que foram significativas segundo o teste de hipótese para coeficiente de correlação ($H_0: \rho(Y_j, Y_{j'}) = 0$).

Os autovalores e autovetores da matriz $\mathbf{R}$ foram calculados (Tabela 14) e cada CP pôde ser obtida como descrito em 3.4.1. Os coeficientes das CP's e suas correlações com as variáveis padronizadas são apresentados nas Tabelas 15 e 16.

Tabela 14 - Autovalores (λ_j , $j = 1, 2, \dots, 13$) e porcentagem da variação explicada pelas componentes principais (VCP)

CP	λ_j	VCP	VCP Acumulada
C_1	5,5398	0,42614	0,42614
C_2	2,9465	0,22665	0,65279
C_3	1,4495	0,11150	0,76429
C_4	1,1193	0,08609	0,85039
C_5	0,7953	0,06118	0,91156
C_6	0,4556	0,03504	0,94661
C_7	0,4364	0,03357	0,98017
C_8	0,1647	0,01267	0,99284
C_9	0,0531	0,00409	0,99693
C_{10}	0,0247	0,00190	0,99883
C_{11}	0,0102	0,00079	0,99962
C_{12}	0,0040	0,00031	0,99993
C_{13}	0,0009	0,00007	1,00000

Tabela 15 - Coeficientes das cinco primeiras componentes principais e suas correlações (valores entre parênteses) com as variáveis Z_p , $p = 1, 2, \dots, 13$

Variável	Componentes Principais				
	C_1	C_2	C_3	C_4	C_5
Z_1	-0,397(-0,935)	0,110(0,189)	0,093(0,112)	0,087(0,092)	0,097(0,087)
Z_2	-0,390(-0,919)	0,164(0,281)	0,069(0,083)	0,140(0,148)	-0,010(-0,009)
Z_3	-0,362(-0,852)	0,059(0,101)	0,057(0,069)	0,081(0,086)	0,508(0,453)
Z_4	-0,221(-0,520)	0,220(0,378)	0,053(0,064)	0,162(0,172)	-0,821(-0,733)
Z_5	-0,386(-0,909)	0,176(0,302)	0,067(0,081)	-0,142(-0,151)	0,022(0,019)
Z_6	-0,272(-0,641)	0,218(0,374)	-0,014(-0,017)	-0,438(-0,463)	-0,070(-0,062)
Z_7	-0,303(-0,712)	0,144(0,247)	-0,423(-0,509)	0,346(0,366)	0,095(0,084)
Z_8	0,107(0,253)	0,058(0,100)	-0,706(-0,849)	0,360(0,381)	-0,008(-0,007)
Z_9	-0,275(-0,647)	-0,397(-0,682)	0,072(0,086)	0,127(0,134)	0,014(0,013)
Z_{10}	-0,209(-0,492)	-0,442(-0,759)	0,073(0,088)	0,161(0,169)	-0,156(-0,139)
Z_{11}	-0,211(-0,497)	-0,399(-0,684)	-0,258(-0,310)	-0,310(-0,328)	-0,043(-0,038)
Z_{12}	-0,139(-0,327)	-0,445(-0,764)	-0,291(-0,350)	-0,294(-0,311)	-0,132(-0,117)
Z_{13}	0,017(0,041)	-0,310(-0,532)	0,370(0,445)	0,505(0,534)	-0,002(-0,002)

O *scree-plot* (Figura 15) e os critérios descritos na seção 3.4.1 indicaram que apenas as três primeiras componentes são suficientes para explicar a maior parte da variação total dos dados, 76,42% (Tabela 14), ou seja, elas podem substituir as variáveis originais em análises subseqüentes (se for de interesse do pesquisador). A redução do número de variáveis já era esperada, pelo fato de se ter variáveis muito correlacionadas (ver matriz **R**).

A primeira CP, apesar dos coeficientes serem menores que 0,50, está mais relacionada às características externas gerais dos animais, como o tamanho da concha (Y_2 , Y_3 e Y_5) e o peso do animal (Y_1). A segunda CP relaciona-se com a parte reprodutiva dos animais, isto é, com as

Figura 15 - *Scree-plot* das 13 CP's (a) e gráfico dos escores dos animais referente às duas primeiras CP's obtidas (b)

glândulas prostática (Y_9 e Y_{10}) e albumina (Y_{11} e Y_{12}). E a terceira CP, com a parte interna, referente à carne dos animais (Y_7 e Y_8). Se a quarta CP fosse de interesse para o pesquisador, ela estaria relacionada à fertilidade do animal, já que as variáveis percentagem do aparelho reprodutor (Y_{13}) e percentagem da concha (Y_6) apresentaram maiores coeficientes nesta componente.

Nas Figuras 15 (b), 16 e 17 percebe-se a existência de animais (pontos) em todos os quadrantes (ou octantes para a figura tri-dimensional) como, por exemplo, animais com alto valor de carne (componente C_3), alto valor de tamanho da concha (componente C_1) e baixo valor referente as glândulas de reprodução (componente C_2).

Figura 16 - Gráficos dos escores dos animais referente a primeira e terceira CP's (a) e a segunda e terceira CP's (b)

Figura 17 - Gráfico dos escores dos 106 animais para as três primeiras CP's vistos por ângulos diferentes

Tabela 16 - Coeficientes de ponderação das variáveis e suas correlações (valores entre parênteses) com as últimas componentes principais

Variável	Componentes Principais			
	C_{13}	C_{12}	C_{11}	C_{10}
Z_1	-0,089(-0,0027)	-0,143(-0,0091)	-0,711(-0,0719)	-0,112(-0,0176)
Z_2	-0,720(-0,0221)	-0,109(-0,0069)	0,145(0,0147)	-0,064(-0,0101)
Z_3	0,572(0,0175)	-0,032(-0,0021)	-0,048(-0,0049)	-0,066(-0,0104)
Z_4	0,347(0,0106)	0,096(0,0061)	-0,056(-0,0057)	-0,025(-0,0040)
Z_5	0,080(0,0025)	0,108(0,0069)	0,493(0,0499)	-0,563(-0,0885)
Z_6	-0,030(-0,0009)	-0,056(-0,0036)	-0,190(-0,0193)	0,335(0,0528)
Z_7	0,053(0,0016)	0,044(0,0028)	0,312(0,0315)	0,607(0,0955)
Z_8	-0,036(-0,0011)	-0,028(-0,0018)	-0,211(-0,0213)	-0,417(-0,0656)
Z_9	-0,100(-0,0031)	0,706(0,0449)	-0,107(-0,0108)	0,022(0,0034)
Z_{10}	0,074(0,0023)	-0,630(-0,0400)	0,089(0,0090)	0,017(0,0027)
Z_{11}	0,010(0,0003)	-0,161(-0,0103)	0,123(0,0125)	-0,070(-0,0109)
Z_{12}	-0,005(-0,0002)	0,133(0,0084)	-0,107(-0,0109)	0,018(0,0029)
Z_{13}	0,003(0,0001)	0,002(0,0001)	0,031(0,0031)	0,027(0,0042)

Variável	Componentes Principais			
	C_9	C_8	C_7	C_6
Z_1	-0,038(-0,0088)	0,456(0,1850)	-0,220(-0,1456)	0,011(0,0074)
Z_2	0,014(0,0033)	-0,461(-0,1873)	-0,168(-0,1109)	-0,005(-0,0036)
Z_3	0,029(0,0066)	-0,502(-0,2037)	-0,102(-0,0671)	-0,011(-0,0074)
Z_4	0,093(0,0215)	-0,178(-0,0723)	-0,159(-0,1047)	0,004(-0,0029)
Z_5	-0,200(-0,0462)	0,370(0,1503)	0,166(0,1095)	0,107(0,0725)
Z_6	0,051(0,0117)	-0,113(-0,0459)	0,654(0,4320)	0,293(0,1980)
Z_7	-0,077(-0,0178)	0,303(0,1229)	-0,106(-0,0701)	0,019(0,0131)
Z_8	0,058(0,0134)	-0,130(-0,0526)	0,327(0,2159)	0,100(0,0677)
Z_9	0,177(0,0409)	0,007(0,0027)	0,264(0,1746)	-0,351(-0,2368)
Z_{10}	-0,141(-0,0324)	0,025(0,0103)	0,320(0,2116)	-0,414(-0,2792)
Z_{11}	0,677(0,1561)	0,088(0,0355)	-0,256(-0,1690)	0,232(0,1567)
Z_{12}	-0,653(-0,1505)	-0,167(-0,0677)	-0,238(-0,1572)	0,232(0,1564)
Z_{13}	0,002(0,0005)	0,026(0,0104)	0,142(0,0936)	0,700(0,4724)

Com o intuito de descartar as variáveis respostas redundantes, utilizou-se o método relatado por Mardia et al. (1997, apud BARBOSA et al., 2006). Conforme a Tabela 14, foram oito as componentes com autovalores menores que 0,7.

Baseando-se nos maiores valores dos coeficientes, em valor absoluto, nas últimas CP's, as variáveis passíveis de descarte, em ordem de menor importância para explicar a variação total, foram: comprimento da concha (Y_2), peso glândula prostática (Y_9), peso do animal (Y_1), peso da carne (Y_7), peso da glândula albumina (Y_{11}), largura da concha (Y_3), percentual de concha (Y_6) e percentual do aparelho reprodutor (Y_{13}).

Segundo Pacheco et al. (1998, apud CAETANO, 2005), em virtude da variação na forma das conchas, havendo animais mais alongados e outros mais arredondados com relação comprimento $\times$ largura de concha, animais mais arredondados apresentam uma maior taxa de postura. Esta afirmação pode ajudar a explicar a correlação significativa ($r = 0,5$) entre as variáveis largura da concha (Y_3) e peso da glândula prostática/útero (Y_9). O comportamento

arredondado poderia ser melhor explicado por uma outra variável, por exemplo, a razão corporal ($\frac{\text{comprimento}}{\text{largura}} = \frac{Y_2}{Y_3}$), o que talvez justifique o descarte das variáveis comprimento da concha (Y_2) e largura da concha (Y_3).

O comprimento da concha (Y_2), apresentou várias correlações elevadas com outras variáveis e foi a variável menos importante para explicar a variação total deste ensaio. A variável percentagem do aparelho reprodutor (Y_{13}), apesar de não apresentar altas correlações com outras variáveis, também não foi considerada importante neste estudo. Tais fatos indicam que a informação que é expressa por estas variáveis, já estejam contempladas em outras variáveis ou em combinações delas.

