Universidade Técnica de Lisboa
Instituto Superior de Economia e Gestão

Identificação de Outliers: uma aplicação ao conjunto das maiores Empresas com actividade em Portugal

Maria Manuela Caria Figueira

Dissertação apresentada para obtenção do grau de Mestre em Matemática Aplicada à Economia e à Gestão

Junho / 1995

RESERVADO

I. S. E. G.
E. Biblioteca
1668-G. 42768

HD2356.P67
F54
1995

ERRATA

Página	*linha*	*onde se lê*	*deve ler-se*
1	10	outlier	*outlier*
13	11	translacção	translação
24	2	... alternativa $H_{j1,...,jp}$	... alternativa $\overline{H}_{j1,...,jp}$
24	8	$H_{j1,...,jp}$, ou seja, $H = \cup H_{j1,...,jp}$.	$\overline{H}_{j1,...,jp}$, ou seja, $\overline{H} = \cup \overline{H}_{j1,...,jp}$.
24	9	$L_{Hj1,...,jp}(x_1,...,x_n$	$L_{\overline{H}j1,...,jp}(x_1,...,x_n$
28	3	$T_3 = \max_j \left[\frac{x_{(n)} - \overline{x}}{s}; \ldots \right.$	$T_3 = \max \left[\frac{x_{(n)} - \overline{x}}{s}; \ldots \right.$
35	18	Barnett e Lewis(1980)	Barnett e Lewis(1978)
36	6 e 11	Barnett e Lewis(1980)	Barnett e Lewis(1978)
36	25	Lewis(1980)	Lewis(1978)
40	17	$0 \leq \Re i \leq 1$	$0 \leq \Re_i \leq 1$
43	23	$\sum_{i=1}^{n} c_i$	$\sum_{i=1}^{n} c_i = 1$
50	3	verificam seguinte	verificam a seguinte
89	15	a análise exploratória	Na análise exploratória
89	20	teste	O teste
90	28	os valor críticos	os valores críticos
91	17	Barnett e Lewis(1980)	Barnett e Lewis(1978)
92	6 e 22	T_7 e T_8	T_6 e T_7
92	7	T_7 e T_8	T_6
92	21	T_7	T_6
92	23	T_1	T_1 e T_6
95	1	duas primeiras ...	as duas primeiras
105	20	Neste caso, só a ordem...	Neste caso, não só a ordem

BIBLIOTECA

AGRADECIMENTOS

Desejo expressar o meu mais profundo agradecimento ao Professor Doutor António Luís Silvestre pela disponibilidade e pela efectiva e preciosa orientação.

Uma palavra de agradecimento, também, à família, amigos e colegas pela compreensão, apoio e estímulo durante o decurso deste trabalho.

ÍNDICE

I.S.E.G. BIBLIOTECA

Página

INTRODUÇÃO

Todo o investigador já deparou com um conjunto de dados em que algumas observações se afastam demasiado das restantes, parecendo que foram geradas por um mecanismo diferente. O estudo destas observações é importante dado que "uma das importantes etapas, em qualquer análise estatística de dados, é estudar a qualidade das observações..." Muñoz-Garcia *et al.*(1990).

As observações que apresentam um grande afastamento das restantes ou são inconsistentes com elas são habitualmente designadas por *outliers*. Pensamos ser necessário, desde já, encontrar uma definição do termo. A definição de outlier não é fácil como se pode verificar pelas definições dadas por alguns dos que mais contribuiram para o seu estudo:

"An observation with an abnormally large residual will be referred to as an *outlier*. Other terms in English are "wild", "straggler", "sport" and "maverick"; one may also speak of a "discordant", "anomalous" or "aberrant" observation." [Anscombe(1960) pág. 125]

"An outlying observation, or "outlier", is one that appears to deviate markedly from other members of the sample in which it occurs." [Grubbs(1969) pág. 1]

"Observations that, in the opinion of the investigator, stand apart from the bulk of the data have been called "outliers", "discordant observations", "rogue values", "contaminants", "surprising values", "mavericks", and "dirty data" ... investigators are rightly concerned when such observations occur." [Beckman e Cook(1983) pág.120]

"Outliers are observations that do not follow the pattern of the majority of the data." [Rousseuw e Zomeren(1990) pág. 633]

"An outlier is an observation which being atypical and/or erroneous deviates decidedly from the general behaviour of experimental data with respect to the criteria which is to be analysed on it." [Muñoz-Garcia *et al.*(1990) pág. 217]

"We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data." [Barnett e Lewis(1994) pág. 7]

Das definições anteriores pode-se concluir que um *outlier* é caracterizado pela sua relação com as restantes observações que fazem parte da amostra. O seu distanciamento em relação a essas observações é fundamental para se fazer a sua caracterização. Estas observações são também designadas por observações "anormais", contaminantes, estranhas, extremas ou aberrantes. De notar que as observações extremas representam o maior e menor valor da amostra ordenada (o máximo e o mínimo), e não são obrigatoriamente *outliers*. Porém, se existirem *outliers* eles são certamente as observações extremas podendo ainda existir outros *outliers* para além do máximo e do mínimo.

Na identificação das observações aberrantes o investigador utiliza tanto a sua experiência como a sua intuição. A detecção das observações *outliers* é feita préviamente (através de análise gráfica, por exemplo), e só posteriormente, tais observações serão objecto de análises mais objectivas para testar a sua real inconsistência em relação aos restantes dados.

A preocupação com observações *outliers* é antiga e data das primeiras tentativas de analisar uma conjunto de dados. Inicialmente pensava-se que a melhor forma de lidar com esse tipo de observação seria através da sua eliminação da análise. Actualmente este procedimento é ainda muitas vezes utilizado,

existindo, no entanto, outras formas de lidar com tal tipo de fenómeno. Conscientes deste facto e sabendo que tais observações poderão conter informações importantes em relação aos dados, sendo por vezes as mais importantes, é nosso propósito apresentar os principais aspectos da discussão deste assunto.

Grande parte dos autores que estudam este fenómeno referem comentários de Bernoulli datados de 1777 [ver por exemplo, Barnett e Lewis(1994) pág.27], como sendo uma das primeiras e mais importantes referências a observações *outliers*. Esses comentários indicam que a prática de rejeitar tal tipo de observação era comum naquela altura (século XVIII). A discussão sobre as observações *outliers* centrava-se na justificação da rejeição daqueles valores. As opiniões não eram unânimes: uns defendiam a rejeição das observações "inconsistentes com as restantes", enquanto outros afirmavam que as observações nunca deviam ser rejeitadas simplesmente por parecerem inconsistentes com os restantes dados e que todas as observações deviam contribuir com igual peso, para o resultado final. Em qualquer dos casos está presente uma certa subjectividade na tomada de decisão sobre o que fazer com os *outliers*.

A primeira tentativa de definição de um critério de rejeição baseado em princípios probabilísticos deveu-se a Pierce e consta num trabalho datado de 1852 [Barnett e Lewis(1994) pág.28].

As regras de rejeição provocaram uma acesa discussão, assim como outras sugestões relativas ao mesmo tema. Como consequência, assistiu-se ao aparecimento de alguns artigos e livros tratando o tema, sempre no contexto da regressão linear. Depois de 1950 assistiu-se ao aparecimento de um grande número de contribuições para o desenvolvimento desta área. Apesar de realizados muitos estudos, a noção de *outlier* parece ser ainda um pouco vaga. *Outlier* é ainda um conceito um pouco subjectivo.

Antes de decidir o que deverá ser feito às observações *outliers* é conveniente ter conhecimento das causas que levam ao seu aparecimento. Em muitos casos as

razões da sua existência determinam as formas como devem ser tratadas. Assim, as principais causas que levam ao aparecimento de *outliers* são: erros de medição, erros de execução e variabilidade inerente dos elementos da população.

No primeiro caso as observações *outliers* são originadas por medições inadequadas ou registo incorrecto de alguns valores.

No segundo caso, os *outliers* resultam da definição imprópria da amostra. Esta pode ser desequilibrada (enviesada) ou conter elementos que não são representativos da população em estudo.

No último caso a variação é uma característica natural e incontrolável. Neste caso, os *outliers* ocorrem naturalmente e por isso contêm informação que pode levar a importantes descobertas.

O estudo de *outliers*, independentemente da(s) sua(s) causa(s) pode ser realizado em várias fases.

A fase inicial é a da identificação das observações que são potencialmente aberrantes. A identificação de *outliers* consiste na detecção, com métodos subjectivos, das observações surpreendentes. A identificação é feita, geralmente, por análise gráfica ou, no caso de o número de dados ser pequeno, por observação directa dos mesmos. São assim identificadas as observações que têm fortes possibilidades de virem a ser designadas por *outliers*.

Na segunda fase, tem-se como objectivo a eliminação da subjectividade inerente à fase anterior. Pretende-se saber se as observações identificadas como *outliers* potenciais o são, efectivamente. São efectuados testes à ou às observações "preocupantes". Devem ser escolhidos os testes mais adequados para a situação em estudo. Estes dependem do tipo de *outlier* em causa, do seu número, da sua origem, do conhecimento da distribuição subjacente à população de origem das observações, etc.

As observações suspeitas são testadas quanto à sua discordância. Se for aceite a hipótese de algumas observações serem *outliers*, elas podem ser designadas como discordantes.

Uma observação diz-se discordante se puder considerar-se inconsistente com os restantes valores depois da aplicação de um critério estatístico objectivo. Muitas vezes o termo discordante é usado como sinónimo de *outlier*. O mesmo será feito por nós. Utilizaremos também os termos *outlier* inferior e *outlier* superior para nos referirmos aos valores considerados aberrantes por serem, respectivamente, muito inferiores ou muito superiores relativamente aos dados restantes.

Na terceira e última fase é necessário decidir o que fazer com as observações discordantes. A maneira mais simples de lidar com essas observações é eliminá-las. Como já foi dito, esta abordagem, apesar de muito utilizada, não é aconselhada. Ela só se justifica no caso de os *outliers* serem devidos a erros cuja correcção é inviável. Caso contrário, as observações consideradas como *outliers* devem ser tratadas cuidadosamente pois contêm informação relevante sobre características subjacentes aos dados e poderão ser decisivas no conhecimento da população à qual pertence a amostra em estudo.

A forma alternativa à eliminação é a acomodação dos *outliers* (*accommodation of outliers*). Tenta-se "viver" com os *outliers*. A acomodação passa pela inclusão na análise de todas as observações, incluindo os possíveis *outliers*. Independentemente de existirem ou não *outliers*, opta-se por construir protecção contra eles. Para tal, são efectuadas modificações no modelo básico e/ou nos métodos de análise. Às observações *outliers* é atribuído um peso reduzido. Ao serem menosprezadas, estas observações não influenciam demasiadamente o valor das estimativas dos parâmetros. Esta abordagem passa pelo menosprezo das observações aberrantes que poderiam, eventualmente, influenciar demasiadamente os resultados.

A utilização de técnicas e procedimentos robustos é a forma mais utilizada para acomodação a *outliers*. O uso de técnicas robustas poderá ser uma exigência, à priori, do investigador com apenas um objectivo em mente: garantir os melhores resultados na inferência. Mas, a acomodação poderá surgir como defesa em relação à influência nefasta que as observações discordantes podem causar. Neste

caso, as técnicas robustas são utilizadas quando existem fortes indícios da presença de observações aberrantes.

De entre estas duas abordagens, identificação e acomodação, a primeira parece-nos ser a de maior importância. Os métodos de acomodação requerem grande informação sobre o processo gerador dos *outliers* e são criados para serem imunes à presença desse tipo de observação. Desta forma, eles tendem a esconder ou menosprezar informação essencial contida nos dados. Pelo contrário, os métodos de identificação pretendem dar a conhecer essa informação e apresentar as características do conjunto de dados em análise.

Por ser um tema de grande importância e interesse, o estudo de *outliers* ocupou e continua a ocupar muitos investigadores das mais diversas áreas. A detecção de *outliers* em amostras univariadas é um dos tópicos de extrema importância na literatura estatística. Os trabalhos mais interessantes devem-se a Anscombe(1960), Grubbs(1969), Tietjen e Moore(1972), Rosner(1975), Cook(1977) e Brant(1990). Sem referência não pode ficar o grande contributo dado pelos livros de Barnett e Lewis(1994) e Hawkins(1980) assim como do artigo de Beckman e Cook(1983).