Todas as combinações, duas a duas, entre variáveis: peso glândula prostática (Y_9), percentagem da glândula prostática (Y_{10}), peso da glândula albumina (Y_{11}) e percentagem da glândula albumina (Y_{12}) apresentaram correlações significativamente relevantes, possivelmente porque as glândulas são responsáveis pela reprodução do animal, dependendo uma da outra. Dentre estas, foram selecionadas como mais importantes a percentagem de glândula prostática (Y_{10}) e percentagem da glândula albumina (Y_{12}).

O peso da concha somado ao peso da carne é aproximadamente o peso do animal. Por sua vez, o peso da concha esta diretamente correlacionado com o comprimento ($r = 0,87$) e a largura da concha ($r = 0,77$). A variável peso do animal (Y_1) apresentou correlações elevadas com as variáveis comprimento (Y_2 , $r = 0,92$), largura (Y_3 , $r = 0,84$) e peso da concha (Y_5 , $r = 0,92$), além da variável peso da carne (Y_7 , $r = 0,73$), o que justifica ter sido selecionada para descarte.

O espaço interno da concha destes animais é o limitante do seu crescimento. Desta forma, compreende-se as altas correlações da variável peso da carne (Y_7) com as variáveis peso do animal (Y_1), peso da concha (Y_5), comprimento (Y_2) e largura (Y_3) da concha. A percentagem de carne (Y_8) apresentou correlações menores que 0,5 com as outras variáveis, indicando uma correlação fraca e por este motivo tenha sido considerada mais importante para o estudo do que a variável peso da carne (Y_7).

Já para a variável peso da concha (Y_5) ocorreu o contrário, sendo considerada mais importante que a percentagem da concha (Y_6); ambas são altamente correlacionadas ($r = 0,81$) e também apresentam altas correlações com as variáveis peso do animal (Y_1), largura (Y_3) e comprimento (Y_2) da concha.

As variáveis a serem mantidas em futuras análises, segundo a ACP foram: razão corporal (Y_4); peso da concha (Y_5); percentual de carne (Y_8); percentual de glândula prostática

(Y_{10}); e percentual de glândula albumina (Y_{12}).

Em resumo, aplicando a ACP em um conjunto de dados é possível selecionar as variáveis mais influentes ou mais informativas em um experimento. Contudo, como se pode observar, algumas variáveis importantes na análise são relações entre variáveis que foram consideradas “pouco importantes” e devem ser descartadas. Concluir que uma característica é descartável não implica em não observá-la em experimentos futuros. Indica que ela deva ser observada para provavelmente compor outras características mais importantes, mas que não deve ser utilizada na análise.

4 CONCLUSÕES

Diante do estudo realizado, pôde-se concluir que:

- O uso da técnica multivariada de análise de agrupamentos pode auxiliar bastante o pesquisador na construção de blocos de animais, baseando-se em informações de mais de uma característica. Na decisão pela melhor solução, recomenda-se que o pesquisador avalie a qualidade dos agrupamentos obtidos, compare as variâncias internas dos blocos e a variância total. Para maior segurança, deve-se utilizar diferentes métodos de agrupamento e um conjunto de animais numeroso, visto que as soluções resultantes nem sempre apresentam blocos de mesmo tamanho.
- A técnica de ACP pode indicar ao pesquisador quais variáveis são mais importantes na explicação do fenômeno observado, sendo possível eliminar de futuras análises, as características redundantes, sem perda demasiada de informação. Nesses casos, sugere-se que o pesquisador observe todas as possíveis características do fenômeno em estudo e calcule as relações entre as variáveis utilizando esta técnica multivariada. As características consideradas pouco importantes pela ACP não precisam ser utilizadas nas análises posteriores.
- A utilização da MANOVA em substituição a p -ANOVA's apresenta como vantagens importantes, a utilização de um nível de significância conjunto nas comparações de médias das diversas características simultaneamente e o aproveitamento na análise das correlações entre as variáveis respostas.
- A experiência do pesquisador na aplicação das técnicas multivariadas é de extrema importância, e não pode ser desprezada em momento algum.
- No presente trabalho, o *software* R mostrou-se uma ferramenta poderosa para a aplicação das técnicas de análises multivariadas, principalmente por seus recursos gráficos. Entretanto, o seu uso requer que o pesquisador tenha um bom conhecimento teórico dos conceitos envolvidos.

Como continuidade deste estudo, seria interessante divulgar outras técnicas multivariadas ao meio agropecuário, o que possivelmente levaria os pesquisadores um novo pensar dos seus experimentos, e também a algumas melhoras nos resultados, pois a natureza é multivariada e não univariada. Em relação ao *software* R, poderia-se criar um único pacote multivariado para

facilitar a utilização do *software* na aplicação das técnicas multivariadas, pois vários pacotes são necessários e alguns geralmente, fazem as mesmas tarefas com algumas diferenças, dificultando a escolha.

REFERÊNCIAS

- ABREU, V.M.N.; SILVA, M. de A. e; CRUZ, C.D.; FIGUEIREDO, É.A.P. de; ABREU, P.G. de. Desempenho e Predição de Híbridos e Análise de Agrupamento de Características de Matrizes de Frangos de Corte. **Revista Brasileira de Zootecnia**, Viçosa, v. 31, n. 2, p. 617-626, 2002.
- ADLER, D.; MURDOCH, D. **The rgl Package - 3D visualization device system (OpenGL)**. R package version 0.74, 51p., 2007. Disponível em: <<http://rgl.neoscientists.org>>. Acesso em: 06 jun. 2007.
- ALEIXO, S.S.; SOUZA, J.G. de; FERRAUDO, A.S. Técnicas de Análise Multivariada na Determinação de Grupos Homogêneos de Produtores de Leite. **Revista Brasileira de Zootecnia**, Viçosa, v. 36, n. 6, p. 2168-2175, dez. 2007.
- BARBIN, D. **Planejamento e Análise Estatística de Experimentos Agronômicos**. Arapongas: Midas, 2003. 194p.
- BARBOSA, L.; LOPES, P.S.; REGAZZI, A. J.; GUIMARÃES, S.E.F.; TORRES, R. de A. Avaliação de característica de carcaça de suínos utilizando-se a análise dos componentes principais. **Revista Brasileira de Zootecnia**, Viçosa, v. 34, n. 6, supl., p. 2209-2217, 2005a.
- BARBOSA, L.; LOPES, P.S.; REGAZZI, A.J.; GUIMARÃES, S.E.F.; TORRES, R. de A. Seleção de variáveis de desempenho de suínos por meio da análise de componentes principais. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 57, n. 6, p. 805-810, 2005b.
- BARBOSA, L.; LOPES, P.S.; REGAZZI, A.J.; GUIMARÃES, S.E.F.; TORRES, R. de A. Estudo da associação entre características de desempenho e de carcaça de suínos por meio de Correlação Canônica. **Revista Brasileira de Zootecnia**, Viçosa, v. 34, n. 6, supl., p. 2218-2224, 2005c.
- BARBOSA, L.; LOPES, P.S.; REGAZZI, A.J.; GUIMARÃES, S.E.F.; TORRES, R. de A. Avaliação de características de qualidade da carne de suínos por meio de componentes principais. **Revista Brasileira de Zootecnia**, Viçosa, v. 35, n. 4 (supl.), p. 1639-1645, 2006.
- BARROSO, L.P.; ARTES, R. Análise multivariada. In: SEAGRO - SIMPÓSIO DE ESTATÍSTICA APLICADA A EXPERIMENTAÇÃO AGRONÔMICA, 10., RBRAS - REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 48., 2003, Lavras. **Minicurso...** Lavras: UFLA, 2003. 156p.
- CAETANO, F.A.M. **Estudo comparativo do aparelho reprodutor do molusco *Achatina fulica* criado em cativeiro e asselvajado**. 2005. 61 p. Dissertação (Reprodução animal) - Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, São Paulo, 2005.
- CARRER, C. da C. **Caracterização e diferenciação regional da pecuária de corte no Brasil no fim do século: gênese, modernização e a reestruturação produtiva e mercadológica**. 2000, 268p. Tese (Doutorado) - Universidade de Campinas, Campinas, 2000.
- CARVALHO, M.S.; STRUCHINER, C.J. Análise de correspondência: Uma Aplicação do método à avaliação de serviços de Vacinação. **Caderno de Saúde Pública**, Rio de Janeiro, v. 8, n. 3, p. 287-301, set. 1992.
- CHINELATTO NETO, A.; CASTRO, G.P.C.; LIMA, J.E. de. Uso de análise estatística multivariada para tipificação de produtores de leite de Minas Gerais. **Organizações Rurais e Agroindustriais**, Lavras, v. 7, n. 1, p. 114-121, 2005.

COOPER, J.C.B. Factor analysis: an overview. **The American Statistician**. Alexandria, v. 37, n. 2, p. 141-147, maio 1983.

CRUZ, C.D. **Aplicação de algumas técnicas multivariadas no melhoramento de plantas**. 1990, 188p. Tese (Doutorado em Genética e Melhoramento de Plantas)- Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba.

CUADRAS, C.M. **Models Estadístics Multivariants**. Apostila. Barcelona, 2006. 249p.

CZERMAINSKI, A.B.C. Análise de Correspondência. In: SEMINÁRIO APRESENTADO NA DISCIPLINA ANÁLISE MULTIVARIADA, ESALQ/USP. 2004, Piracicaba. **Seminário...** Piracicaba, 2004. 17p.

DAHER, R.F.; MORAES, C.F.; CRUZ, C.D. et al. Seleção de caracteres morfológicos discriminantes em capim-elefante (*Pennisetum purpureum* schum). **Revista Brasileira de Zootecnia**, Viçosa, v. 26, p. 247-254, 1997.

DALGAARD, P. **Introductory Statistics with R**. Statistics and Computing. New York-USA: Springer, 267p., 2002.

DAOYU, Z.; LAWES, G.S. Manova and discriminant analyses of phenotypic data as a guide for parent selection in kiwifruit (*Actinidia deliciosa*) breeding. **Euphytica**, Netherlands, v. 114, p. 151-157, 2000.

DEMÉTRIO, C.G.B. **Análise multidimensional para dados de cana-de-açúcar**. 1985, 144p. Tese (Estatística e Experimentação Agrônômica) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1985.

DESTEFANIS, G.; BARGE, M.T.; BRUGIAPAGLIA, A. et al. The use of principal component analysis (PCA) to characterize beef. **Meat Science**, Savoy, Illinois, v. 56, p. 255-259, 2000.

EVERITT, B.S. **An R and S-Plus® Companion to Multivariate Analysis**. New York: Springer, 2005. 221p.

FAGUNDES, A.C.A. **Influência da temperatura ambiente na eficiência de utilização da energia da dieta e níveis séricos de T3, T4 e cortisol em suínos**. 1999, 54p. Tese (Doutorado em Zootecnia - Produção Animal) - Universidade Estadual Paulista “Júlio de Mesquita Filho”, Jaboticabal, 1999.

FERNANDES, T.A.G.; LIMA, J.E. Uso de análise multivariada para identificação de sistemas de produção. **Pesquisa Agropecuária Brasileira**, Brasília, v. 26, n. 10, p. 1823-1836, out. 1991.

FERREIRA, C.A.; FERREIRA, R.L.C.; SANTOS, D.C. dos; SANTOS, M.V.F. dos; SILVA, J.A.A. da; LIRA, M. de A.; MOLICA, S.G. Utilização de Técnicas Multivariadas na Avaliação da Divergência Genética entre Clones de Palma Forrageira (*Opuntia ficus-indica* Mill.). **Revista Brasileira de Zootecnia**, Viçosa, v. 32, n. 6, supl.1, p. 1560-1568, 2003.

FERREIRA, D.F. **Análise Multivariada**. Apostila. Lavras, 1996. 394p.

FIGUEIRA, M.M.C. **Identificação de outliers**. MILLENIUM, n.º 12 - Out/1998. Disponível em: <<http://www.ipv.pt/millennium/arq12.htm>>. Acesso em: 03 ago. 2007.

FONSECA, R.; PIRES, A.V.; LOPES, P.S.; TORRES, R.A., EUCLYDES, R.F. Estudo da divergência genética entre raças suínas utilizando técnicas de análise multivariada. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 52, n. 4, p. 403-409, ago. 2000.

FOX, J. **The car Package - Companion to Applied Regression**. R package version 1.2-2, 103p., 2007. Disponível em: <<http://www.r-project.org>, <http://socserv.socsci.mcmaster.ca/jfox/>>. Acesso em: 21 abr. 2007.

GENZ, A.; BRETZ, F.; HOTHORN, T. **The mvtnorm Package - Multivariate Normal and T Distribution**. R package version 0.7-5, 10p., 2006. Disponível em: <<http://cran.r-project.org>>. Acesso em: 26 abr. 2007.

GROSSMAN, G.D.; NICKERSON, D.M.; FREEMAN, M.C. Principal component analyses of assemblage structure data: utility of tests based on eigenvalues. **Ecological Society of America**, Washington, v. 72, n. 1, p. 341-347, fev. 1991.

HAIR JR., J.F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **Análise Multivariada de Dados**. Tradução de A.S. Sant'anna e A. Cloves Neto, 5. ed. Porto Alegre: Bookman, 2006. 593 p.

HARDOON, D.R.; SZEDMAK, S.; SHAW-TAYLOR, J. **Canonical correlation analysis; An overview with application to learning methods**. Department of Computer Science Royal Holloway, University of London, England. Technical Report CSD-TR-03-02. 39p., may 2003. Disponível em: <http://eprints.ecs.soton.ac.uk/9225/01/tech_report03.pdf>. Acesso em: 05 out. 2007.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 5. ed. New Jersey: Prentice Hall, 2002. 767p.

KHATTREE, R.; NAIK, D.N. **Applied Multivariate Statistics With SAS Software**. 2. ed. Cary: SAS Publishing, John Wiley, 1999. 338p.

LAFORGE, H. **Analyse Multivariée**. Canadá: Études Vivantes, 1981. 317p.

LANG, D.T.; SWAYNE, D. **The rggobi Package - Interface between R and GGobi**. Version 2.1.6, 69p., 2007. Disponível em: <<http://cran.r-project.org>>. Acesso em: 15 maio 2007.

LEDO, C.A. da S.; FERREIRA, D.F.; RAMALHO, M.A.P. Análise de variância multivariada para os cruzamentos dialéticos. **Ciências Agrotécnicas**, Lavras, v. 27, n. 6, p. 1214-1221, dez. 2003.

LIBERATO, J.R.; VALE, F.X.R.; CRUZ, C.D. Técnicas estatísticas de análise multivariada e a necessidade de o fitopatologista conhecê-las. **Fitopatologia Brasileira**, Brasília, v.24, p. 5-8, 1999.

LIGGES, U. **The scatterplot3d Package - 3D Scatter Plot**. Version 0.3-25, 6p., 2007. Disponível em: <<http://cran.r-project.org>>. Acesso em: 10 ago. 2007.

LIMA, E.L. **Elementos de topologia geral**: Topologia. 2. ed. Rio de Janeiro: IMPA, 1969. 299p.

LUCAS, A. **The amap Package - Another Multidimensional Analysis Package**. Version 0.7-3.mvtnorm, 23p., 2007. Disponível em: <<http://mulcyber.toulouse.inra.fr/projects/amap/>>. Acesso em: 14 jun. 2007.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M. **The cluster Package - Cluster Analysis Basics and Extensions**. Version 1.11.8, unpublished, 73p., 2007. Disponível em: <<http://cran.r-project.org>>. Acesso em: 13 jun. 2007.

MAGALHÃES, M.N.; LIMA, A.C.P. de. **Noções de probabilidade e estatística**. 6. ed. São Paulo: EDUSP, 2004. 392p.

MALUCHE-BARETTA, C.R.D.; AMARANTE, C.V.T. do; KLAUBERG FILHO, O. Análise Multivariada de Atributos do Solo em Sistemas Convencional e Orgânico de Produção de Maçãs. **Pesquisa Agropecuária Brasileira**, Brasília, v. 41, n. 10, p. 1531-1539, out. 2006.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: UFMG, 2005. 295p.

MOITA NETO, J.M. **Estatística multivariada - Uma visão didática-metodológica**. 2004. Disponível em: <http://criticanarede.com/cien_estatistica.html>. Acesso em: 25 abr. 2007.

MORRISON, D.F. **Multivariate statistical methods**. 2.ed. New York: McGraw-Hill Company, 1976. 415p.

MOURA, W. de M.; CASALI, V.W.D.; CRUZ, C.D.; LIMA, P.C. de Divergência genética em linhagens de pimentão em relação à eficiência nutricional de fósforo. **Pesquisa Agropecuária Brasileira**, Brasília, v. 34, n. 2, p. 217-224, fev. 1999.

MURRELL, P. **R Graphics - Computer Science and Data Analysis Series**. Londres: Chapman & Hall/CRC. 2006, 301p.

OKSANEN, J.; KINDT, R.; LEGENDRE, P.; O'HARA, B.; STEVENS, M.H.H. **The vegan Package - Community Ecology Package**. R package Version 1.8-7, 136p., 2007. Disponível em: <<http://cran.r-project.org>>. Acesso em: 01 ago. 2007.

PENNY, K.I. Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance. **Applied Statistics**, Londres, v. 45, n. 1, p. 73-81, 1996.

PINTO, L.F.B.; ALMEIDA, F.Q. de A.; AZEVEDO, P.C.N. de; QUIRINO, C.R.; CABRAL, G.C.; SANTOS, E.M. Análise Multivariada das Medidas Morfométricas de Potros da Raça Mangalarga Marchador: Análise Fatorial. **Revista Brasileira de Zootecnia**, Viçosa, v. 34, n. 2, p. 613-626, 2005a.

PINTO, L.F.B.; ALMEIDA, F.Q. de; QUIRINO, C.R.; CABRAL, G.C.; AZEVEDO, P.C.N. de; SANTOS, E.M. Análise Multivariada das Medidas Morfométricas de Potros da Raça Mangalarga Marchador: Análise Discriminante. **Revista Brasileira de Zootecnia**, Viçosa, v. 34, n. 2, p. 600-612, 2005b.

POWER, J.P.; CAMPBELL, B.M.S. Cluster Analysis and the Classification of Medieval Domesne-Farming Systems. **Transactions of the Institute of British Geographers**, New Series. Great Britain, v. 17, n. 2, p. 227-245, 1992.

R Development Core Team. R Foundation for Statistical Computing. R: A language and environment for statistical computing., Vienna, Austria. ISBN 3-900051-07-0, URL. Disponível em: <<http://www.R-project.org>>. Acesso em: 10 fev. 2007.

REGAZZI, A.J. **Análise Multivariada**. Universidade Federal de Viçosa, Viçosa. (INF-766) notas de aula, 2002.

REIS, E. **Estatística Multivariada Aplicada**. Lisboa: Edições Silabo, 1997. 343p.

RENCHE, A.C.; SCHAALJE G.B. **Linear Models in Statistics**. 2. ed. New Jersey: John Wiley, 2008. 672p.

RODRIGUES, L.R.F.; ANDO, A. Melhoramento Genético Vegetal. Uso da Sensitividade à Radiação Gama na Discriminação de Variedades de Arroz-de-Sequeiro dos Grupos Índica e Japônica. **Bragantia**, Campinas, v. 62, n. 2, p. 179-188, 2003.

RODRIGUES, L.S.; ANTUNES, I.F.; TEIXEIRA, M.G.; SILVA, J.B. da. Divergência genética entre cultivares locais e cultivares melhoradas de feijão. **Pesquisa agropecuária brasileira**, Brasília, v. 37, n. 9, p. 1275-1284, set. 2002.

ROLIM, G. de S.; CAMARGO, M.B.P. de; LANIA, D.G.; MORAES, J.F.L. de. Classificação Climática de Köppen e de Thornthwaite e sua Aplicabilidade na determinação de Zonas Agroclimáticas para o Estado de São Paulo. **Bragantia**, Campinas, v. 66, n. 4, p. 711-720, 2007.

SANTOS, C.G.P.; MATO, L.F.; CLENNELL, B. Análise Discriminante aplicada à caracterização do reservatório do Campo de Namorado (Bacia de Campos/RJ, Brasil). In: CONGRESSO BRASILEIRO DE P&D EM PETRÓLEO E GÁS, 2., 2003, Rio de Janeiro. **Anais eletrônicos...** Rio de Janeiro, 2003, 6p. Disponível em: <<http://www.portalabpg.org.br/PDPetro/2/2083.pdf>>. Acesso em: 16 jun. 2007.

SILVA, I.C. da; MESSAGE, D.; CRUZ, C.D.; SILVA, M.V.G.B. da. Aplicação de Análises Multivariadas para Determinação da Casta de Abelhas *Apis mellifera* L. (Africanizadas), Obtidas em Laboratório. **Revista Brasileira de Zootecnia**. Viçosa, v. 34, n. 2, p.635-640, 2005.

SOUZA, G.N., BRITO, J.R.F., BASTOS, R.R.; RUBIALE, L. Avaliação de associações e similaridades em epidemiologia veterinária por meio da análise de correspondência. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 54, n. 5, p. 539-542, out. 2002.