Porém, menos trabalho foi desenvolvido em relação aos *outliers* multivariados. Um *outlier* multivariado é aquela observação que apresenta um "grande" distanciamento das restantes no espaço p-dimensional definido por todas as variáveis. No entanto, um *outlier* multivariado não necessita ter valores extremos em qualquer uma das variáveis.

No estudo de *outliers* multivariados é necessária a existência de mais uma fase que as necessárias na análise univariada. Assim, para além da detecção e teste formal das observações aberrantes em relação ao modelo básico e utilização de métodos de acomodação na inferência, é necessária a utilização de um princípio apropriado de ordenação das observações de forma a expressar a sua aberrância. O objectivo é transformar as observações multivariadas, de dimensão p, num escalar.

Geralmente, com este tipo de transformação, perde-se alguma informação relativa à estrutura multivariada dos dados.

Apenas nas últimas duas décadas foi dada alguma atenção a este tema. A principal razão parece ser o acréscimo de dificuldade com a passagem de uma amostra univariada para uma multivariada. Muitas das primeiras propostas para a identificação de *outliers* multivariados referem-se a métodos baseados na análise gráfica. As contribuições mais importantes devem-se a Gnanadesikan(1977), Atkinson(1981), Rousseuw e Zomeren(1990) e Hadi(1992).

Desde os primórdios do estudo de *outliers*, o modelo de regressão linear foi o contexto que monopolizou os trabalhos mais importantes. São muitos os autores com trabalhos neste domínio. A título de exemplo, refiram-se os seguintes: Cook(1977), Andrews e Pregibon(1978), Draper e John(1981), Chambers e Heathcote(1981), Cook e Weisberg(1982), Rosner(1983), Marasinghe(1985) e Barnett e Lewis(1994).

O estudo de *outliers* tem sido realizado em outros domínios além da regressão linear, por exemplo: dados circulares [Collet(1976)], análise discriminante [Campbell(1978)], tabelas de contigência, [Gentleman(1980) e Galpin e Hawkins(1981)]; "Factorial experiments" [Daniel(1960)]; distribuições não Normais, Gama e Exponencial [Lewis e Fieller(1979) e Kimber(1982)] e componentes principais com o contributo de Jolliffe(1986) e Gnanadesikan(1977).

O nosso trabalho surge na sequência do interesse que este tema nos desperta. Pretendemos analisar um conjunto de dados reais já disponiveis, e pareceu-nos necessário, antes de mais, proceder à identificação de *outliers* que, eventualmente, estejam presentes nesse conjunto de dados.

Assim, depois de termos presente quais os últimos desenvolvimentos sobre *outliers*, quisemos aplicar alguns dos métodos de identificação de observações *outliers* a dados reais.

O objectivo é detectar e testar outliers segundo diferentes perspectivas. Os dados representam 400 empresas que foram observadas em relação a quatro

variáveis: rotação do activo, solvabilidade, produtividade do trabalho e rentabilidade dos capitais próprios. Pretende-se verificar se dentro deste conjunto de empresas, com actividade no nosso país, existem algumas que se distanciam significativamente das restantes e podem ser consideradas como *outliers*.

O trabalho por nós realizado divide-se em duas partes, cada uma com dois capítulos. Na primeira, são abordados os aspectos teóricos e metodológicos na identificação e tratamento de *outliers*. No primeiro capítulo merecem especial atenção os modelos de discordância. São indicadas as estatísticas de teste mais importantes na detecção de *outliers* no caso de os dados serem provenientes de populações Normais, Gama e Exponenciais. São também indicados alguns testes multivariados e alguns testes não paramétricos. Finalmente, aborda-se o tema da acomodação de *outliers*, ainda que de forma breve.

No segundo capítulo apresenta-se a análise de componentes principais. Esta técnica é adequada quando se está perante dados multivariados e sem qualquer tipo de relação visível entre as variáveis, por exemplo do tipo linear. A análise de componentes principais é também utilizada como método de identificação de observações aberrantes através da utilização de algumas estatísticas de teste que incluem informação dada por várias componentes. Será revelado o seu papel importante na identificação de *outliers*.

Na segunda parte, analisam-se os dados com o objectivo de detectar e testar *outliers* segundo diferentes perspectivas.

No terceiro capítulo faz-se a análise exploratória dos dados. A distribuição de cada uma das variáveis é analisada quanto à simetria, normalidade e presença de *outliers* potenciais. São utilizados, principalmente, instrumentos gráficos e alguns testes formais. Ao verificar-se um distanciamento significativo em relação à distribuição Normal são operadas algumas transformações na distribuição inicial das variáveis de modo a aproximá-las da normalidade para posterior aplicação de alguns métodos para identificação de *outliers*. Depois de transformadas, as

variáveis são novamente analisadas. É verificada a normalidade através do teste do Qui-Quadrado e de Kolmogorov-Smirnov.

No quarto capítulo são efectuados testes de discordância às observações com maior afastamento das restantes, supondo a normalidade da população. A utilização de métodos não paramétricos será evidenciada. Apenas um número muito reduzido de empresas é claramente identificado como discordante.

A análise de componentes principais ao permitir analisar cada empresa com a informação de todas as variáveis, em simultâneo, permite identificar empresas *outliers* que passariam despercebidas na análise univariada, por isso são aplicadas as estatísticas de teste mais importantes utilizando as componentes principais.

Finalmente, nas "conclusões" são resumidos os resultados mais importantes que foram obtidos ao longo da dissertação.

I PARTE - ASPECTOS TEÓRICOS E METODOLÓGICOS

Desde há algum tempo que as observações que, num conjunto de dados, parecem afastar-se das restantes, merecem a atenção de vários investigadores. Tais observações podem ter uma grande influência nos resultados de modo que as conclusões podem ser distorcidas ou pouco realistas. É natural a preocupação com as consequências que este tipo de observação pode causar.

Existem diferentes perspectivas de abordar o problema das observações outliers. Em primeiro lugar, deve ter-se em conta que os métodos utilizados devem ser adequados ao tipo de dados em análise, que podem ser univariados ou multivariados. Com o aumento da dimensão dos dados as dificuldades são crescentes.

Iremos abordar, nesta primeira parte alguns dos métodos mais importantes de identificação de observações outliers. São abordados os aspectos da acomodação de outliers assim como o caso paramétrico. A identificação de outliers multivariados também merece alguma atenção. É nesse contexto que é usada a análise de componentes principais. Porém, é evidenciado o papel importante dos modelos de discordância, com maior ênfase para o caso de os dados pertencerem a populações normais.

CAPÍTULO 1 - MÉTODOS DE DETECÇÃO E ACOMODAÇÃO DE OUTLIERS

1 - MÉTODOS DE DETECÇÃO E IDENTIFICAÇÃO DE OUTLIERS

1. 1 – Modelos de discordância

Uma abordagem muito usada na identificação de outliers é a utilização de modelos de discordância. Num modelo de discordância considera-se que num dado conjunto de dados, se existirem observações aberrantes elas têm uma distribuição diferente da das restantes observações ou distribuição idêntica mas com parâmetros distintos. Assim, em cada modelo de discordância é considerada a hipótese nula, H, segundo a qual a amostra foi retirada de uma população com distribuição específica F, que pode ou não ser conhecida e ser especificada completamente ou não, e onde não existem observações "anormais". Em oposição, a hipótese alternativa, $\overline{H}$ considera que todas as observações ou apenas as "anormais" têm uma distribuição diferente da da hipótese nula. A hipótese nula será rejeitada em favor da hipótese alternativa se existirem observações aberrantes.

Para decidir pela aceitação ou rejeição da hipótese nula, da não existência de outliers, é necessário utilizar testes de discordância que tenham distribuição conhecida ou valores críticos tabelados. A construção destes testes depende fundamentalmente do tipo de hipótese alternativa que se está a utilizar no modelo de discordância.

Segundo Barnett e Lewis(1994) a hipótese alternativa pode ter diversas formas:

i) Alternativa determinística

Este tipo de hipótese alternativa é indicada para os casos em que as observações "anormais" são originadas por erros de medição ou de registo. Tal alternativa é específica de cada conjunto de dados. Se um dado conjunto de n observações, $x_1, x_2, ..., x_n$ contém uma observação outlier x_i , que foi originada

por erro de registo, então deve-se rejeitar o modelo básico F para todas as observações em favor do modelo alternativo. O modelo alternativo considera que todas as observações x_j ($j \neq i$) provêm do modelo F e que isso não acontece com x_i sendo necessário fazer nova leitura, corrigir ou até rejeitar algumas observações. Não é necessário nenhum teste de discordância, a rejeição do modelo inicial é deterministicamente correcta.

ii) Alternativa inerente

Este caso contempla as situações em que os outliers são originados não por erros de medição ou de execução mas pela variabilidade inerente da população à qual pertencem. Considera-se a hipótese nula de ausência de outliers, de os dados pertencerem a uma população com distribuição específica F. Ao surgirem valores aberrantes considera-se que a hipótese inicial não é adequada e supõe-se que todas as observações seguem uma nova distribuição alternativa, G, onde as observações aberrantes deixam de o ser. F e G podem ser distribuições diferentes ou apenas uma única distribuição com parâmetros distintos.

iii) Alternativa de "mistura" ou por contaminação

À hipótese inicial H:F opõe-se em alternativa a hipótese de as observações seguirem a distribuição $(1-\lambda)F+\lambda G$, que é uma mistura das distribuições F e G, onde λ é o parâmetro de mistura, e $0 \leq \lambda \leq 1$.

Em vez de admitir que os outliers reflectem a variabilidade inerente à população admite-se a possibilidade de "erros de execução" permitirem a contaminação da amostra por alguns membros de uma população que é diferente da representada pelo modelo básico. Esses membros estranhos ao modelo básico são outliers–discordantes. Rosado(1984) define λ como coeficiente de contaminação que introduz no modelo as possíveis observações contaminantes vindas da população G.

iv) Alternativa por deslizamento (slippage)

É o tipo de hipótese alternativa mais estudado. Supõe-se que todas as observações, excepto um pequeno número k (1 ou 2), provêm do modelo inicial com μ e σ^2 os parâmetros de localização e escala, respectivamente.

As k observações são observações independentes de uma versão modificada de F em que μ e σ^2 foram alterados: μ em qualquer direcção e σ^2, quase sempre, aumentando. Geralmente considera-se F como sendo a distribuição Normal.

Os modelos A e B de Ferguson, definidos num artigo datado de 1961[ver por exemplo Barnett e Lewis(1994) pág. 49], permitem justificar a presença de observações "anormais": o modelo A admite que o outlier resultou de uma translacção enquanto que no modelo B se supõe ter havido um aumento na variância.

Modelo A (efeito na média): $x_1, x_2, ..., x_n$ vêm independentemente de populações normais com variância comum σ^2 e sob H têm média comum, μ. Há constantes conhecidas $a_1, a_2, ..., a_n$ (a maioria das quais são zero), um parâmetro desconhecido Δ e uma permutação desconhecida $(\gamma_1, \gamma_2, ..., \gamma_n)$ de (1,2, ...,n) tal que a distribuição Normal correspondente a x_i tem média $\mu_i=\mu+\sigma\Delta a_{\gamma i}$, (i=1,2,...,n).

A alternativa $\overline{H}$ admite $\Delta\neq 0$ (ou apenas a hipótese unilateral, por exemplo $\Delta>0$ quando os a_i tiverem o mesmo sinal).

Modelo B (efeito na variância): $x_1, x_2, ..., x_n$ vêm independentemente de populações normais com média comum μ e sob H têm variância comum σ^2. Há constantes positivas conhecidas $a_1, a_2, ..., a_n$ (a maioria das quais são nulas), um parâmetro desconhecido Δ e uma permutação desconhecida $(\gamma_1,\gamma_2,...,\gamma_n)$ de (1,2,....,n) tal que a distribuição Normal correspondente a x_i tem variância $\sigma_i^2=\sigma^2\exp(\Delta a_{\gamma i})$, (i=1,2,...,n). Pela alternativa $\overline{H}$ é fixado $\Delta>0$. Barnett & Lewis(1994), consideram que $\Delta<0$ é irrelevante no estudo de outliers, no entanto,

Rosado(1984) defende que a hipótese $\Delta<0$ deve ser considerada no problema de outliers.