STEARNS, T. M.; BEEVER, J.E.; SOUTHEY, B.R.; ELLIS, M.; MCKEITH, F.K.; RODRIGUEZ-ZAS, S.L. Evaluation of approaches to detect quantitative trait loci for growth, carcass, and meat quality on swine chromosomes 2, 6, 13, and 18. II. Multivariate and principal component analysis. **American Society of Animal Science**, Savoy, v. 83, p. 2471-2481, 2005.

TEIXEIRA, L. L. **O uso de técnicas de estatística multivariada no prognóstico de desistência de alunos em IES privadas: um estudo de caso na cidade de Foz do Iguaçu-PR**. 2006. 79p. Dissertação (Métodos Numéricos em Engenharia – Programação Matemática, Setores de Tecnologia e Ciências Exatas) - Universidade Federal do Paraná, Curitiba, 2006.

TIKU, M.L.; BALAKRISHNAN, N. Testing the equality of variance-covariance matrices the robust way. **Communications in Statistics**, Canadá, v. 14, n. 12, p. 3033-3051, jan. 1985.

TOLEDO, L.G. de; NICOLELLA, G. Índice de Qualidade de Água em Microbacia Sob Uso Agrícola e Urbano. **Scientia Agricola**, Piracicaba, v. 59, n. 1, p. 181-186, jan./mar. 2002.

TORRES FILHO, R. de A.; EUCLYDES, R.F. ; TORRES, R. de A.; LOPES, P.S.; BREDAS, F.C. Estudo da divergência genética entre linhas de suínos utilizando técnicas de análise multivariada. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 57, n. 3, p. 390-395, 2005.

TRIVELLONI, C.A.P.; HOCHHEIM, N. Avaliação de Imóveis com Técnicas de Análise Multivariada. In: COBRAC - CONGRESSO BRASILEIRO DE CADASTRO TÉCNICO MULTIFINALITÁRIO, 98., out. 1998, Florianópolis. **Anais eletrônicos...** Florianópolis: UFSC, 1998. Disponível em: <<http://geodesia.ufsc.br/Geodesia-online/arquivo/cobrac98/106/106.HTM>>. Acesso em: 19 abr. 2007.

TRUGILHO, P.F.; LIMA, J.T.; MORI, F.A. Correlação Canônica das Características Químicas e Físicas da Madeira de Clones de *Eucalyptus grandis* e *Eucalyptus saligna*. **CERNE**, Lavras, v. 9, n. 1, p. 066-080, 2003.

VAINIONPÄÄ, J.; KERVINEN, R.; PRADO, M. de; LAURILA, E.; KARI, M.; MUSTONEN, L.; AHVENAINEN, R. Exploration of storage and process tolerance of different potato cultivars using principal component and canonical correlation analyses. **Journal of Food Engineering**, Amsterdam, v. 44, p. 47-61, 2000.

VESSONI, F. **Correlação Canônica**. MV2 Sistemas de Informação, 1998. 13p. Disponível em: <<http://www.mv2.com.br/artigos.htm>>. Acesso em: 09 mar. 2007.

VIANA, C.F.A.; SILVA, M.A.; PIRES, A.V.; LOPES, P.S.; TORRES, R.A. Análise de Variância Multivariada na análise de grupos genéticos de matrizes de frangos de corte. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 53, n. 4, p. 1-6., 2001.

ZÜGE, M.; CHAVES NETO, A. Utilização de Métodos Estatísticos Multivariados na Avaliação do Desempenho Empresarial. **Revista Paranaense de Desenvolvimento**, Curitiba, n. 97, p. 101-112, dez. 1999.

BIBLIOGRAFIA CONSULTADA

ASSIS, G.M.L. da; EUCLYDES, R.F.; CRUZ, C.D.; VALLE, C.B. do. Discriminação de Espécies de *Brachiaria* Baseada em Diferentes Grupos de Caracteres Morfológicos. **Revista Brasileira de Zootecnia**, Viçosa, v. 32, n. 3, p. 576-584, 2003.

BAIRD, D. Exploratory Factor Analysis, Instruments and Logic of Discovery. **British Journal for the Philosophy Science**, Great Britain, v. 38, p. 319-337, 1987.

BELLAVER, C.; GUIDONI, A.L.; BRUM, P.A.R. de; ROSA, P.S. Estimativas das exigências de lisina e de energia metabolizável em frangos de corte de 1 a 21 dias de idade, utilizando-se uma variável multivariada canônica. **Revista Brasileira de Zootecnia**, Viçosa, v. 31, n. 1, p. 71-78, 2002.

BRITO, L.T. de L.; SILVA, A. de S.; SRINIVASAN, V.S.; GALVÃO, C. de O.; GHEYI, H.R. Uso de Análise Multivariada na Classificação das Fontes Hídricas Subterrâneas da Bacia Hidrográfica do Salitre. **Engenharia Agrícola**, Jaboticabal, v. 26, n. 1, p. 36-44, jan./abr. 2006.

CARNEIRO, P.L.S., FONSECA, R., PIRES, A.V.; TORRES FILHO, R.A.; TORRES, R.A.; Peixoto, J.O.; LOPES, P. S.; EUCLYDES, R.F. Estudo da divergência genética entre linha-gens de matrizes de frangos de corte por meio de análise multivariada. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 54, n. 1, p. 75-83, 2002.

CÔRTEZ, C.; DAMASCENO, J.C.; FUKUMOTO, N.M.; SAKAGUTI, E.S.; SANTOS, G.T. dos; AL-CALDE, C.R. Potencial Discriminatório dos N-alcanos em Plantas Forrageiras Tropicais por Análises Multivariadas. **Revista Brasileira de Zootecnia**, Viçosa, v. 34, n. 4, p. 1079-1087, 2005.

DAHER, R.F.; VÁZQUEZ, H.M.; PEREIRA, A.V.; FERNANDES, A.M. Introdução e Avaliação de Clones de Capim-Elefante (*Pennisetum purpureum* Schum.) em Campos dos Goytacazes, RJ. **Revista Brasileira de Zootecnia**, Viçosa, v. 29, n. 5, p. 1296-1301, 2000.

FONSECA, R.; TORRES FILHO, R.A.; TORRES, R.A.; PEIXOTO, J.O.; PIRES, A.V.; CARNEIRO, P.L.S.; SOUZA, G.H.; BUENO, R.S.; LOPES, P.S.; EUCLYDES, R.F. Avaliação de frangos de corte utilizando técnicas de análise multivariada: I - Características de carcaça. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 54, n. 5, p. 525-529, 2002.

FOX, J. **An R and S-Plus Companion to Applied Regression**. California: Sage Publications, 2002. 312p.

GIANNOTTI, J.D.G.; PACKER, I.U.; MERCADANTE, M.E.Z.; LIMA, C.G. de. Análise de Agrupamento para Implementação da Meta-Análise em Estimativas de Herdabilidade para Características de Crescimento em Bovinos de Corte. **Revista Brasileira de Zootecnia**, Viçosa, v. 34, n. 4, p. 1165-1172, 2005.

GIMENES, F.M.P.; GIMENES, R.M.T.; OPAZO, M.A.U. Os processos de integração econômica sob a ótica da análise estatística de agrupamento. **Revista FAE**, Curitiba, V. 7, n. 2, p. 19-32, jul./dez. 2004.

LIGGES, U.; MÄCHLER, M. Scatterplot3d - an R Package for Visualizing Multivariate Data. **Journal of Statistical Software**, Alexandria, v. 8, n. 11, p. 1-20, 2003.

OKSANEN, J. **Multivariate Analysis of Ecological Communities in R**: vegan tutorial. Version 1.8-7, 2007. 39p. Disponível em: <<http://cran.r-project.org>>. Acesso em: 08 jul. 2007.

OLIVEIRA, V.R.; CASALI, V.W.D.; PEREIRA, P.R.G.; CRUZ, C.D.; PIRES, N. de M. Tolerância de Genótipos de Pimentão ao Baixo Teor de Fósforo no Solo. **Bragantia**, Campinas, v. 58, n. 1, p. 125-139, 1999.

PINTO, L.F.B.; ALMEIDA, F.Q. de; QUIRINO, C.R.; AZEVEDO, P.C.N. de; CABRAL, G.C.; CORASSA, A. Análise multivariada das medidas morfométricas de potros da raça Mangalarga Marchador: análise de componentes principais. **Revista Brasileira de Zootecnia**. Viçosa, v. 34, n. 2, p. 589-599, abr. 2005.

PIRES, A.V., CARNEIRO, P.L.S., TORRES FILHO, R.A.; FONSECA, R.; TORRES, R.A.; EUCLYDES, R.F.; LOPES, P.S.; BARBOSA, L. Estudo da divergência genética entre seis linhas de aves Legorne utilizando técnicas de análise multivariada. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 54, n. 3, p. 314-319, 2002.

SHARMA, S. **Applied Multivariate Techniques**. United States: John Wiley & Sons, 1996. 493p.

SILVA, H.; RESENDE, A.; ROSA, C.; SIMÕES, R. Dinâmica agropecuária e urbanização: Uma análise multivariada para Minas Gerais 1995-2000. In: ENCONTRO NACIONAL DE ECONOMIA, 33., 2005, Rio de Janeiro. **Anais eletrônicos...** Disponível em: <<http://econpapers.repec.org/paper/anpen2005/140.htm>>. Acesso em: 10 out. 2007.

WERNER, M. **Identification of multivariate outliers in large data sets**. 2003. 241p. Thesis (Philosophy Applied Mathematics) - University of Colorado at Denver in partial fulfillment, Colorado, 2003.

APÊNDICES

APÊNDICE A - Algumas métricas para variáveis quantitativas

Tabela 17 - Algumas distâncias para variáveis quantitativas

Distância	Fórmula
Minkowski	$d_{ij} = \left(\sum_{k=1}^p y_{ik} - y_{jk} ^m \right)^{1/m}$
City-Block ($m = 1$)	$d_{ij} = \sum_{k=1}^p y_{ik} - y_{jk} $
Euclidiana ($m = 2$)	$d_{ij} = \sqrt{\sum_{k=1}^p (y_{ik} - y_{jk})^2}$
Mahalanobis	$d_{ij}^2 = (\mathbf{y}_i - \mathbf{y}_j)' \mathbf{S}^{-1} (\mathbf{y}_i - \mathbf{y}_j)$
Chebishev	$d_{ij} = \max_k y_{ik} - y_{jk} $
Canberra	$d_{ij} = \sum_{k=1}^p \frac{ y_{ik} - y_{jk} }{(y_{ik} + y_{jk})}$
Coeficiente de Czekanowski	$d_{ij} = 1 - \frac{\sum_{k=1}^p \min(y_{ik}, y_{jk})}{\sum_{k=1}^p (y_{ik} + y_{jk})}$

$\mathbf{y}_i = (y_{i1}, \dots, y_{ip})'$ é o vetor de observações do indivíduo i , $i = 1, 2, \dots, n$, no qual y_{ik} representa o valor assumido pela variável k , no indivíduo i ($k = 1, 2, \dots, p$).

d_{ij} é a distância do elemento i ao j .

$\mathbf{S}$ é a matriz de variâncias e covariâncias amostrais, comum a todas as unidades.

O valor m é referente a métrica de Minkowsky. Para $m = 1$, obtém-se a distância City Block (Manhattan ou quarteirão). Se $m = 2$, a distância de Minkowski coincide com a distância Euclidiana. A métrica de Minkowsky é menos afetada pela presença de valores discrepantes na amostra do que a distância Euclidiana.

APÊNDICE B - Estatísticas Descritivas: Experimento I

Tabela 18 - Suínos: Estatísticas descritivas das variáveis originais

Estatística Descritiva	Variável	
	Peso Inicial (kg)	Idade Inicial (dias)
Mínimo	23,00	70.00
1º quartil	24,00	72.75
Média	25,02	73.60
Mediana	25,00	73.00
3º quartil	26,00	75.00
Máximo	27,00	78.00
Desvio Padrão	1,06	1.78

APÊNDICE C - Estatísticas Descritivas: Experimento II e III

Tabela 19 - Escargot: Estatísticas descritivas das variáveis originais

Estatística Descritiva	Variável						
	Peso (g)	Comprimento (mm)	Largura (mm)	Razão corporal	Concha (g)	Concha (%)	Carne (g)
Mínimo	18,80	53,00	30,00	1,53	2,94	11,31	2,98
1º quantil	25,98	60,00	35,00	1,68	5,52	19,77	5,52
Média	38,03	68,94	38,77	1,78	10,82	27,49	6,80
Mediana	32,05	67,00	38,00	1,77	8,00	23,34	7,57
3º quantil	42,86	76,00	41,25	1,85	14,16	34,04	9,05
Máximo	97,94	97,00	56,00	2,23	47,00	161,61	17,80
Desvio Padrão	17,78	10,42	5,24	0,12	8,02	15,18	3,03

Tabela 20 - Escargot: continuação

Estatística Descritiva	Variável					
	Carne (%)	GlProst (g)	GlProst (%)	GlAlbum (g)	GlAlbum (%)	ApReprod (%)
Mínimo	11,57	0,04	0,11	0,01	0,03	0,17
1º quantil	17,07	0,55	1,53	0,16	0,52	2,09
Média	19,64	0,85	2,23	0,46	1,24	4,76
Mediana	20,83	0,92	2,33	0,94	2,29	7,198
3º quantil	24,27	1,10	2,93	0,87	2,98	8,33
Máximo	43,57	6,30	17,50	5,78	12,27	51,88
Desvio Padrão	5,56	0,72	1,80	1,29	2,60	7,93

ANEXOS

ANEXO A - Padronização das Variáveis

A padronização consiste em expressar as variáveis respostas em termos de unidades de desvio padrão. A operacionalização consiste em subtrair a média e dividir pelo desvio padrão, de modo que tenham média *zero* e variância igual a *um*, isto é:

$$Z_j = \frac{(Y_j - \bar{y}_j)}{\sqrt{\sigma_{jj}}},$$

onde Z_j é a “nova” variável padronizada; Y_j é a “antiga” variável resposta; $\bar{y}_j$ e σ_{jj} são, respectivamente, a média e a variância da variável resposta Y_j , $j = 1, 2, \dots, p$.

Esta padronização corresponde exatamente a aplicar a técnica “desejada” a matriz de correlações $\mathbf{R}$, e não mais na matriz de variâncias e covariâncias $\mathbf{S}$.

ANEXO B - Matriz Positiva Definida

O estudo da variação e das inter-relações em dados multivariados é baseado, muitas vezes, em *distâncias* e na suposição que os dados têm distribuição normal multivariada. Distâncias ao quadrado e a densidade normal multivariada podem ser expressa em termos de produtos de matrizes chamadas *formas quadráticas*.

Definição 4.1 : Quando uma matriz simétrica $\mathbf{A}$ de ordem k , tem propriedade de $\mathbf{y}'\mathbf{A}\mathbf{y} > 0$ para qualquer $\mathbf{y}' = [y_1, y_2, \dots, y_k]$ (vetores de observações $\mathbf{y}$), com exceção de $\mathbf{y} \neq 0$, então a forma quadrática $\mathbf{y}'\mathbf{A}\mathbf{y}$ é dita *positiva definida* (pd) e a matriz $\mathbf{A}$ é dita *positiva definida*. Se $\mathbf{y}'\mathbf{A}\mathbf{y} \geq 0$, então $\mathbf{A}$ é dita *positiva semidefinida* (psd).

Alguns resultados importantes sobre matriz pd (RENCER; SCHAAALJE, 2008):

Teorema 4.1 : Se $\mathbf{A}$ é pd, então todos os elementos a_{ii} de sua diagonal são positivos.

Teorema 4.2 : Uma matriz simétrica $\mathbf{A}$ é pd, se e somente se existe uma matriz não-singular $\mathbf{P}$ tal que $\mathbf{A} = \mathbf{P}'\mathbf{P}$.

Uma método de fatorar uma matriz pd $\mathbf{A}$ em um produto $\mathbf{P}'\mathbf{P}$ é chamado de *Decomposição de Cholesky*. Nesta decomposição, a matriz $\mathbf{A}$ pode ser fatorada de modo único em $\mathbf{A} = \mathbf{T}'\mathbf{T}$, onde $\mathbf{T}$ é uma matriz não-singular e triangular superior.

Corolário 4.1 : Uma matriz pd é não-singular.

Teorema 4.3 : Se $\mathbf{A}$ é pd, então $\mathbf{A}^{-1}$ é pd.

Teorema 4.4 : Se Σ é p.d., de tal modo que Σ^{-1} existe, então:

$$\Sigma e = \lambda e \quad \Rightarrow \quad \Sigma^{-1}e = \left[\frac{1}{\lambda} e \right]$$

tal que (λ, e) é um par autovalor-autovetor (Anexo I) de Σ correspondente ao par $(\frac{1}{\lambda}, e)$ de Σ^{-1} , que também é p.d.

ANEXO C - Distância de Mahalanobis

A Distância de Mahalanobis foi criada em 1936 por Prasanth Chandra Mahalanobis (Figura 18), sendo baseada na correlação entre as variáveis. Sua escala é invariante, isto é, não depende da escala de medida. Na estatística multivariada, esta distância é muito rica em informações. É usada em análises de agrupamento e outras técnicas de classificação, como também na distribuição de Hottelling's T^2 , que é usada em testes multivariados. Além disso, é utilizada para detectar outliers, especialmente no desenvolvimento de modelos de regressão linear (*Chi-Square plot* ou *Q-Q plot*).

Figura 18 - Prasanth Chandra Mahalanobis (Índia - 1893★ ; 1972†)

Formalmente, a distância de Mahalanobis para um grupo de valores com média $\boldsymbol{\mu}' = [\mu_1 \ \mu_2 \ \dots \ \mu_p]$ e Σ a matriz de variâncias e covariâncias de um vetor multivariado $\mathbf{y}' = [y_1 \ y_2 \ \dots \ y_p]$ é definido como:

$$D^2(\mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

e representa o quadrado da distância generaliza de $\mathbf{y}$ a $\boldsymbol{\mu}$. Observe que esta distância também está presente no expoente da função densidade de probabilidade da distribuição normal multivariada. O conjunto de valores de $\mathbf{y}$ tais que:

$$(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{c} ,$$

onde $\mathbf{c}$ é um vetor constante, representa a superfície de um elipsóide centrado em $\boldsymbol{\mu}$ (REIS, 1997). Quanto mais “esticada” esta elipsóide, maior é a correlação entre as variáveis.

Se $\mathbf{y} \sim N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, com determinante da matriz $\boldsymbol{\Sigma}$ maior que *zero*, então o quadrado da distância generaliza de $\mathbf{y}$ a $\boldsymbol{\mu}$ terá distribuição $\chi^2_{(p)}$, onde p é o número de variáveis respostas.

A distância de Mahalanobis também pode ser definida como uma medida de dissimilaridade entre dois vetores aleatórios $\mathbf{y}$ e $\mathbf{x}$ de mesma distribuição com matriz de covariância Σ :

$$d(\mathbf{y}, \mathbf{x}) = \sqrt{(\mathbf{y} - \mathbf{x})' \Sigma^{-1} (\mathbf{y} - \mathbf{x})} .$$

Se $\Sigma = \mathbf{I}$, onde $\mathbf{I}$ é igual a matriz identidade, a distância de Mahalanobis se reduz a distância Euclidiana. Observe que a Distância de Mahalanobis difere da distância Euclidiana porque leva-se em consideração a correlação do conjunto de dados. Se Σ é uma matriz diagonal, então a distância medida resultará na chamada distância Euclidiana normalizada:

$$d(\mathbf{y}, \mathbf{x}) = \sqrt{\sum_{i=1}^p \frac{(\mathbf{y}_i - \mathbf{x}_i)^2}{\sigma_i^2}}$$

onde σ_i é o desvio padrão de y_i sobre o conjunto amostral.

ANEXO D - Boxplot Bivariado

O *boxplot bivariado* é análogo ao familiar *boxplot* univariado, porém para duas dimensões (EVERITT, 2005). Este gráfico consiste na construção de um gráfico de dispersão do tipo Y_j versus $Y_{j'}$, $j \neq j'$, sobreposto a intervalos de confiança em formato de elipse.