Estes modelos são bastante gerais sendo possível a sua utilização sem qualquer restrição quanto ao número de outliers presentes na amostra. Geralmente considera-se F como sendo a distribuição Normal.

v) Alternativa permutável (com variáveis permutáveis, de origem Bayesiana)

$x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n$ são observações independentes de distribuição F do modelo inicial e x_i é uma observação aleatória de distribuição G. Assume-se que o índice i da observação discordante poderá ter qualquer um dos valores 1,2,...,n. As variáveis aleatórias $X_1,X_2,...,X_n$ neste modelo não são independentes, mas sim permutáveis.

Outra questão a considerar é a maior ou menor tendência dos modelos – distribuições – para gerarem observações aberrantes. Assim as distribuições podem ser "outlier-prone", ter tendência para gerar outliers ou "outlier-resistant", ou seja, resistentes ao aparecimento de outliers.

Segundo Neyman e Scott, referido por Green(1976), uma família de distribuições $\mathcal{J}$ é completamente "outlier-prone" se para cada $\varepsilon>0$, $k>0$ e $n>2$ existe uma distribuição $F_{\varepsilon,k,n} \in \mathcal{J}$ tal que para uma amostra de tamanho n de $F_{\varepsilon,k,n}$ tem-se,

$$P[X_{(n)} - X_{(n-1)} > k(X_{(n-1)} - X_{(1)})] > 1-\varepsilon ,$$

onde $X_1, X_2, ..., X_n$ são as variáveis aleatórias com distribuição comum F e que deram origem às n observações independentes da amostra e $X_{(1)}, X_{(2)}, ..., X_{(n)}$ são esses valores ordenados de forma crescente. Se uma família de distribuições não é "outlier-prone" ela será resistente a outliers.

A definição anterior não se aplica a distribuições individuais, por isso nenhuma distribuição individual pode ser completamente "outlier-prone". Porém,

o interesse reside nas distribuições individuais uma vez que as observações provêm de uma distribuição e não de uma família de distribuições.

Em relação às distribuições individuais, Green(1976) apresenta as definições de distribuições absoluta e relativamente resistentes a outliers e distribuições absoluta e relativamente "outlier-prone", em termos da função de distribuição.

Para as distribuições conhecidas, cuja densidade existe e é conhecida ela é preferivel à função de distribuição para representar a distribuição. Neste caso são as seguintes as condições para se dizer que uma distribuição tem tendência para outliers ou é resistente a eles assumindo que f(x) é a função densidade, $F(\infty)=1$ e $F(x)<1$ para todo o x finito (seguindo Green(1976)):

i) $f(x+\varepsilon)/f(x) \to 0$ assim que $x \to \infty$ para todo o $\varepsilon>0$

ii) $f(kx)/f(x) \to 0$ assim que $x \to \infty$ para todo o $k>1$

iii) Existem constantes $\varepsilon>0$, $\delta>0$ e x_0 tal que $f(x+\varepsilon)/f(x) \geq \delta$ para todo o $x>x_0$.

iv) Existem constantes $k>1, \delta>0$ e x_0 tal que $f(kx)/f(x) \geq \delta$ para todo o $x>x_0$.

Para definir a resistência absoluta e relativa a outliers são suficientes i) e ii), respectivamente e para definir a tendência relativa e absoluta para gerar outliers, são suficientes iii) e iv), respectivamente.

Os conceitos anteriores, de tendência e resistência absoluta e relativa a outliers referem-se à aba direita da distribuição (pode ser feito o mesmo para a aba esquerda).

Existem seis classes de distribuições, segundo as suas propriedades relativamente aos outliers, considerando a aba direita:

Classe I – Distribuições que são absolutamente e também relativamente resistentes a outliers, e nunca podem ser absolutamente "outlier-prone".Um exemplo é a distribuição Normal.

Classe II – Distribuições que são relativamente resistentes a outliers e não são absolutamente resistentes a outliers nem absolutamente "outlier-prone". Um exemplo é a distribuição de Poisson.

Classe III – Distribuições que são absolutamente " outlier-prone" e relativamente resistentes a outliers. A distribuição Gama pertence a este grupo.

Classe IV – Distribuições que são absolutamente outlier-prone e não são relativamente resistentes a outliers nem relativamente outlier-prone.

Classe V – Distribuições que são relativamente "outlier-prone" e são também absolutamente "outlier-prone", mas não podem ser relativamente resistentes a outliers. A distribuição de Cauchy pertence a esta classe.

Classe VI – Distribuições que não são relativamente resistentes a outliers nem absolutamente "outlier-prone".

1.1.1 – Tipos de testes

Independentemente da hipótese nula e da alternativa, e por vezes ignorando-as, podem ser definidas algumas estatísticas para os testes de discordância a

outliers. Sendo uns mais apropriados que outros em diferentes situações, vamos apresentar os seis tipos básicos de testes estatísticos:

1) Estatística do tipo excesso/dispersão

São rácios de diferenças entre um outlier e o seu vizinho mais próximo e uma medida de dispersão. Como exemplo temos uma estatística de Dixon[Barnett e Lewis(1994) pág. 38] indicada para testar $x_{(n)}$ sem usar $x_{(1)}$,

$$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$$

e

$$\frac{x_{(n)} - x_{(n-1)}}{\sigma}$$

onde σ é o desvio padrão na hipótese inicial. σ pode ser substituído por uma estimativa independente, se for conhecida, ou então por uma estimativa baseada numa amostra restrita que exclui observações contra as quais nos queremos proteger como o outlier $x_{(n)}$ ou outras observações extremas.

2) Estatísticas de amplitude/dispersão

Aqui o numerador é a amplitude amostral,

$$\frac{x_{(n)} - x_{(1)}}{s}.$$

Usando a amplitude tem-se a desvantagem de não se saber, sem análises posteriores, se os resultados representam discordância de um outlier como observação extrema máxima, mínima ou ambas. O desvio padrão da amostra, s, pode ser substituído por um valor equivalente da amostra restrita, estimativa independente, valor conhecido ou por uma medida de dispersão da população.

3) Estatística de desvio/dispersão

Neste tipo de estatísticas coloca-se no numerador uma medida de distância entre o potencial outlier e alguma medida de tendência central dos dados. Por exemplo,

$$\frac{\bar{x} - x_{(1)}}{s}$$

em que se tem a observação com o menor valor como possível outlier. Tal como s, $\bar{x}$ pode ser calculado em função da amostra restrita, ou substituído por uma estimativa independente, pelo valor observado na população ou alguma medida de localização. Uma variante deste tipo de estatística, consiste em usar no numerador o desvio máximo, isto é $\max_i \frac{|x_i - \bar{x}|}{s}$.

4) Estatísticas de "soma de quadrados"

São rácios de somas de quadrados da amostra restrita e da amostra total. Por exemplo:

$$\frac{\sum_{i=1}^{n-2} (x_{(i)} - \bar{x}_{n,n-1})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{onde} \quad \bar{x}_{n,n-1} = \frac{\sum_{i=1}^{n-2} x_{(i)}}{n-2},$$

é a média dos $x_{(i)}$ com a exclusão de $x_{(n-1)}$ e $x_{(n)}$ e foi proposto pela primeira vez por Grubbs para testar dois outliers $x_{(n-1)}$ e $x_{(n)}$.

5) Estatísticas dos momentos de ordem superior

Podem ser usadas medidas tais como medidas de enviesamento e achatamento.

Exemplo:

$$\frac{n^{\frac{1}{2}} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{\frac{3}{2}}} \quad \text{e} \quad \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2}$$

6) Estatísticas de localização/extremos

São rácios entre valores extremos e medidas de localização. São importantes para examinar outliers quando a hipótese nula considera a família de distribuições Gama. Exemplo: $\frac{x_{(n)}}{\bar{x}}$

A maior parte dos testes anteriores enfrenta dificuldades quando a amostra contém mais que um outlier. Um desses problemas é o chamado efeito de camuflagem ("masking"), onde a existência de um outlier torna difícil a detecção de outros outliers que estão assim disfarçados, camuflados. Logo, são identificados menos outliers do que na realidade existem.

Outro problema, e oposto ao anterior, é o efeito de "swamping", em que as observações "anormais" fazem com que algumas observações "normais" sejam consideradas como aberrantes. São identificados mais outliers do que os que realmente existem.

Segundo vários autores, e sob o ponto de vista da robustez, verifica-se o efeito de camuflagem porque os testes (procedimentos) têm um ponto crítico ("breakdown-point") muito baixo. O ponto crítico é definido, grosso-modo, como sendo a mais pequena quantidade de contaminação que faz com que o estimador tome um valor arbitrariamente grande e aberrante. Assim, tendo pontos críticos elevados as medidas amostrais de localização e dispersão resistentes evitam o efeito de camuflagem na maioria das situações.

Segundo Hoaglin *et al.*(1986), existe uma classe de testes resistentes aos tipos de problemas definidos anteriormente, ou seja, não são muito sensíveis a observações aberrantes. Um conjunto básico de estatísticas sumárias usado na análise exploratória de dados é constituído pela mediana amostral, e pelos quartos amostrais (ou quartis). O uso de medidas pouco sensíveis, como é o caso da mediana e dos quartis, na construção dos testes faz com que eles sejam relativamente resistentes a observações outliers.

Os 1° e 3° quartis coincidem com os quartos ou seja, $q_{0,25}=F_l$ e $q_{0,75}=F_u$ onde F_l é o quarto inferior (lower fourth) e F_u o quarto superior (upper fourth). A dispersão quartal $d_F=F_u-F_l$, que no caso das distribuições de frequência de variáveis contínuas se confunde praticamente com a amplitude inter-quartis, representa a amplitude do intervalo que compreende 50% das observações centrais da amostra. Não são consideradas, no seu cálculo, 25% das observações menores e 25% das observações maiores.

Segundo Murteira(1993), qualquer valor da amostra é considerado outlier severo quando

$$x_i<F_l-3d_F \quad \text{ou} \quad x_i>F_u+3d_F$$

e outlier moderado quando

$$F_l-3d_F<x_i<F_l-1,5d_F \quad \text{ou} \quad F_u+1,5d_F<x_i<F_u+3d_F.$$

Os valores F_l-3d_F e F_u+3d_F são as barreiras externas inferior e superior e $F_l-1,5d_F$ e $F_u+1,5d_F$ são as barreiras internas inferior e superior, respectivamente.

Na utilização de testes formais de outliers deve ter-se em conta que eles dividem-se em duas classes:

– aqueles que testam a presença de outliers mas não identificam observações particulares como outliers, e

– aqueles em que as observações discordantes da amostra são identificadas como sendo outliers.

Os últimos têm a forma típica de

$$T = \text{máx}_{1 \le i \le n}\ h_i\,(X_i, U)$$

onde U é uma estatística baseada na amostra total (não restrita) e h_i uma função dessa estatística e das observações. Rosado(1984) na sua tese de doutoramento dedica grande atenção a este último grupo de testes.

A identificação da observação ou das observações aberrantes passa por várias etapas. Em primeiro lugar há que formular as hipóteses que definem o modelo. As hipóteses devem ser devidamente testadas e, finalmente, se existirem outliers é necessário decidir o que fazer. A sua rejeição ou eliminação, a acomodação a essas observações ou simplesmente a sua identificação são alguns dos caminhos possíveis.

Seja a hipótese H, em que se admite que todas as observações $x_1,...,x_n$ têm a mesma densidade $f(x_i;\theta)$, $(i=1,2,...,n)$. A função de verosimilhança nesta hipótese, da não presença de outliers é dada por

$$L_H(x_1,...x_n;\theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Na hipótese alternativa $\overline{H}$ admite-se a presença de uma observação "anormal" que poderá ser qualquer uma das n observações.

Se $\overline{H}_j$ é a hipótese de que x_j é a observação aberrante, então:

$-x_j$ tem densidade de probabilidade $f(x_j;\theta')$ para algum $j \in (1,2,...,n)$;

$-$as restantes observações $x_1,...,x_{j-1},x_{j+1},...,x_n$ têm função densidade de probabilidade $f(x_i;\theta)$ para $i \neq j$.

Então a hipótese $\overline{H}$ pode considerar-se como reunião das n hipóteses alternativas $\overline{H}_j$, ou seja, $\overline{H} = \bigcup_{j=1}^{n} \overline{H}_j$. Neste caso tem-se a função de verosimilhança :

$$L_{\overline{H}_j}(x_1,...,x_n;\theta,\theta') = \left[\prod_{i \neq j}^{n} f(x_i;\theta)\right] f(x_j;\theta')$$

Se representarmos por $\hat{\theta}$ o estimador da máxima verosimilhança para θ em H e $\hat{\theta}_j$ e $\hat{\theta}'_j$ os estimadores da máxima verosimilhança para θ e θ' na hipótese $\overline{H}$, então os máximos das funções de verosimilhança sob H e $\overline{H}_j$ são respectivamente:

$$\hat{L}(x_1,...,x_n;\ \hat{\theta}) \quad \text{e} \quad \hat{L}_j(x_1,...,x_n;\ \hat{\theta}_j;\ \hat{\theta}'_j)$$

onde o índice j indica que a observação x_j é aberrante (o chapéu pode omitir-se).

Para testar a homogeneidade da amostra constrói-se um teste baseado na razão das veromilhanças máximas,

$$\lambda = \frac{\max_H L(x_1,...,x_n;\ \theta)}{\max_{H\cup\overline{H}} L(x_1,...,x_n;\ \theta;\ \theta')}$$

$$= \frac{L}{\max(L,\ \max_j L_j)}$$

Sabe-se que $0 \le \lambda \le 1$ e que quanto mais próximo de 1 for o valor de λ mais provável é a aceitação da hipótese de homogeneidade das observações, uma vez que neste caso numerador e denominador tendem a ser aproximadamente iguais. $\lambda < c$ (com $c<1$) define uma região de rejeição para H.

Se dividirmos o numerador e o denominador por L tem-se

$$\lambda = \frac{1}{\max\left(1,\ \max_j \frac{L_j}{L}\right)}$$

e seja $T(X_1,...,X_n) = \max_j \frac{L_j}{L} = \frac{\max_j L_j}{L}$ então a regra de rejeição de H é:

$$\frac{1}{\max\left[1,\ T(X_1,...,X_n)\right]} < c \quad (<1),$$

ou seja

$$T(X_1,...,X_n) > \frac{1}{c} \quad (>1).$$

Se para a amostra $x_1,...,x_n$ a estatística $T(X_1,...,X_n)$ não verifica a condição anterior então o teste leva-nos à conclusão que a amostra é homogénea e não há, assim, nenhuma observação aberrante. Pelo contrário, se a hipótese de homogeneidade das observações da amostra é rejeitada existe então uma observação aberrante, uma vez que se verifica,

$$T(X_1,...,X_n) = \frac{\max L_j}{L} > c' = \frac{1}{c} \quad (>1).$$

O índice j onde $T(X_1,...,X_n)$ atinge o máximo define a observação com esse índice, x_j, como responsável pela não homogeneidade da amostra e x_j é então apontado como outlier. Geralmente x_j é o máximo ou mínimo da amostra.

Este método, usado para definir e testar outliers, designado método GAN (generativo com alternativa natural) por Rosado(1984) difere dos métodos tradicionais pretendendo eliminar alguma subjectividade existente nesses métodos, identificando ele próprio os outliers, mas numa fase posterior.

Nos métodos tradicionais, as observações discordantes são fixadas previamente como sendo outliers potenciais. Os testes são desenvolvidos para essas observações específicas tendo como objectivo a confirmação de que aqueles valores são ou não efectivamente aberrantes. Tais testes só servem para testar essas observações específicas.

Para além disso, este método permite também testar uma dada observação como outlier, como é feito nos métodos tradicionais.

A exposição anterior refere-se ao estudo e tratamento de apenas um outlier. Numa grande parte dos estudos existentes considera-se apenas 1 ou 2 outliers. Rosado(1984) faz a formulação para p outliers.

Seguindo de perto Rosado(1984), tem-se na hipótese nula da homogeneidade que todas as observações têm densidade de probabilidade $f(x_i; \theta)$. A função de verosimilhança é dada por:

$$L_H(x_1,...,x_n;\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

Na hipótese alternativa $\overline{H}$ admite-se a existência de p outliers que podem ser quaisquer p das n observações. Considera-se então a hipótese alternativa $H_{j1,...jp}$ que admite serem $x_{j1},...,x_{jp}$, para alguma combinação (j1,...,jp) dos índices (1,...,n) , os outliers existentes no modelo, ou seja:

– $x_{j1},...,x_{jp}$ têm densidade de probabilidade $f(x_i; \theta')$ para $i \in (j1,...,jp)$;

– as restantes observações têm f.d.p. $f(x_i; \theta)$.

A hipótese alternativa pode considerar-se como reunião das $\binom{n}{p}$ hipóteses $H_{j1,...jp}$, ou seja, $H = \cup H_{j1,...jp}$. Neste caso tem-se a verosimilhança:

$$L_{Hj1,...jp}(x_1,...,x_n; \theta; \theta') = [\prod_{i \in (j1,...jp)} f(x_i; \theta')] \times [\prod_{i \notin (j1,...jp)} f(x_i; \theta)].$$

O máximo da função de verosimilhança sob H é:

$$\hat{L} = \hat{L}(x_1,...,x_n; \hat{\theta})$$

e sob $\overline{H}_{j1,...jp}$ e' $\hat{L}_{j1,...jp} = \hat{L}_{j1,...jp}(x_1,...,x_n; \hat{\theta}_{j1,...jp}; \hat{\theta}'_{j1,...jp})$

onde $\hat{\theta}_{j1,...jp}$ é o estimador da máxima verosimilhança(EMV) para θ na hipótese $\overline{H}_{j1,...jp}$; $\hat{\theta}'_{j1,...jp}$ é o EMV para θ'na hipótese $\overline{H}_{j1,...jp}$ e $\hat{\theta}$ é o EMV para θ na hipótese H.

Fazendo o quociente das verosimilhanças máximas obtém-se

$$\lambda = \frac{\hat{L}}{\max(\hat{L}, \max_{j1,...jp} \hat{L}_{j1,...jp})}$$

sendo $\lambda < c$ (<1) a região de rejeição de H. Ou ainda,

$$\lambda = \frac{1}{\max\left(1, \dfrac{\max_{j1,...jp} \hat{L}_{j1,...jp}}{\hat{L}}\right)}$$

se for $T(X_1,...,X_n)=\max_{j1,...,jp}\frac{\hat{L}_{j1,...,jp}}{\hat{L}}=\frac{\max_{j1,...,jp}\hat{L}_{j1,...,jp}}{\hat{L}}$,

equivalentemente a região de rejeição de H será $T(X_1,...,X_n)>c$ (>1). Se a condição anterior não for verificada, então conclui-se pela não existência de observações discordantes na amostra $x_1,...,x_n$.

Rejeitada a hipótese da homogeneidade tem-se então que

$$\frac{\max_{j1,...,jp}\hat{L}_{j1,...,jp}}{\hat{L}} > c$$

A combinação (j1,...,jp) dos índices onde a estatística $T(X_1,...,X_n)$ atinge o máximo define as observações $x_{j1},...,x_{jp}$ como discordantes.

1.1.2 – Outliers em Populações normais

Fazendo jus à posição central ocupada pela distribuição Normal nos estudos estatísticos, essa distribuição foi desde o ínicio a que mereceu mais atenção nos estudos relativos a outliers. Com efeito, apenas últimamente (2 ou 3 décadas) outras distribuições não–normais mereceram a atenção dos investigadores para o estudo de outliers.

Muitas vezes usa-se também a distribuição Normal como aproximação assintótica de famílias de distribuições como a Poisson, Gama,... . É muito perigosa a sua utilização em casos não assintóticos, pois essas aproximações tendem a produzir piores resultados nas abas da distribuição que é exactamente a zona mais importante de localização dos outliers.

A aproximação à distribuição Normal embora excelente, em muitas situações e para variados objectivos, no estudo de outliers pode não ser desejável. Por isso, essas aproximações à Normal devem ser usadas, se o forem, com bastante cuidado.

Supondo, em H, que as observações foram geradas por um modelo Normal com média μ e variância σ^2, $N(\mu,\sigma^2)$, existem várias alternativas possíveis (por exemplo, alternativa de slippage) consoante os outliers são devidos a alteração na localização ou na escala, e os valores dos parâmetros são ou não conhecidos. O caso em que o outlier é devido a alteração em ambos os parâmetros, média e variância, não é estudado aqui, como acontece na maioria dos estudos, por ser de difícil manuseio e identificação.

Barnett e Lewis(1994) proporcionam-nos uma extensa listagem de testes para outliers devidos a deslizamento na média e na variância (separadamente), para 1, 2 e k($\geq$2) outliers mas em que essas observações aberrantes são identificadas previamente pelo analista e não pelos testes de discordância. Os testes são feitos para essas observações específicas e apenas confirmam ou negam a condição de outlier.

Seguidamente, apresentam-se algumas estatísticas de teste indicadas para situações em que a amostra em estudo se supõe ser proveniente de uma população Normal. Dado existir uma grande variedade de testes foram seleccionados, para apresentação neste trabalho, os mais representativos de cada caso, aqueles que apresentam uma relativa facilidade de cálculo e uma leitura intuitiva e simples. Para um estudo mais profundo aconselha-se Barnett e Lewis(1994), onde um grande número de testes é analisado com grande pormenor.

Os parâmetros μ e σ^2 podem ser ambos conhecidos, desconhecidos ou apenas um deles conhecido.

Se os valores dos parâmetros da população são desconhecidos, são estimados da seguinte forma: μ é estimado por

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \quad \text{e} \quad \sigma^2 \text{ por } s^2 = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n-1}.$$

Em alguns testes usam-se as expressões seguintes:

$$S^2 = \sum_{j=1}^{n} (x_j - \bar{x})^2 ; \quad s^2 = \frac{S^2}{n-1}; \quad S^2(\mu) = \sum_{j=1}^{n} (x_j - \mu)^2 ; \quad s^2(\mu) = \frac{S^2(\mu)}{n}, \quad S^2_{n-1,n}$$

é igual a S^2 quando são omitidas as observações $x_{(n-1)}$ e $x_{(n)}$ e $S^2_{1,n}$ é igual a S^2 quando são omitidas as observações $x_{(1)}$ e $x_{(n)}$.

- Testes para um único outlier superior ($x_{(n)}$)

A estatística de teste,

$$T_1 = \frac{x_{(n)} - \mu}{\sigma}$$

obtém-se através do critério da razão de verosimilhanças e é indicada quando a hipótese alternativa é a de deslizamento no parâmetro de localização em que uma observação é oriunda de uma distribuição Normal, $N(\mu+a;\sigma^2)$, com $a>0$. Esta estatística é relativamente vulnerável quando existe mais que um outlier e não é indicada para utilizações sucessivas. Os parâmetros, se desconhecidos, podem ser substituidos por estimativas baseadas na amostra. Tabelas para níveis de significância de 1% e 5% podem ser consultadas em Barnett e Lewis(1994, pág. 485-486) nas tabelas XIIIa(ambos parâmetros desconhecidos), XIIIe(apenas μ desconhecido) e XIIIg(ambos parâmetros conhecidos).

A estatística,

$$T_2 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}},$$

é habitualmente designada por "estatística r_{10} de Dixon" e serve para testar a discordância de $x_{(n)}$, podendo ser obtidas outras estatísticas substituindo em T_2 $x_{(1)}$ por $x_{(2)}$ ou $x_{(3)}$. É eficiente quando existe no máximo um outlier, caso contrário é vulnerável ao efeito de masking. É indicada nas situações em que a hipótese alternativa é a mesma de T_1. Os valores críticos estão tabelados na pág. 498 de Barnett e Lewis(1994), tabela XIXa.