De acordo com a teoria sobre distribuição normal multivariada, todos os pares de variáveis $(Y_j, Y_{j'})$ devem ter uma distribuição normal bivariada e, portanto, os gráficos de dispersão devem indicar a forma de uma elipse. Por este motivo, é justificado os intervalos de confiança traçados no gráfico de dispersão. Ele também auxilia na identificação de dados discrepantes. Abaixo, apresenta-se o programa, no R, para confecção deste gráfico.

```
#####
biweight<- function(a,const1=9,const2=36,err=0.0001) {
#a é uma matriz de dados com duas colunas
#const1=common tuning constant
#const2=bivariate tuning constant
#err= critério de convergência.
#
  x<-a[,1];  y<-a[,2]
  n<-length(x);  mx<-median(x);  my<-median(y)
  madx<-median(abs(x-mx));  mady<-median(abs(y-my))
  if(madx != 0) { ux<-(x-mx)/(const1*madx)
    ux1<-ux[abs(ux)<1]
    tx<-mx+(sum((x[abs(ux)<1]-mx)*(1-ux1*ux1)^2)/
      sum((1-ux1^2)^2))
    sx<- sqrt(n)*sqrt(sum((x[abs(ux)<1]-mx)^2*
      (1-ux1*ux1)^4))/abs(sum((1-ux1*ux1)*
      (1-5*ux1*ux1)))
  }
  else { tx<-mx;  sx<-sum(abs(x-mx))/n }
  if(mady != 0) { uy<-(y-my)/(const1*mady)
    uy1<-uy[abs(uy)<1]
    ty<-my+(sum((y[abs(uy)<1]-my)*(1-uy1*uy1)^2)/
      sum((1-uy1^2)^2))
    sy<- sqrt(n)*sqrt(sum((y[abs(uy)<1]-my)^2*
      (1-uy1*uy1)^4))/abs(sum((1-uy1*uy1)*
      (1-5*uy1*uy1)))
  }
  else { ty<-my;  sy<-sum(abs(y-my))/n }
  z1<-(y-ty)/sy+(x-tx)/sx;  z2<-(y-ty)/sy-(x-tx)/sx
  mz1<-median(z1);  mz2<-median(z2)
  madz1<-median(abs(z1-mz1));  madz2<-median(abs(z2-mz2))
  if(madz1 !=0) { uz1<-(z1-mz1)/(const1*madz1);  uz11<-uz1[abs(uz1)<1]
    tz1<-mz1+(sum((z1[abs(uz1)<1]-mz1)*(1-uz11*uz11)^2)/
      sum((1-uz11^2)^2))
    sz1<- sqrt(n)*sqrt(sum((z1[abs(uz1)<1]-mz1)^2*
```

```

      (1-uz11*uz11)^4))/abs(sum((1-uz11*uz11)*
      (1-5*uz11*uz11)))
    }
    else { tz1<-mz1;   sz1<-sum(abs(z1-mz1))/n }
if(mady != 0) { uz2<-(z2-mz2)/(const1*madz2)
  uz21<-uz2[abs(uz2)<1]
  tz2<-mz2+(sum((z2[abs(uz2)<1]-mz2)*(1-uz21*uz21)^2)/
    sum((1-uz21^2)^2))
  sz2<- sqrt(n)*sqrt(sum((z2[abs(uz2)<1]-mz2)^2*
    (1-uz21*uz21)^4))/abs(sum((1-uz21*uz21)*
    (1-5*uz21*uz21)))
  }
  else { tz2<-mz2;   sz2<-sum(abs(z2-mz2))/n }
esq<-((z1-tz1)/sz1)^2+((z2-tz2)/sz2)^2
w<-numeric(length=n);   c2<-const2
for(i in 1:10) {
  w[esq<const2]<-(1-esq[esq<const2]/const2)^2
  w[esq>=const2]<-0;   l<-length(w[w==0])
  if(l<0.5*n) break;   else const2<-2*const2
}
tx<-sum(w*x)/sum(w);   sx<-sqrt(sum(w*(x-tx)^2)/sum(w))
ty<-sum(w*y)/sum(w);   sy<-sqrt(sum(w*(y-ty)^2)/sum(w))
r<-sum(w*(x-tx)*(y-ty))/(sx*sy*sum(w))
const2<-c2;   wold<-w
for(i in 1:100) {
  z1<-((y-ty)/sy+(x-tx)/sx)/sqrt(2*(1+r))
  z2<-((y-ty)/sy-(x-tx)/sx)/sqrt(2*(1+r))
  esq<-z1*z1+z2*z2
  for(j in 1:10) {
    w[esq<const2]<-(1-esq[esq<const2]/const2)^2
    w[esq>=const2]<-0;   l<-length(w[w==0])
    if(l<0.5*n) break;   else const2<-2*const2
  }
  tx<-sum(w*x)/sum(w);   sx<-sqrt(sum(w*(x-tx)^2)/sum(w))
  ty<-sum(w*y)/sum(w);   sy<-sqrt(sum(w*(y-ty)^2)/sum(w))
  r<-sum(w*(x-tx)*(y-ty))/(sx*sy*sum(w));
  term<-sum((w-wold)^2)/(sum(w)/n)^2
  if(term<-err) break;   else { wold<-w;   const2<-c2 }
}
param<-c(tx,ty,sx,sy,r);   param
}

#
####==#####==#####==#####==#####==#####==#####==#####==#####
###                               Boxplot Bivariado                               ###
####==#####==#####==#####==#####==#####==#####==#####==#####
bivbox<- function(a, d = 7, mtitle = "Bivariate Boxplot",
  method = "robust",xlab="Y1",ylab="Y2")
{
  #a é uma matriz de dados
  #d é uma constante(usualmente 7)
  #
  p <- length(a[1, ])

```

```

if(method == "robust") {
  param <- biweight(a[,1:2])
  m1 <- param[1]; m2 <- param[2]
  s1 <- param[3]; s2 <- param[4]; r <- param[5]
}
else { m1 <- mean(a[,1]); m2 <- mean(a[,2])
  s1 <- sqrt(var(a[,1])); s2 <- sqrt(var(a[,2]))
  r <- cor(a[,1:2])[1,2]
}
x <- (a[,1]-m1)/s1 ; y <- (a[,2]-m2)/s2
e <- sqrt((x*x + y*y - 2*r*x*y)/(1 - r*r))
e2 <- e*e; em <- median(e)
emax <- max(e[e2 < d*em*em])
r1 <- em*sqrt((1+r)/2); r2 <- em*sqrt((1-r)/2)
theta <- ((2*pi)/360)*seq(0,360,3)
xp <- m1+(r1*cos(theta) + r2*sin(theta))*s1
yp <- m2+(r1*cos(theta) - r2*sin(theta))*s2
r1 <- emax*sqrt((1+r)/2); r2 <- emax*sqrt((1-r)/2)
theta <- ((2*pi)/360)*seq(0,360,3)
xpp <- m1+(r1*cos(theta) + r2*sin(theta))*s1
ypp <- m2+(r1*cos(theta) - r2*sin(theta))*s2
maxxl <- max(xpp); minxl <- min(xpp)
maxyl <- max(ypp); minyl <- min(ypp)
b1 <- (r*s2)/s1; a1 <- m2-b1*m1
y1 <- a1 + b1*minxl; y2 <- a1 + b1*maxxl
b2 <- (r*s1)/s2; a2 <- m1 - b2*m2
x1 <- a2 + b2*minyl; x2 <- a2 + b2*maxyl
maxx <- max(c(a[,1], xp, xpp, x1, x2))
minx <- min(c(a[,1], xp, xpp, x1, x2))
maxy <- max(c(a[,2], yp, ypp, y1, y2))
miny <- min(c(a[,2], yp, ypp, y1, y2))
plot(a[,1],a[,2], xlim=c(minx,maxx), ylim=c(miny,maxy), xlab=xlab,
  ylab=ylab, lwd=2, pch=1)
lines(xp, yp, lwd=2); lines(xpp, ypp, lty=2, lwd=2)
segments(minxl, y1, maxxl, y2, lty=3, lwd=2)
segments(x1, minyl, x2, maxyl, lty=4, lwd=2)
}
####FIM!

```

ANEXO E - Q-Q plot (*Chi-Square plot*)

Este gráfico, também conhecido como gráfico qui-quadrado (ou *Chi-Square plot*), auxilia na verificação de multinormalidade e na identificação de possíveis *outliers*. Este consiste em:

- Calcular a Distância de Mahalanobis: $d_i^2 = (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}})$, $i = 1, 2, \dots, n$, para todos os elementos da amostra e ordenar estes valores em ordem crescente, isto é, $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$, onde $d_{(i)}^2$ representa a i -ésima estatística de ordem;
- Fazer o gráfico dos pares $(d_{(i)}^2; \chi_p^2((i - \frac{1}{2})/n))$, onde $\chi_p^2((i - \frac{1}{2})/n)$ representa a ordenada do percentil de ordem $100((i - \frac{1}{2})/n)$ da distribuição qui-quadrado com p graus de liberdade, isto é,

$$P \left[\chi_p^2 \leq \chi_p^2 \left(\frac{(i - \frac{1}{2})}{n} \right) \right] = \frac{(i - \frac{1}{2})}{n}.$$

Quando a normalidade p -variada é coerente com os dados amostrais, este gráfico deve resultar em algo próximo a uma reta (Figura 19). Curvas diferentes da reta indicam falta de normalidade.

Figura 19 - Ilustração do gráfico Chi-Square plot

A justificativa deste gráfico é que quando o número de elementos amostrais n é grande, a variável d_i^2 tem aproximadamente uma distribuição qui-quadrado com p graus de liberdade, onde $\mathbf{y}_i$ representa os valores observados das p -variáveis do i -ésimo elemento amostral, $\bar{\mathbf{y}}$ é o vetor de médias amostrais e $\mathbf{S}$ é a matriz de covariâncias amostral de ordem p . Assim, pode-se comparar os

valores das distâncias d_i^2 com as respectivas ordenadas dos percentis da distribuição qui-quadrado (MINGOTI, 2005). O programa em código R para a obtenção do gráfico *Q-Q plot* é apresentado a seguir:

```
#####
###                               Gráfico Q-Q plot (Chi-Square plot)                               ###
#####

chisplot <- function(x) {
  if (!is.matrix(x)) stop("x is not a matrix")
# dimensões
  n <- nrow(x)
  p <- ncol(x)
#
  xbar <- apply(x, 2, mean)
  S <- var(x)
  S <- solve(S)
#
  index <- (1:n)/(n+1)
  xcent <- t(t(x) - xbar)
  di <- apply(xcent, 1, function(x,S) x %*% S %*% x,S)
#
  quant <- qchisq(index,p)
  qqplot(quant, sort(di), ylab = "Distâncias Ordenadas - dj^2",
  xlab = "Quantil qui-quadrado", lwd=2, pch=1)
  abline(0,1, col="gray")
  title("Q-Q plot")
  locator()
}
#####FIM!
```


ANEXO F - Teste de Esfericidade de Bartlett para matriz de correlações

A aplicação das técnicas multivariadas muitas vezes exige que as p -variáveis observadas sejam correlacionadas de algum modo. Quando a distribuição do vetor aleatório $\mathbf{y}$ é normal p -variada, é possível construir testes de hipótese para avaliação da matriz de correlações. Um destes testes é o de Bartlett (1954), ou teste da independência das variáveis respostas. Geometricamente, este teste serve para verificar se os contornos da densidade são esferas quando $\Sigma = \sigma^2 \mathbf{I}$.