- Testes para um outlier, $x_{(1)}$ ou $x_{(n)}$

$$T_3 = \max_j \left[\frac{x_{(n)} - \bar{x}}{s}; \frac{\bar{x} - x_{(1)}}{s} \right]$$

A estatística anterior é uma generalização de T_1 para a forma bilateral e obtém-se através do critério da razão de verosimilhanças. Serve para testar a discordância de $x_{(1)}$ ou $x_{(n)}$, sendo uma dessas observações oriunda da distribuição normal $N(\mu+a;\sigma^2)$, com $a \neq 0$, quando a hipótese alternativa admite ter havido deslizamento na média. Valores críticos podem ser consultados em Barnett e Lewis(1994), tabela XIIIb(pág.485). Se for conhecido o desvio padrão da população, σ, será usado em vez de s. Neste caso, os valores críticos encontram-se na tabela XIIIf da pág. 486 de Barnett e Lewis(1994). É relativamente imune em relação ao efeito de camuflagem por parte de outros outliers.

Rosado(1984) considera para este caso a estatística

$$T_4 = \max_j \left[\frac{x_j - \bar{x}}{\sigma} \right]^2$$

em que o outlier será a observação $x_{(1)}$ ou $x_{(n)}$ que maximize o critério. Também esta estatística é obtida pelo critério da máxima verosimilhança.

No caso de a média da população ser conhecida, esta será usada em vez da média amostral. Então, a estatística de teste transformar-se-à em:

$$T_5 = \text{máx} \left[\frac{|x_{(n)} - \mu|}{s(\mu)}; \frac{|\mu - x_{(1)}|}{s(\mu)} \right].$$

T_5 é uma estatística de teste bilateral obtida pelo critério da razão de verosimilhanças máximas para a hipótese alternativa de a distribuição de uma observação ser $N(\mu;b\sigma^2)$, $b>1$. É necessária a consideração dos módulos dado que as diferenças entre as observações extremas e a média da população podem, eventualmente, ser negativas. Os valores críticos são os da tabela XIIId da pág. 485 de Barnett e Lewis(1994).

A estatística de teste seguinte,

$$T_6 = \frac{n^{\frac{1}{2}} \sum_{j=1}^{n} (x_j - \bar{x})^3}{\left[\sum_{j=1}^{n} (x_j - \bar{x})^2\right]^{\frac{3}{2}}}$$

representa a assimetria da amostra e pode ser usada para testar um outlier que será $x_{(n)}$ ou $x_{(1)}$ conforme o sinal de $\sum_{j=1}^{n} (x_j - \bar{x})^3$ seja positivo ou negativo, respectivamente. Para mais que um outlier deve ser aplicado sucessivamente. Encontram-se os valores críticos na tabela XXa na pág. 499 em Barnett e Lewis (1994).

A estatística de teste T_7,

$$T_7 = \frac{n \sum_{j=1}^{n} (x_j - \bar{x})^4}{\left[\sum_{j=1}^{n} (x_j - \bar{x})^2\right]^2}$$

representa o achatamento (kurtosis) da amostra. É testada como outlier a observação $x_{(n)}$ ou $x_{(1)}$ que esteja mais distante de $\bar{x}$. Para mais que um outlier aplica-se o teste sucessivamente. Valores críticos encontram-se na tabela XXb na pág.499 de Barnett e Lewis(1994). As observações são consideradas discordantes para valores elevados da estatística.

A estatística seguinte representa o achatamento amostral baseado nos desvios das observações em relação à média da população, μ.

$$T_8 = \frac{\sum_{j=1}^{n} (x_j - \mu)^4}{n s^4(\mu)}$$

É testada uma observação de cada vez. No caso de existir mais do que uma observação potencialmente aberrante o teste é aplicado sucessivamente. Testa-se primeiro $x_{(n)}$ se $|x_{(n)} - \mu|$ for maior que $|\mu - x_{(1)}|$ ou $x_{(1)}$ no caso contrário. Os valores críticos a utilizar no teste são os da tabela XXc pág. 500 de Barnett e

Lewis (1994).

-Teste para dois outliers, $x_{(1)}$ e $x_{(n)}$

A estatística de teste,

$$T_9 = \frac{x_{(n)} - x_{(1)}}{s}$$

é usada para uma alternativa inerente e tem potência em relação a uma variedade de distribuições simétricas e alternativas à Normal. O desempenho é mais fraco em relação a distribuições assimétricas. Os valores críticos são apresentados em Barnett e Lewis(1994) na tabela XVIIa da página 494.

$$T_{10} = \frac{S^2_{1,n}(\mu)}{S^2(\mu)}$$

T_{10} serve para testar se $x_{(1)}$ e $x_{(n)}$ são outliers quando a média da população é conhecida. Os valores críticos encontram-se na tabela XVIb da pág.493 de Barnett e Lewis(1994). Se a variância da população for conhecida, ela será usada em substituição de $S^2(\mu)$. Neste caso, os valores críticos são os da tabela XVId de Barnett e Lewis (1994) que consta da página 493. No caso de a média ser desconhecida e a variância conhecida a estatística transforma-se em

$$T_{11} = \frac{S^2_{1,n}}{\sigma^2}$$

e os valores críticos respectivos são os da tabela XVIc da página 493 de Barnett e Lewis(1994). Valores dos testes menores que os valores críticos levam à aceitação da hipótese de aquelas observações serem outliers.

-Teste para dois outliers, $x_{(n)}$ e $x_{(n-1)}$

A estatística de teste,

$$T_{12} = \frac{S^2_{n-1,n}(\mu)}{S^2(\mu)}$$

serve para testar a aberrância de $x_{(n-1)}$ e $x_{(n)}$ quando apenas se conhece o valor da média da população e encontram-se tabelados os valores críticos na tabela XVd da pág. 492 de Barnett e Lewis(1994). Se for conhecida a variância, σ^2, ela será usada em vez de $S^2(\mu)$ e, neste caso, usa-se a tabela XVh da pág.493 da mesma referência. Se a média for desconhecida usa-se a estatística

$$T_{13}=\frac{S^2_{n-1,n}}{\sigma^2}$$

e os valores críticos da tabela XVf de Barnett e Lewis(1994, pág.492). Se os valores observados da estatística de teste são inferiores aos valores tabelados então as duas maiores observações, $x_{(n-1)}$ e $x_{(n)}$, são outliers.

- Teste para k(≥2) outliers superiores

A estatística de teste seguinte,

$$T_{14}=\frac{\left[x_{(n-k-1)}+...+x_{(n-1)}+x_{(n)}-k\bar{x}\right]}{s}$$

foi obtida através do critério da razão de verosimilhanças máximas e testa a discordância dos k potenciais outliers superiores $x_{(n-k+1)}$, ..., $x_{(n-1)}$, $x_{(n)}$. É indicada para a alternativa de deslizamento na localização, na qual k observações têm distribuição Normal $N(\mu+a; \sigma^2)$, com a>0 e ambos parâmetros são desconhecidos. No caso de μ e/ou σ conhecidos $\bar{x}$ e/ou s serão substituídos por μ e ou σ. Nas tabelas XVa, XVc, XVe, e XVg de Barnett e Lewis(1994, pág. 491-493) estão indicados os valores críticos para os testes nas situações de μ e σ desconhecidos, μ conhecido e σ desconhecido, μ desconhecido e σ conhecido, e μ e σ conhecidos, respectivamente.

Apesar de a distribuição Normal ser a mais utilizada, isso não impede que outras distribuições tenham um papel importante no estudo de outliers. É o caso da distribuição Gama e dos seus casos particulares.

1.1.3 – Outliers em populações Gama

Como se disse, a distribuição Normal tem sido a mais estudada, no contexto do estudo de outliers. Todavia, recentemente outras distribuições foram objecto desse estudo, nomeadamente a distribuição Gama e os seus casos particulares: distribuição Exponencial e distribuição do Qui-Quadrado (esta com menos interesse no estudo de outliers).

As áreas de grande importância para o estudo de outliers, são:

– Testes de "sobrevivência " (ou "life testing") com a utilização da distribuição Exponencial

– análise da variância em amostras com distribuição do Qui-Quadrado

– amostras assimétricas para as quais o uso da distribuição Gama é indicado

– outros contextos onde os modelos básicos são os de Poisson, como por exemplo o estudo de fluxo de tráfego, falhas em equipamento,...

Considere-se a distribuição Gama com função densidade de probabilidade

$$f(x)=[\lambda^r \Gamma(r)]^{-1} (x-a)^{r-1} \exp[-(x-a)/\lambda] \ , \quad (x>0, x>a, \lambda>0, a\geq 0)$$

onde r é o parâmetro de forma, λ o parâmetro de escala e "a" o parâmetro de localização (origem). Para facilidades de cálculo, geralmente considera-se a origem em 0 ou seja a=0.

A distribuição Exponencial com média λ é um caso particular da distribuição Gama quando r=1. Nos testes seguintes supõe-se que λ é desconhecido, r conhecido e a=0.

1) Teste para um único outlier "superior" $x_{(n)}$ numa amostra Gama ou Exponencial

$$T_{15}= \frac{x_{(n)}}{\sum_{j=1}^{n} x_j},$$

Esta estatística de teste representa o quociente entre o potencial outlier e a soma de todas as observações da amostra. A sua distribuição está em Barnett e Lewis(1994) na pág.197 e os valores críticos encontram-se na tabela III das páginas 473 e 474 da mesma referência. É um teste de máxima verosimilhança e é usado para a alternativa de deslizamento.

2) Teste para um único outlier "inferior" $x_{(1)}$ numa amostra Gama ou Exponencial

$$T_{16} = \frac{x_{(1)}}{\sum_{j=1}^{n} x_j}$$

Esta estatística de teste, obtida pelo critério da razão de verosimilhanças, tem a distribuição indicada em Barnett e Lewis(1994), na pág. 199 e os valores críticos na tabela V da pág. 476 da mesma referência.

3) Teste para k(≥2) outliers "superiores" numa amostra Gama ou Exponencial

$$T_{17} = \frac{\left[x_{(n-k-1)} + \ldots + x_{(n)}\right]}{\sum_{j=1}^{n} x_j}$$

Para conhecimento da distribuição da estatística anterior pode ser consultada a página 201 de Barnett e Lewis(1994) e os valores críticos na tabela VII da pág. 477. Essa tabela apenas contém valores críticos para o caso de dois outliers, $x_{(n)}$ e $x_{(n-1)}$, da distribuição Exponencial com r=1 e para os níveis de significância de 1% e 5%. Também esta estatística de teste foi obtida pelo critério da razão de verosimilhanças.

4) Teste para 2 outliers, um "inferior" e outro "superior" numa amostra Gama ou Exponencial

$$T_{18}= \frac{X_{(n)}}{X_{(1)}}$$

Esta estatística tem distribuição na pág. 204 de Barnett e Lewis (1994) e valores críticos na tabela VIII da pág. 479 da mesma referência. Os valores tabelados são para dimensões de amostras muito pequenas (até 12, apenas).

Os testes anteriores (e os restantes em Barnett e Lewis(1994) têm a desvantagem de os outliers serem identificados previamente pelo analista e não o serem pelos testes de discordância, logo os testes de discordância não são gerais, eles destinam-se a observações específicas.

Para além dos modelos de discordância outras abordagens podem ser utilizadas no estudo de outliers. O desconhecimento da distribuição da população originária dos dados poderá levar à utilização de testes não paramétricos.

1.2 – Testes não paramétricos

Procedimentos não paramétricos podem ser utilizados para identificar observações outliers. Apesar de menos potentes que os métodos paramétricos eles constituem um instrumento útil em situações de desconhecimento da distribuição dos dados. Os dois testes não paramétricos mais usados, para problemas de slippage, devem-se a Mosteller e a Doornbos. O primeiro é de maior facilidade de aplicação enquanto que o último tem mais potência. Outros testes são também abordados em Barnett e Lewis(1980), nomeadamente um teste de Walsh destinado a grandes amostras.

Teste de Mosteller

Em H supõe-se que a amostra é sub-dividida em *n* grupos de *m* observações cada, e que cada um provém de uma distribuição com função densidade f(x). Na

hipótese alternativa o grupo de ordem i (pode ser qualquer um), tem origem numa distribuição com função densidade de probabilidade f(x−a), com a>0 e desconhecido.

O teste proposto para testar H é baseado nas estatísticas de ordem das observações da amostra total (n×m observações). Supondo que os grupos (ou sub–amostras) são ordenados segundo a sua máxima observação, $G_{(1)}$, $G_{(2)}$,...,$G_{(n)}$, onde $G_{(1)}$ contém a observação de valor mais elevado de entre todas as que fazem parte da amostra ,$G_{(2)}$ tem o segundo maior valor,....., então $G_{(i)}$ é o grupo de ordem i, isto é, o grupo que contém o i-ésimo maior valor observado.

Seja M(i,j), com j>i, o nº de observações em $G_{(i)}$ que excedem as observações em $G_{(j)}$,$G_{(j+1)}$,...,$G_{(n)}$.

A estatística usada por Mosteller é M(1,2), ou seja, o nº de observações em $G_{(1)}$ que excedem todas as outras observações pertencentes aos outros grupos. Se o valor da estatística for suficientemente grande rejeita-se H, o que significa que o grupo $G_{(1)}$ que contém o maior valor observado é proveniente de uma população que deslizou em localização para a direita (existe distribuição conhecida para este teste). Encontram-se os valores críticos para a estatística de teste M(1,2) na tabela XXI em Barnett e Lewis(1980, pág.324).

É suposto que cada um dos n grupos tem a mesma dimensão: m observações. A suposição de que as dimensões de cada uma das sub-populações é diferente levaria a inúmeras dificuldades de cálculo.

Estatística de Doornbos

A estatística de Doornbos é uma generalização a *n* grupos (ou amostras) da estatística de Wilcoxon para duas amostras. Ordenam-se as *m*×*n* observações dos *n* grupos de *m* observações cada. Define-se

$$T_j = \sum_{l=1}^{m} r_{jl}$$

como a soma das ordens no grupo de ordem j (j=1,2,...,n), em que se tem r_{jl} como a ordem global da l-ésima observação do j-ésimo grupo (considerando todas as observações).

Na situação de deslizamento para a direita, rejeita-se H para valores elevados de máx T_j e no caso de deslizamento para a esquerda rejeita-se H para pequenos valores de min T_j. Seguindo de perto Barnett e Lewis(1980), rejeita-se a hipótese nula da homogeneidade dos dados a um nível de significância α se

$$\text{máx}_{j=1,2,...,n}\ T_j = T_{\text{máx}} > \lambda_\alpha$$

quando, segundo a hipótese nula, α é o menor possível sujeito a $P(T_{\text{máx}} > \lambda_\alpha) \leq \alpha$. São apresentados valores críticos para níveis de significância de 1 e 5% em Barnett e Lewis (1980, pág.325).

Se a direcção do deslizamento não é especificada, então faz-se o teste bilateral a T_j. É suposto neste caso, de deslizamento na localização que a dispersão dos *n* grupos é constante.

Analisar o deslizamento no parâmetro de dispersão, mantendo-se a localização constante é também de grande interesse. É usado o teste de Siegel-Tukey: ordenam-se os dados atribuindo-se a ordem 1 para o maior valor, 2 para o mais pequeno, 3 para o 2° mais pequeno, 4 para o 2° maior, 5 para o 3° maior,... . Calcula-se a estatística de Wilcoxon T_j, com esses valores. Para pequenos valores de T_j têm-se grandes dispersões e vice-versa. Se num grupo tanto a localização como a dispersão deslizarem para a direita (para cima) ambos os testes (localização + escala) perdem potência.

Um teste para grandes amostras foi proposto por Walsh [Barnett e Lewis(1980) pág. 284]. Este teste é indicado para deslizamento na localização em amostras sem distribuição especificada. Supondo que se prevêm k outliers inferiores (pode ser considerado o caso de os outliers serem superiores), a estatística de Walsh rejeita a hipótese nula se

$$x_{(k)} - (1+A)x_{(k+1)} + Ax_{(N)} < 0,$$

onde $N=k+(2n)^{1/2}$ e

$$A=\frac{1+B\sqrt{\left\{(\sqrt{2n}-B^2)/(\sqrt{2n}-1)\right\}}}{\sqrt{2n}-B^2-1},$$

e B é tal que $\alpha=1/B^2$, em que α é o nível de significância.

Apesar das hipóteses sobre a distribuição serem muito "simples", o teste é de aplicação restrita devido às suposições do tamanho da amostra serem muito severas. Com efeito, a dimensão da amostra, n, deve ser tal que $(2n)^{1/2}>B^2+1$. Outra restrição diz respeito à exigência da suposição prévia do valor de k. No entanto, nas situações em que se desconhece a distribuição do modelo, este método é um instrumento útil na identificação de outliers.

Qualquer que seja o caso em análise, paramétrico ou não paramétrico, a dificuldade aumenta se a amostra em estudo respeita a dados multivariados.

1.3 – Outliers em amostras multivariadas

A existência de observações discordantes com as restantes é de relativamente fácil determinação em amostras univariadas. Algumas vezes, por observação dos valores que constituem a amostra ou pela análise de alguns gráficos, é fácil identificar as observações que se afastam da maioria. Noutros casos, é necessária a aplicação de técnicas mais sofisticadas. Em ambos os casos, esta análise prévia tem de ser seguida de testes apropriados para confirmar as suspeitas de existência de observações outliers.

Quando se passa para um conjunto de dados em que foram observadas, não uma mas *p* variáveis, há um acréscimo significativo de dificuldades. No entanto, a luta contra essas dificuldades é justificada pela necessidade de obter conhecimentos, uma vez que em termos práticos é muito usual e necessário trabalhar com dados multidimensionais em vez de dados com uma dimensão apenas.

Em dados multidimensionais, uma observação é considerada outlier se está "muito" distante das restantes no espaço *p*-dimensional definido pelas variáveis.

Um grande problema na identificação de outliers multivariados surge pelo facto de que uma observação pode não ser "anormal" em nenhuma das variáveis originais estudadas isoladamente e sê-lo na análise multivariada, ou pode ainda ser outlier por não seguir a estrutura de correlação dos restantes dados. É impossível detectar este tipo de outlier observando cada uma das variáveis originais isoladamente.

Podem ser levadas a cabo análises gráficas para identificar potenciais outliers ou grupos de outliers, que serão aquelas observações que estão isoladas ou se afastam do principal (maior) grupo de valores.

A observação multivariada $\mathbf{x}$ pode ser representada por uma medida de distância

$$R(\mathbf{x},\mathbf{x}_0,\Gamma)=(\mathbf{x}-\mathbf{x}_0)'\Gamma^{-1}(\mathbf{x}-\mathbf{x}_0),$$

onde $\mathbf{x}_0$ representa a localização dos dados ou da distribuição que lhe está subjacente, e Γ representa a variabilidade das observações. $\mathbf{x}_0$ poderá ser o vector nulo $\mathbf{0}$, a média da amostra $\bar{\mathbf{x}}$ ou a média da população, μ, e Γ poderá ser a matriz de variância-covariância($\mathbf{V}$) ou a matriz de var-cov da amostra ($\mathbf{S}$), dependendo do facto de μ e $\mathbf{V}$ serem ou não conhecidos.

Esta medida de distância é habitualmente designada por distância de Mahalanobis e tem, aproximadamente, distribuição do Qui-Quadrado com p graus de liberdade.

Outra forma de identificar potenciais outliers é ordenar a amostra ordenando os valores $R_i(\mathbf{x}_0,\Gamma)=R(\mathbf{x}_i,\mathbf{x}_0,\Gamma)$. Observações com valores exageradamente elevados de $R_i(\mathbf{x}_0,\Gamma)$ podem considerar-se como outliers. Aplicando este tipo de redução aos dados, geralmente perde-se informação sobre a estrutura multivariada contida nos dados.

Também neste ponto, sobre outliers multivariados, a maior parte do trabalho desenvolvido supõe a normalidade do modelo básico. Sejam $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ as n

observações de uma amostra p-dimensional da variável aleatória X com distribuição $N(\mu,V)$ onde μ é o vector, de dimensão p, das médias e $V(p\times p)$ é a matriz de dispersão.

Duas hipóteses alternativas definidas por Ferguson [ver por exemplo Barnett e Lewis(1994) pág. 284], são:

Modelo A:

$$E(\mathbf{x}_j)= \mu+a, \quad \text{algum } j, \text{ e}$$
$$E(\mathbf{x}_i)= \mu \ (i\neq j) \quad \text{com} \quad V(\mathbf{x}_i)=\mathbf{V} \quad (i=1,2,...,n)$$

Modelo B:

$$V(\mathbf{x}_j)=bV, \quad \text{algum } j \text{ e } b>1, \text{ e}$$
$$V(\mathbf{x}_i)=\mathbf{V} \ (i\neq j) \quad \text{e} \quad E(\mathbf{x}_i)= \mu \quad (i=1,2,...,n)$$

Modelo A (V conhecido)

Sob H, o máximo da verosimilhança logaritmizada (à parte as constantes), é

$$L(\mathbf{x}|V) = -\frac{1}{2} \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})' \, \mathbf{V}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$$

Sob a alternativa, de existência de um outlier, o máximo do logaritmo da verosimilhança (à parte as constantes) é dado por

$$L_A(\mathbf{x}|V) = -\frac{1}{2} \sum_{j\neq i}^{n} (\mathbf{x}_j - \bar{\mathbf{x}}_i)' \, \mathbf{V}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}_i)$$

em que $\bar{\mathbf{x}}_i$ é a média amostral excluíndo a observação $\mathbf{x}_i$ considerada como potencial outlier e o índice i foi escolhido por maximizar $L_A(\mathbf{x}|V)-L(\mathbf{x}|V)$.

Então a observação $\mathbf{x}_i$ é dada como outlier superior, $x_{(n)}$, se maximizar $R_i(\mathbf{x},V)$. Consequentemente e implicitamente houve ordenação baseada na medida de distância $R(\mathbf{x},\bar{\mathbf{x}},V)$.

Como a suposição de conhecer V é muito irrealista, vamos examinar o caso em que V é desconhecido.

Modelo A (V desconhecido)

Sob H, o máximo da função de verosimilhança logaritmizada (à parte as constantes), é

$$L(\mathbf{x}) = -\frac{n}{2}\log|A| \quad \text{onde} \quad A = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'.$$

Sob $\overline{H}$, o máximo da verosimilhança logaritmizada é $L_A(\mathbf{x}) = -\frac{n}{2}\log\left|A^{(i)}\right|$

onde $A^{(i)}$ é calculado do mesmo modo que A mas omitindo $\mathbf{x}_i$ e i é escolhido para maximizar $L_A(\mathbf{x})-L(\mathbf{x})$, (é máximo quando $L_A(\mathbf{x})$ é máximo e isso acontece quando $|A^{(i)}|$ é mínimo).

Então neste caso, de **V** desconhecido, parece ter havido uma ordenação das observações segundo os valores de $|A^{(i)}|$. A observação correspondente ao menor $|A^{(i)}|$ é declarada como outlier.

Se fizermos $\Re_j = |A^{(j)}| / |A|$ e ordenarmos a amostra segundo $\Re_j$ a observação outlier é aquela que tem o menor $\Re_j$, ou seja, $\Re_{(1)}$.

A distribuição do teste é muito complicada, no entanto Wilks [ver por exemplo Barnett e Lewis(1994), pág. 288)] mostra que os $\Re_i$ são variáveis Beta identicamente distribuídas com $B[(n-p-1)/2; p/2)]$ e a distribuição conjunta é simétrica em R^n sujeita a $\sum\Re_i = n(1-p/n-1)$ $0 \le \Re i \le 1$ $(i=1,2,...,n)$.

Wilks testa em bloco a existência de 2, 3 ou 4 outliers usando a estatística:

$$\Re_{i1,i2,...,is} = |A^{(i1,i2,...,is)}| / |A|, \quad s=2,3,4$$

e $|A^{(i1,i2,...,is)}|$ é calculado como $|A|$ mas omitindo $x_{i1}, x_{i2}, ..., x_{1s}$. O conjunto de observações que minimizar $\Re_{i1,...,is}$ é considerado como outlier.

Para a detecção de um outlier, quando μ e V são desconhecidos é irrelevante adoptar a alternativa do modelo A ou do modelo B, uma vez que se chega aos mesmos resultados com ambos os modelos, dado que o teste tem a mesma forma [Barnett e Lewis(1994), pág. 292].

Outros métodos podem ser usados na análise de amostras multivariadas tendo em vista a detecção de observações outliers. Um desses métodos, actualmente já muito desenvolvido, é a análise de componentes principais.