Suponha que se tenha uma amostra aleatória de tamanho n do vetor aleatório $\mathbf{y}$. Considere as seguintes hipóteses:

$$H_0 : \mathbf{R} = \mathbf{I};$$

$$H_1 : \mathbf{R} \neq \mathbf{I}.$$

onde $\mathbf{I}$ é a matriz identidade de ordem p e $\mathbf{R}$ é a matriz de correlações teórica das p -variáveis. A hipótese nula, H_0 , equivale a dizer que as p variáveis são independentes ou que a matriz de variâncias e covariâncias $\mathbf{S}$ de $\mathbf{y}$ é diagonal²⁷; e H_1 que as p variáveis não são independentes. A estatística de teste χ é definida por:

$$\chi = - \left[n - \frac{1}{6} (2p + 11) \right] \left[\sum_{j=1}^p \ln(\lambda_j) \right]$$

onde $\ln(\cdot)$ denota a função logaritmo neperiano e λ_j são os autovalores (Anexo I) da matriz de correlações amostrais $\mathbf{R}$, $j = 1, 2, \dots, p$. Sob H_0 e n grande, a estatística χ tem uma distribuição aproximadamente qui-quadrado com $\frac{1}{2}p(p-1)$ graus de liberdade. Assim, rejeita-se H_0 se o valor observado de χ for maior ou igual ao valor crítico da distribuição qui-quadrado para o nível de significância escolhido para o teste (MINGOTI, 2005).

```
#####
###                               Teste de Esfericidade de Bartlett                               ###
#####
# lembre-se que a matriz de cov/var correta é aquela dos resíduos!!!
## Teste de Esfericidade -> S=I
#                               H0: R = I
#
# mod<- manova(y~trat) # modelo multivariado ajustado
#   E<- mod$res; E      # Matriz de Resíduos
#   S<- cov(E); S
#   R<- cor(E); R
#   n= nrow(E); n
```

²⁷Quando têm-se distribuição normal, correlação igual a zero implica em independência entre as variáveis.

```

    p= ncol(E); p
#
    auto<-eigen(R); auto
    a<- log(auto$values); a
    UM<- cbind(rep(1,p)); UM
    lambda=a%*%UM; lambda
#
    Xcalc= -(n-(1/6)*(2*p+1))*(lambda); Tcalc # valor
    gl=(1/2)*p*(p-1); gl
    X_95<-qchisq(0.95,gl); T_95 # valor
    X_99<-qchisq(0.99,gl); T_99 # valor
# Se Xcalc>X rejeita-se H0 ao nível de signif. 5%, e conclui-se que
# R é diferente da matriz I (identidade), isto é, as variáveis
# respostas não são independentes.
####==###==###==###==###==###==###==###==###==###==###==FIM!

```

ANEXO G - Teste de Box para Igualdade de Matrizes de Variâncias e Covariâncias

A pressuposição das matrizes de variâncias e covariâncias dos grupos serem iguais é a de maior importância na realização da MANOVA. O teste clássico, o teste da razão de verossimilhança, para testar a igualdade de matrizes de variâncias e covariâncias foi desenvolvido por Box (1950), sendo uma generalização do teste univariado de igualdade de variâncias de Bartlett (1947). As hipóteses a testar são:

$$H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$$

$$H_1 : \Sigma_l \neq \Sigma_k, \text{ para algum } l \neq k \text{ } (l, k = 1, 2, \dots, g).$$

Sejam n a dimensão total da amostra, $v_k = n_k - 1$ os graus de liberdade associados a cada grupo, $\mathbf{S}_k$ a matriz de variâncias e covariâncias do grupo k , $k = 1, 2, \dots, g$, e $\mathbf{S} = \frac{\mathbf{W}}{n - g}$ a matriz de variâncias e covariâncias total, onde g é o número de populações, e $\mathbf{W}$ é a matriz de SQPC do resíduo. O teste é definido do seguinte modo:

$$M = (n - g) \ln |\mathbf{S}| - \sum_{k=1}^g v_k \ln |\mathbf{S}_k|$$

Box sugeriu duas aproximações para o teste:

- i. **A aproximação à χ^2 :** é indicada quando as dimensões dos grupos são superiores a 20, o número de variáveis e de grupos inferior a seis. Esta aproximação é dada por:

$$M C \sim \chi^2_{\left(\frac{1}{2}p(p+1)(g-1)\right)}$$

onde

$$C = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \left(\sum_{k=1}^g \frac{1}{v_k} - \frac{1}{n-g} \right)$$

- ii. **A aproximação à F :** indicada em todas as outras situações. Esta é formulada por:

$$\frac{M \left(1 - a_1 - \frac{v}{v_0} \right)}{v} \sim F_{(v, v_0)}$$

sendo

$$a_1 = 1 - C;$$

$$a_2 = \frac{(p-1)(p+2)}{6(g-1)} \left(\sum_{k=1}^g \frac{1}{v_k^2} - \frac{1}{(n-g)^2} \right);$$

$$v = \frac{p(p+2)(g-1)}{2}; \text{ e}$$

$$v_0 = \frac{v+2}{a_2 - a_1^2}.$$

Um nível de significância de 0,01 ou menos é usado como ajuste para a sensibilidade da estatística. (HAIR JR. et al., 2006, p.275).

O teste de Box é muito sensível a violações do pressuposto de populações normais multivariadas. Assim, a hipótese de igualdade de matrizes pode ser rejeitada apenas por se violar o pressuposto de normalidade e não por se tratarem de matrizes significativamente diferentes. Quando esta situação se verificar, deverá proceder-se a transformações dos dados de modo a que se verifique o pressuposto anterior e só depois testar a igualdade de matrizes de variâncias e covariâncias (REIS, 1997). Na situação multivariada, o tamanho crítico das discrepâncias das matrizes de covariâncias deve depender do número de variáveis p . Logo, uma transformação de variáveis pode melhorar a situação quando as variâncias marginais são pouco diferentes (JOHNSON; WICHERN, 2002).

Tiku e Balakrishnan (1985) propuseram um teste mais robusto para testar à igualdade de duas matrizes de variância e covariância, sendo menos sensível a pressuposição de multinormalidade. O teste é uma modificação da estatística T_D^2 proposta por Tiku e Singh (1982) para testar a igualdade de vetores de média de duas populações multivariadas.

Entretanto, para n_1 e n_2 grandes, não é preciso se preocupar com o problema de matrizes de covariâncias diferentes. Para esta situação, foi criada uma estatística para testar a hipótese nula de igualdade de vetores de grupos (detalhes em JOHNSON; WICHERN, 2002).

As linhas de comando, no R, para aplicar o teste de Box para igualdade de matrizes de variâncias e covariâncias são:

```
#####
###          Teste de Box (Igualdade de matrizes var/cov)          ###
#####
### H0: S1=S2=...=Sg
#
#      E                                # matriz de erros
S1<- cov(E[a:b, ]); S1              # matriz de covariâncias do grupo 1, a:b é
                                     # o intervalo dos elementos do grupo 1.
S2<- cov(E[c:d, ]); S2              # matriz de covariâncias do grupo 2.
                                     #... continua até o grupo g.
                                     #Sg<- cov(E[x:z, ]) # matriz de covariâncias do grupo g.
#
n1<- nrow(E[a:b,]); n1              # n. de elementos do grupo 1.
n2<- nrow(E[c:d,]); n2              # n. de elementos do grupo 2.
```

```

# ... continua até grupo g.
#ng<- nrow(E[x:z,]) # n. de elementos do grupo g.
k<- g; k # número de grupos
#
# Para ilustrar, considere g=2:
Sp<- ((n1-1)*S1+(n2-1)*S2)/(n-k); Sp # matriz de covariâncias totais.
M<- (n-k)*log(det(Sp))-((n1-1)*log(det(S1))+(n2-1)*log(det(S2)))
C<- 1-((2*(p^2)+3*p-1)/(6*(p+1)*(k-1)))*((1)/(n1-1)+(1)/(n2-1)-(1)/(n-k))
#
####=====####
#### aproximação à F:
a1<- 1-C
a2<- ((p-1)*(p+2))/(6*(k-1))*((1)/((n1-1)^2)+(1)/((n2-1)^2)-(1)/((n-k)^2))
v<- (p*(p+1)*(k-1))/2; v
v0<- (v+2)/(a2-((a1)^2)); v0
#
f_calc<- M*(1-a1-((v)/(v0)))/(v); f_calc
f_90<- qf(0.90,v,v0); f_90
f_95<- qf(0.95,v,v0); f_95
f_99<- qf(0.99,v,v0); f_99
f_999<- qf(0.999,v,v0); f_999
# Se f_calc > f_95 rejeita-se H0 ao nível de signif. 5%.
#
####=====####
#### aproximação à X^2:
X2_calc<- M*C; X2_calc
X2_90<- qchisq(0.90,v); X2_90
X2_95<- qchisq(0.95,v); X2_95
X2_99<- qchisq(0.99,v); X2_99
# Se X2_calc > X2_95 rejeita-se H0 ao nível de signif. 5%.
#
####=====####=====####=====####=====####=====####=====FIM!

```

ANEXO H - Distribuições de Wishart e de T^2 de Hotelling

H.1 - Distribuição de Wishart

A distribuição amostral da matriz de variâncias e covariâncias amostrais é chamada de *distribuição de Wishart* e é definida como a soma de produtos de vetores independentes com distribuição normal multivariada. Especificamente:

$$\begin{aligned} W_p(.|\Sigma) &= \text{distribuição de Wishart com } p \text{ graus de liberdade} \\ &= \text{distribuição de } \sum_{j=1}^p \mathbf{Z}_j \mathbf{Z}_j', \text{ onde } \mathbf{Z}_j \sim_{i.i.d.} N_p(\mathbf{0}, \Sigma). \end{aligned} \quad (1)$$

A distribuição de Wishart na realidade é uma generalização multivariada da distribuição qui-quadrado.

H.2 - Distribuição T^2 de Hotelling

Essa distribuição foi inicialmente proposta por Harold Hotelling, tendo como objetivo generalizar a distribuição *t-Student* no campo multivariado.

Teorema 4.5 : Se $\mathbf{u} \sim N_p(\mathbf{0}_{(p \times 1)}; \Sigma)$,

$\mathbf{V} \sim W_p(\Sigma; n)$, e $\mathbf{u}$ e $\mathbf{V}$ forem independentes, então:

$$n \mathbf{u}' \mathbf{V}^{-1} \mathbf{u} \sim T_{(p,n)}^2,$$

sendo $T_{(p,n)}^2$ a distribuição de Hotelling (distribuição escalar) com p e n graus de liberdade.