2 – MÉTODOS DE ACOMODAÇÃO DE OUTLIERS

A existência de observações outliers é inevitável na maior parte dos casos. Para além da identificação e/ou rejeição dessas observações existe outra abordagem importante: a acomodação de outliers. Para os defensores da acomodação, todas as observações são importantes e não se deve prescindir de nenhuma delas. Todas as observações, à excepção das que foram originadas por erros, devem contribuir para a análise do fenómeno em causa.

A existência de valores aberrantes poderá distorcer as análises estatísticas. Em vez da eliminação de tais observações podem ser utilizados métodos robustos na inferência estatística. Deste modo, obtém-se protecção contra os valores "anormais" da amostra.

Métodos robustos são procedimentos que conduzem a estimadores robustos. Usando estimadores robustos, as estimativas (valores assumidos pelos estimadores) não se alteram significativamente se forem alteradas as hipóteses iniciais. Robustez significa insensibilidade em relação a pequenos desvios nas hipóteses iniciais.

Os métodos estatísticos robustos protegem contra os outliers, embora não sejam os outliers a principal preocupação ou o objectivo principal. A redução da influência das observações outliers no valor das estimativas é alcançada com a utilização destes métodos.

Se o interesse reside apenas na inferência sobre as características do modelo básico, independentemente da presença e natureza dos outliers, então tais observações têm um papel perturbador e os métodos robustos servem para minimizar o seu impacto.

A utilização de métodos robustos reveste-se de extrema importância, especialmente se o conjunto de dados em análise incluir observações que podem ser consideradas outliers. As observações outliers poderão ter uma grande influência sobre os parâmetros da distribuição, podendo transformá-los por completo.

Os últimos desenvolvimentos na área da acomodação de outliers no processo de inferência dividem-se em duas direcções principais. A primeira diz respeito aos métodos de estimação que proporcionam protecção contra os outliers, dando menos importância a essas observações do que às restantes. A segunda refere-se à preocupação com a robustez na presença de outliers. Os métodos de estimação e testes específicos têm em conta a natureza do modelo inicial e do modelo alternativo, que admite a existência de outliers.

2.1 – Estimação de parâmetros de localização

Em relação ao parâmetro de localização μ, uma forma de protecção consiste na utilização de estimadores robustos. A ideia de ponderar as observações de forma a obter estimadores robustos foi progredindo ao longo do séc.XX assumindo diversas formas. Para a estimação deste parâmetro de localização pode recorrer-se a diversos tipos de estimadores: estimadores–L, estimadores–M e estimadores–R.

i) Estimadores–L

Este tipo de estimador tem a forma de uma combinação linear dos valores ordenados da amostra, $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. Ou seja,

$$\mu = \sum_{i=1}^{n} c_i x_{(i)},$$

onde c_i são os ponderadores, $x_{(i)}$ são os valores ordenados da amostra e $\sum_{i=1}^{n} c_i$.

Os valores considerados para os ponderadores, c_i, determinam as características dos estimadores. Assim, este tipo de estimador tem uma grande flexibilidade uma vez que permite atribuir uma menor ponderação às observações que são outliers potenciais.

Se $c_i = \frac{1}{n}$, i=1,2,...,n, obtém-se a média aritmética como um caso particular do estimador–L.

Outro caso particular deste tipo de estimador é o chamado estimador–L de Huber, que consiste na mediana amostral, onde c_i=0 para todas as observações ordenadas excepto para a observação central (ou duas centrais, no caso de o número de observações ser par).

Os estimadores–L mais importantes são as chamadas *médias aparadas* ou "α-trimmed mean". T(α), a média aparada a 100.α%, obtém-se através da eliminação das 100.α% menores observações e das 100.α% maiores observações e calculando a média aritmética simples dos restantes valores ou de algumas combinações específicas de alguns quantis.

A *meia-média* é um caso particular que corresponde à média aparada com α=0.25, ou seja, a meia-média é a média aritmética da metade central das observações ordenadas.

Murteira(1993, pág.78) refere ainda o caso da *mediana alargada* (broad median) como um estimador–L e define-a da seguinte forma:

– Para um número de observações (n) ímpar:

.Se 5≤n≤12 é a média aritmética das três estatísticas de ordem centrais,

.Se n≥13 é a média aritmética das cinco estatísticas de ordem centrais.

– Para um número de observações (n) par:

.Se 5≤n≤12 é a média aritmética das quatro estatísticas de ordem centrais com ponderações 1/6, 1/3, 1/3 e 1/6,

.Se n≥13 é a média aritmética das seis estatísticas de ordem centrais com ponderações 1/10, 1/5, 1/5, 1/5, 1/5 e 1/10.

A mediana alargada pode considerar-se uma média variavelmente aparada:

se 5≤n≤12 → α=0,5–1,5/n e

se n≥13 →α=0,5–2,5/n

O valor de α a ser escolhido, ou seja, a quantidade ideal a aparar está intimamente relacionada com o peso das caudas da distribuição da amostra. Segundo Murteira(1993, pág.82):

– se as caudas são neutras (a distribuição é aproximadamente Normal), a média aritmética é a melhor medida de localização em termos de eficiência. O valor para α é zero.

– se as caudas são ligeiramente pesadas, para pequenas amostras (n=5), a medida mais eficiente é a média aparada a 25%. Para amostras ligeiramente maiores ($10 \leq n \leq 20$), a medida mais eficiente é a média aparada a 10%.

– se as caudas são pesadas (distribuição de Cauchy, por exemplo), a medida mais eficiente é a mediana alargada ou a mediana.

ii) Estimadores – M

Este tipo de estimador baseia-se no princípio da máxima verosimilhança. É uma generalização do princípio dos mínimos quadrados. No modelo básico, a suposição é a de que os dados da amostra seguem uma distribuição contínua com função de distribuição F(x) e função densidade f(x). A média da população, μ, é estimada através do estimador $T_n=T_n(x_1,...,x_n)$ escolhido de forma a minimizar a equação do tipo:

$$\sum_{j=1}^{n} \psi(u_j),$$

onde $u_j = x_j - T_n$ é uma função de valores reais não constantes com determinadas características. Por exemplo, se fizermos $\psi(u)=u^2$, obtém-se a média amostral, se $\psi(u)=|u|$ obtém-se a mediana da amostra enquanto que com $\psi(u)= -\log f(u)$ se obtém o estimador da máxima verosimilhança. Geralmente, escolhe-se uma função $\psi(u)$ convexa de tal forma que a sua derivada seja monótona e T_n seja único.

Alguns exemplos de estimadores–M obtidos através de certas funções $\psi(u)$ podem ser vistos em Barnett e Lewis(1994, pág.148-150).

iii) Estimadores – R

Um estimador de teste de dimensão (rank), T_n baseado nos n valores $x_j - T_n$ e nos n valores $T_n - x_j$ (j=1,2...,n), diz-se um estimador–R. O mais conhecido é o estimador Hodges-Lehmann e consiste na mediana do conjunto das n(n + 1)/2 médias do tipo $(x_j + x_k)/2$ (j<k; j=1,2,...,n; k=1,2,...,n).[ver por exemplo, Barnett e Lewis(1994, pág. 152)]

2.2 – Estimação de parâmetros de dispersão

Para estimar parâmetros de dispersão é comum usarem-se alguns dos métodos de estimação robusta para parâmetros de localização. Dadas n observações, $x_1, x_2, ..., x_n$ de uma variável aleatória X, determina-se um estimador robusto para a localização das observações. Seja x esse estimador robusto obtido através de um dos métodos vistos anteriormente e calculem-se os desvios d_i, de cada observação em relação ao estimador da localização. Escolhendo uma função apropriada h(d), ponderam-se os n valores $h(d_1),...,h(d_n)$ de forma a obter o estimador de dispersão,

$$S = \frac{1}{n^*} \sum_{i=1}^{n} h(d_i)$$

onde n^* é um divisor apropriado e não necessariamente igual a n. Por exemplo, ao estimar a variância da população através da variância amostral é habitual fazer $h(d)=d^2$ e $n^*=n-1$.

2.3 – Medidas de eficiência dos estimadores

Qualquer que seja o tipo de estimador considerado, deve ter-se em conta o seu desempenho. Para tal devem ser usadas medidas de eficiência dos estimadores. O rácio seguinte dá uma indicação quantitativa da necessidade de um estimador alternativo a $\hat{\mu}$.

$$e(\hat{\mu}) = \frac{var(\hat{\mu}|\overline{H})}{var(\hat{\mu}|H)},$$

onde $\hat{\mu}$ é um estimador de μ, $\overline{H}$ o modelo alternativo e H o modelo básico. Se o valor do rácio estiver perto da unidade, $\hat{\mu}$ é robusto e não haverá necessidade de outro estimador.

Uma forma alternativa de abordar este problema é em termos de prémio e protecção. Tais conceitos são abordados em Guttman *et al.*(1971).

CAPITULO 2 – IDENTIFICAÇÃO DE OUTLIERS EM COMPONENTES PRINCIPAIS

Uma das técnicas mais usadas na análise de dados multivariados é a análise de componentes principais (ACP). O interesse principal reside na redução da dimensão de um conjunto de dados constituído por um grande número de variáveis correlacionadas entre si, mantendo tanto quanto possível a variabilidade inicial dos dados.

Como outras técnicas clássicas, a ACP produz resultados que podem ser bastante sensíveis e por isso influenciados pela existência de observações "anormais". Deste modo, tem um papel importante na identificação de observações outliers.

Seguidamente, caracterizamos a análise de componentes principais, definindo a técnica e indicando o método de obtenção das componentes. Especial atenção será dada ao papel importante das componentes principais na identificação de outliers, nomeadamente quanto à utilização de algumas estatísticas de teste destas observações.

2.1 – Caracterização da ACP

Com o objectivo de estudar o tipo de relações existentes num conjunto de dados resultantes da observação de p variáveis correlacionadas, é comum fazer a transformação das variáveis originais em novas variáveis não correlacionadas entre si e designadas por componentes principais(CP). Essa transformação não é mais do que uma rotação ortogonal no espaço p-dimensional. Como objectivos principais da análise de componentes principais, têm-se:

1) obter novas variáveis não correlacionadas

2) reduzir a dimensão dos dados, isto é, passar de p variáveis para m ($<p$) variáveis

3) tentar estabelecer qualquer tipo de relação entre as variáveis, não explícita nos dados.

As novas variáveis, as CP, são combinações lineares das variáveis originais e aparecem por ordem decrescente de importância de tal forma que, por vezes, a primeira componente explica uma grande parte da variação dos dados originais. Sendo assim, poderão ser consideradas aquelas CP que em conjunto expliquem uma variação significativa dos dados (80 ou 90%) podendo as restantes ser desprezadas, uma vez que a informação perdida não é significativa. Deste modo, pode considerar-se um número de variáveis menor do que o número inicial com a vantagem de as mesmas não estarem correlacionadas entre si.

Não será de grande utilidade a análise em CP de um conjunto de dados que não apresente "elevada" correlação entre as variáveis observadas. No caso extremo de variáveis não correlacionadas seriam obtidas tantas CP como variáveis originais mas por ordem decrescente de importância, ou seja, ordenadas segundo as respectivas variâncias. A informação dada por cada componente seria semelhante à informação contida em cada uma das variáveis iniciais.

A análise de componentes principais é apenas uma técnica de cálculo (técnica matemática), e não requer a especificação de um modelo estatístico subjacente para explicar a estrutura dos dados. Geralmente, não são formuladas hipóteses sobre a distribuição das variáveis originais.

Vejamos agora como são obtidas as componentes principais.

Seja $\mathbf{x}^T = [X_1\ X_2 \ldots X_p]$ um vector de p variáveis aleatórias com média μ e matriz de covariâncias Σ. Pretende-se determinar um novo conjunto de variáveis, sejam $Z_1, Z_2, \ldots, Z_p$, não correlacionadas e ordenadas por forma decrescente da sua variância.