Não há necessidade de tabelar a distribuição de Hotelling porque ela facilmente se aproxima à distribuição F de Snedecor (REIS, 1997) pela expressão:

$$\frac{(n-p+1)}{p n} T_{(p,n)}^2 = F_{(p, n-p+1)}$$

Quando $p=1$ esta distribuição coincide com o quadrado da distribuição *t-Student* com n graus de liberdade, ou seja: se $p = 1 \implies T_{(p,n)}^2 = t_n^2$.

A aplicação deste teorema a uma população normal multivariada é de grande importância para os testes de hipóteses (TEIXEIRA, 2006) e regiões de confiança multivariadas.

ANEXO I - Autovalores e Autovetores de uma Matriz

Os autovalores e autovetores de uma matriz estão presentes em muitas das técnicas multivariadas, pois acabam sendo “soluções” de otimização de vários métodos multivariados, devido as imposições de cada técnica. Eles são obtidos da seguinte maneira:

Para qualquer matriz quadrada $\mathbf{A}$ de dimensão p , um escalar λ e um vetor não-nulo $\mathbf{w}$ podem ser encontrados, de tal forma que:

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{w}, \quad (2)$$

λ é chamado de *autovalor* (ou *valor característico*) de $\mathbf{A}$ e $\mathbf{w}$, de *autovetor* (ou *vetor característico*) de $\mathbf{A}$. Para encontrar λ e $\mathbf{w}$ para uma matriz $\mathbf{A}$, escreve-se a equação (2) como:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{w} = 0. \quad (3)$$

Como a matriz $(\mathbf{A} - \lambda\mathbf{I})\mathbf{w}$ é uma combinação das colunas de $\mathbf{A} - \lambda\mathbf{I}$, então as colunas desta matriz são linearmente dependentes. Assim, a matriz quadrada $\mathbf{A} - \lambda\mathbf{I}$ é singular, e o seu determinante é igual a zero. Logo, pode-se resolver (3) para λ usando:

$$|\mathbf{A} - \lambda\mathbf{I}| = 0$$

que é conhecida como *equação característica*.

Se $\mathbf{A}$ é $(n \times n)$ a equação característica terá n raízes, isto é, $\mathbf{A}$ terá n autovalores $\lambda_1, \lambda_2, \dots, \lambda_n$. Depois de calcular os autovalores, os autovetores poderão ser encontrados resolvendo a equação (3) (RENCER; SCHAALJE, 2008). Portanto, a matriz $\mathbf{A}$ tem n pares $(\lambda_1, \mathbf{w}_1), \dots, (\lambda_n, \mathbf{w}_n)$, onde $\mathbf{w}_i' \mathbf{w}_j = 0$, isto é, os autovetores são ortogonais. Se $\mathbf{w}_i' \mathbf{w}_j = 1$, diz-se que os autovetores são normalizados.

O comando *eigen*, no software R, calcula os autovalores e autovetores de uma determinada matriz.

ANEXO J - Decomposição Espectral e Decomposição em Valores Singulares

A *Decomposição Espectral* (DE) de uma matriz $\mathbf{A}$ simétrica de ordem p , é dada por:

$$\mathbf{A}_{(p \times p)} = \sum_{j=1}^p \lambda_j \mathbf{e}_j \mathbf{e}_j' = \mathbf{C} \mathbf{D} \mathbf{C}' ,$$

sendo $\mathbf{D}$ a matriz diagonal de ordem p , com os autovalores (Anexo I) de $\mathbf{A}$, $\lambda_1, \lambda_2, \dots, \lambda_p$; e $\mathbf{C} = (\mathbf{e}_1 \dots \mathbf{e}_p)$ a matriz de ordem p , com os respectivos autovetores de $\mathbf{A}$.

Se uma matriz $\mathbf{A}$ é positiva definida (pd) é possível encontrar uma *raiz quadrada* de $\mathbf{A}$, denotada por $\mathbf{A}^{1/2}$. Desde que os autovalores são positivos, pode-se substituir a raiz quadrada $\sqrt{\lambda_j}$ em λ_j na decomposição espectral de $\mathbf{A}$ para obter

$$\mathbf{A}^{1/2} = \mathbf{C} \mathbf{D}^{1/2} \mathbf{C}' ,$$

onde $\mathbf{D}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_p})$. A matriz $\mathbf{A}^{1/2}$ é simétrica e tem a propriedade $\mathbf{A}^{1/2} \mathbf{A}^{1/2} = (\mathbf{A}^{1/2})^2 = \mathbf{A}$ (RENCHEER; SCHAALJE, 2008).

A *Decomposição em Valores Singulares* (DVS) é a generalização da DE, isto é, qualquer matriz $\mathbf{A}_{(n \times p)}$ pode ser decomposta como:

$$\mathbf{A}_{(n \times p)} = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{v}_i' = \mathbf{U} \mathbf{D} \mathbf{V}' ,$$

para $i = 1, 2, \dots, r$, e $r = \min(n \times p)$, onde:

$\mathbf{U}$ e $\mathbf{V}$ são ortogonais;

$\mathbf{U}$ é a matriz dos autovetores de $\mathbf{A} \mathbf{A}'_{(n \times n)}$, $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$, associados aos $\lambda_i \neq 0$;

$\mathbf{V}$ é a matriz dos autovetores de $\mathbf{A}' \mathbf{A}_{(p \times p)}$, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r$, associados aos $\lambda_i \neq 0$; e

$\mathbf{D}$ é a matriz quadrada dos autovalores não nulos de $\mathbf{A} \mathbf{A}'_{(n \times n)}$ ou $\mathbf{A}' \mathbf{A}_{(p \times p)}$, ou seja, é diagonal com elementos reais e não negativos.

ANEXO K - Programa para a AA

```

#####
###           Análise de Agrupamentos - Métodos Hierárquicos           ###
#####

## Métrica Euclidiana - Distâncias (Medidas de Dissimilaridade)
  d<-dist(dados, method="euclidean"); # Matriz de Distâncias.
  round(d,2)
#
  require(cluster)                # Pacote para realizar a AA
  hc1 <- hclust(d, "centroid"); hc1 # Centróide.
  hc2 <- hclust(d, "single");  hc2  # Método do vizinho mais próximo.
  hc3 <- hclust(d, "complete"); hc3  # Método do vizinho mais distante.
  hc4 <- hclust(d, "average"); hc4  # Método da Ligação Média.
  hc5 <- hclust(d, "ward");     hc5  # Método de Ward.
#
#####
## Dendrograma: os elementos do agrupamento hc1.
  plot(hc1, hang=-1)
  y1<- rect.hclust(hc1, k=g, border="gray"); y1 # k=número de grupos.
#
#####
## correlação cofenética do agrupamento hc1.
  require(vegan)
  d.cofen1<-cophenetic(hc1); cor(d,d.cofen1)
#
## os elementos de cada grupo, obtido pelo hc1.
  c4<- cutree(hc1,k=2); c4          # k= número de grupos
  plot(dados, col=c4)
#
#####
###           Análise de Agrupamentos - Método não-hierárquico           ###
#####

### k-médias
  ANH<- kmeans(dados,g,nstart=1981); ANH # nstart = semente.
#
  plot(dados,col=ANH$cluster); points(ANH$centers,col=1:2,pch=8)
#
#####FIM!

```

ANEXO L - Programa para a MANOVA

```

#####
###           Análise de Variância Multivariada - MANOVA           ###
#####
### Lendo os Dados como 1 data.frame:
      dados<- read.table("dados.txt",head=T, dec=",")
      head(dados)           # para ler o cabeçalho.
      dim(dados)            # dimensão da matriz de dados.
      attach(dados)         # para chamar as variáveis pelo nome.
#
      Trat<- factor(trat); Trat  # o vetor de tratamentos 'trat' foi
                                # identificado como fator.
#
#####
### Ajustando o modelo multivariado:
      # y<- cbind(y1,y2,...,yp) # colocar as p variáveis respostas.
      mod1<- manova(y~Trat); mod1
      E<- mod1$res; E          # matriz de resíduos (erros)
#
#####
### MANOVA:                H0: não há efeito de trat.
#                           vs
#                           H1: existe efeito de trat.
#
      summary(mod1,test="Pillai")          # ou summary(mod1)
      summary(mod1,test="Wilks")
      summary(mod1,test="Hotelling")
      summary(mod1,test="Roy")
#
      B<- summary(mod1)$SS$Trat; B  # matriz B (SQPCHip ou SQPCTrat)
      W<- summary(mod1)$SS$Res; W  # matriz W (SQPCRes)
      SQPCTotal<- B+W; SQPCTotal
#
##### FIM!

```

ANEXO M - Programa para a ACP

```

#####
###           Análise de Componentes Principais - ACP           ###
#####

    dados                # matriz de dados (nxp)
    diag(cov(dados)); cor(dados)
#
### considere a ACP sobre a matriz de correlações R.
    summary(princomp(dados,cor=T)) # Porcentagem de variação total
                                   # explicada pela i-ésima CP
    require(amac)                  # chamando o pacote amap.
    cp<- acp(dados); cp$loadings
#
#####
### scree plot
    lambda<- eigen(cor(dados))$value
    plot(lambda,type="l",main="Scree plot",xlab="CP",ylab="autovalores")
    points(lambda,col='gray')
#
#####
## as primeiras CP
    require(rgl)                  # chamando o pacote rgl.
    plot3d(cp$score[,1],cp$score[,2],cp$score[,3],type="s",col="gray")
    plot(cp$score[,1],cp$score[,2])
#
#####
## variáveis padronizadas Z_i
    m<- colMeans(dados); m        # média das variáveis.
    s<- diag(cov(dados)); s        # variância das variáveis.
#
    z1<-(y1-m[1])/(sqrt(s[1]))     # primeira variável padronizada.
    z2<-(y2-m[2])/(sqrt(s[2]))     # segunda variável padronizada.
                                   # ... continua até a p-ésima variável.
                                   #zp<-(yp-m[p])/(sqrt(s[p]))
    lambda                        # var das componentes principais
    e<- eigen(cor(dados))$vector; e
#
#####
## correlação entre as variáveis padronizadas Z_i e as CP's
    a1<- e[,1]*sqrt(lambda[1])
    a2<- e[,2]*sqrt(lambda[2])
                                   # ... continua até a p-ésima variável.
                                   # ap<- e[,p]*sqrt(lambda[p])
# A<- rbind(a1,a2,...,ap)
#
##### FIM!

```