Cada Z_i ($i=1,2,\ldots,p$), é uma combinação linear da variável p-dimensional $\mathbf{x}$, isto é,

$$Z_i = a_{1i}X_1 + a_{2i}X_2 + \ldots + a_{pi}X_p$$
$$= \mathbf{a}_i^T \mathbf{x}$$

onde $\mathbf{a}_i^T = [a_{1i} \quad a_{2i} \ldots a_{pi}]$ é um vector de constantes que verificam seguinte condição de normalização

$$\mathbf{a}_i^T \mathbf{a}_i = \sum_{k=1}^{p} a_{ki}^2 = 1$$

A primeira componente, Z_1, é obtida, escolhendo $\mathbf{a}_1$ tal que Z_1 tenha a maior variância possível, ou seja, $\mathbf{a}_1$ é escolhido de forma a maximizar a variância de $\mathbf{a}_1^T \mathbf{x}$ sujeita à condição $\mathbf{a}_1^T \mathbf{a}_1 = 1$

A segunda componente é encontrada , escolhendo $\mathbf{a}_2$ de forma que Z_2 tenha a maior variância possível de todas as combinações da forma $\mathbf{a}_2^T \mathbf{x}$ e não seja correlacionada com Z_1.

De modo semelhante, obtêm-se as restantes componentes principais, Z_3,..., Z_p, que têm variância decrescente e não estão correlacionadas.

Determinação da primeira componente (Z_1)

Pretende-se determinar $\mathbf{a}_1$ de modo a maximizar a variância de Z_1 impondo a restrição $\mathbf{a}_1^T \mathbf{a}_1 = 1$. A variância de Z_1 vem dada por

$$V(Z_1) = V(\mathbf{a}_1^T \mathbf{x}) = \mathbf{a}_1^T V(\mathbf{x}) \mathbf{a}_1 = \mathbf{a}_1^T \Sigma \mathbf{a}_1.$$

Tem-se então, o seguinte problema

$$\max_{\mathbf{a}_1} \mathbf{a}_1^T \Sigma \mathbf{a}_1$$

$$\text{s.a. } \mathbf{a}_1^T \mathbf{a}_1 = 1$$

A resolução deste problema recorrendo ao método dos multiplicadores de Lagrange conduz-nos à seguinte condição necessária de primeira ordem (derivada nula):

$$(\Sigma-\lambda I)\mathbf{a}_1=\mathbf{0}$$

A equação anterior tem uma solução diferente de zero se $(\Sigma-\lambda I)$ for uma matriz singular, logo λ deve ser tal que $|\Sigma-\lambda I| = 0$. Então, existe uma solução diferente de zero se e só se λ é um valor próprio de Σ. Mas Σ terá, geralmente, p valores próprios todos não negativos visto Σ ser semi-definida positiva. Admitindo que $\lambda_1,\lambda_2,...\lambda_p$ são os valores próprios, por hipótese todos diferentes, tal que $\lambda_1>\lambda_2>...\lambda_p\geq 0$ e uma vez que se quer maximizar a variância, escolhe-se o maior dos valores próprios ou seja, λ_1. Logo, $\mathbf{a}_1$ é o vector próprio associado ao maior valor próprio, λ_1, da matriz Σ.

A segunda componente, $Z_2=\mathbf{a}_2^T\mathbf{x}$, é obtida de forma análoga à anterior com a inclusão da restrição de que Z_1 e Z_2 não devem estar correlacionadas. Obtém-se $(\Sigma-\lambda I)\mathbf{a}_2=\mathbf{0}$ e escolhe-se λ como sendo o segundo maior valor próprio de Σ e $\mathbf{a}_2$ o correspondente vector próprio. Continuando com este processo obtêm-se todas as componentes principais.

Não há dificuldade no caso de existirem alguns valores próprios de Σ iguais. Neste caso, não há uma única forma de escolher os correspondentes vectores próprios. Deve, no entanto, ter-se em conta que, nesse caso, os vectores próprios associados com as raízes múltiplas devem ser escolhidos de forma a serem ortogonais.

Os valores próprios representam as variâncias das CP a que estão associados. A soma das variâncias das variáveis originais é igual à soma das variâncias das componentes principais, logo, não se perdeu nada em termos de variabilidade e agora as variáveis estão ordenadas segundo a sua importância.

A parte da variância total explicada pela i-ésima componente principal é dada por $\lambda_i / \sum_{j=1}^{p} \lambda_j$ enquanto que a contribuição das primeiras m componentes para a variância total é dada por $\sum_{i=1}^{m} \lambda_i / \sum_{j=1}^{p} \lambda_j$.

Muitas vezes, em vez de Σ, utiliza-se a matriz de correlações **P**, o que significa que se pretendem determinar as CP de um conjunto de variáveis que foram previamente estandardizadas para terem variância unitária. O processo para determinar as componentes é o mesmo que foi usado anteriormente.

É importante ter em conta que os valores próprios e vectores próprios de **P** não serão os mesmos que se obtêm com Σ. Ao utilizarmos **P** está-se a tomar a decisão arbitrária de considerar todas as variáveis com o mesmo peso ou importância. Neste caso, a proporção da variância total explicada pela i-ésima componente é dada por λ_i/p, uma vez que a soma dos valores próprios de **P** é igual a p.

Como geralmente Σ e **P** são desconhecidos, é comum utilizarem-se as correspondentes matrizes amostrais, **S** e **R** respectivamente. Neste caso, obtêm-se os valores próprios de **S** (ou **R**), $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ e os correspondentes vectores próprios $\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \ldots, \hat{\mathbf{a}}_p$.

Como **S** é semi-definida positiva, os valores próprios são todos não negativos e representam as variâncias estimadas das componentes. $(\hat{\lambda}_i)$ e $(\hat{\mathbf{a}}_i)$ podem ser vistos como estimativas dos valores teóricos: valores próprios (λ_i) e vectores próprios (a_i) de Σ.

Se se admitir que as observações seguem a distribuição Normal multivariada então, segundo Morrison (1990),

$$\hat{\lambda}_k \sim N(\lambda_k;\ 2\lambda_k^2/(n-1)) \quad \text{e} \quad \frac{\hat{\lambda}_k - \lambda_k}{\lambda_k[2/(n-1)]^{1/2}} \sim N(0,1).$$

Porém, estes resultados são de pouco interesse prático dado que são resultados assintóticos ($n\to\infty$) e a normalidade por vezes levanta dúvidas.

Assim, a tendência moderna é considerar a ACP como uma técnica matemática sem nenhum modelo estatístico subjacente. As CP obtidas através de **S** são vistas como as componentes principais e não como estimativas do que seria obtido com Σ, por isso geralmente omitem-se os chapéus em $\hat{\lambda}_i$ e $\hat{\mathbf{a}}_i$.

Se algumas das variáveis iniciais são linearmente dependentes, alguns dos valores próprios de Σ serão nulos. A dimensão do espaço contendo as observações é igual à característica de Σ, que é dada por (p–o número de valores próprios nulos). Se existirem k valores próprios nulos, podem encontrar-se k restrições lineares e independentes nas variáveis.

A existência de dependência linear exacta é rara, mais importante é detectar dependência linear aproximada. Se o menor valor próprio λ_p é muito próximo de zero então a p-ésima componente é "quase" constante e a dimensão de **x** é "quase" menor que p.

As componentes principais correspondentes a valores próprios "pequenos" são variáveis para as quais os membros da população (amostra) têm valores aproximados (quase iguais). Essas componentes podem ser consideradas como estimativas de relações lineares subjacentes. Se $\lambda_{m+1}, ...,\lambda_p$ são "pequenos", pouca informação é perdida se forem apenas consideradas as primeiras m componentes.

É usual determinar e usar os vectores próprios estandardizados, $\mathbf{a}_k^* = \lambda_k^{1/2}\,\mathbf{a}_k$, em vez dos vectores próprios normalizados, $\mathbf{a}_k$. Os primeiros são tais que a soma dos quadrados dos elementos é igual ao valor próprio correspondente, em vez da unidade, como acontece com os segundos, porque

$$\mathbf{a}_k^{*T}\mathbf{a}_k^* = \lambda_k \mathbf{a}_k^T\mathbf{a}_k = \lambda_k .$$

Está-se a dar maior peso aos coeficientes das componentes mais importantes.

Uma característica importante das componentes principais é que elas não dependem dos valores absolutos das correlações mas sim dos rácios entre elas. Uma consequência prática muito importante resultante deste facto é que podemos ter diferentes matrizes de correlações que levam às mesmas CP. Se se dividirem todos os elementos fora da diagonal principal da matriz **R** (este resultado também se aplica a **P**), por uma constante, k>1, pode-se provar que os valores próprios são alterados mas o mesmo não acontece aos vectores próprios nem às componentes.

Outra característica importante das CP é que elas dependem das unidades em que estão expressas as variáveis originais. Se, por exemplo, uma variável tem variância muito maior que as restantes, então essa variável dominará a primeira componente principal (se for usada a matriz de covariâncias na sua determinação), qualquer que seja a estrutura de correlação, enquanto que se todas as variáveis forem transformadas de modo a terem variância unitária (têm de certa forma importância igual), então a primeira componente será bastante diferente.

Por este facto, não é conveniente utilizar a análise de componentes principais a não ser que as variáveis tenham variâncias "similares", o que normalmente acontece quando as variáveis estão medidas em percentagens ou na mesma medida.

Depois de determinar os valores próprios e as componentes principais utilizando a matriz de correlação ou de covariâncias, devem-se analisar as primeiras componentes principais que deverão explicar uma "grande" proporção da variância total. É então necessário determinar quais as CP que devem (podem) ser desprezadas por não representarem acréscimo significativo de informação.

A questão é então a de saber quantas componentes são significativas e quantas se devem considerar em análises subsequentes. Existem algumas regras, todas elas um pouco subjectivas:

–fixa-se a percentagem de variância total que se quer ver explicada e considera-se um número de componentes até essa percentagem ser atingida. Geralmente considera-se de 80 a 90%.

–consideram-se tantas componentes quantos os valores próprios iguais ou superiores a um. É a chamada regra de Kaiser e só é válida se se utiliza a matriz de correlações para obter as componentes. Jolliffe, num trabalho datado de 1972, [Jolliffe(1986) pág.95] impõe como limite 0,7.

–fazem-se ensaios de hipóteses para verificar se um determinado número de CP é significativo. Pretende-se testar se os últimos q de p valores próprios são iguais e se isso acontecer as componentes principais correspondentes podem ser eliminadas.

2.2 – Papel das componentes principais no estudo de outliers

As primeiras e as últimas componentes principais são as que têm mais interesse no estudo de outliers. As primeiras são especialmente sensíveis a outliers que inflacionam as variâncias e covariâncias, se se utiliza **S**, ou as correlações, se se utilizar **R**.

Através da análise de gráficos a duas e três dimensões, das componentes principais, podem ser detectadas observações outliers que adicionam dimensões sem importância aos dados ou escondem singularidades presentes neles.

Mas, se um outlier é causa de um grande aumento na variância de uma ou mais variáveis originais, então ele terá valores extremos – muito elevados ou muito pequenos – nessas variáveis e, por isso, é detectável pela observação de gráficos das variáveis originais.

Segundo Joliffe(1986), se uma observação inflacionar a covariância ou correlação entre duas variáveis isso poderá ser descrito num gráfico dessas duas variáveis e essa observação será aberrante em relação a uma ou ambas as variáveis consideradas isoladamente.

Em oposição, as últimas componentes principais podem detectar outliers que não aparentam sê-lo se se considerarem as variáveis originais. Isso acontece

porque uma forte estrutura de correlação entre as variáveis originais implica que haja funções lineares dessas variáveis (as componentes principais), com variâncias pequenas, se comparadas com as variâncias das variáveis originais. Examinando os valores das últimas componentes podem detectar-se observações que violam a estrutura de correlação imposta pelo conjunto de todos os dados, não sendo necessariamente aberrantes se forem consideradas as variáveis originais.

Para além da análise gráfica é possível construir testes formais para detectar outliers supondo que as componentes principais são normalmente distribuídas.

Rigorosamente, **x** devia ter distribuição Normal multivariada, mas de facto como, **Z**, as componentes principais são funções lineares das p variáveis aleatórias $X_1, X_2, ..., X_p$, pode invocar-se o teorema do limite central (se p for grande) para justificar a normalidade aproximada das CP mesmo quando as variáveis originais não têm esta distribuição.

2.3 – Tipos de testes

Estatísticas para testar a presença de outliers numa amostra univariada [algumas vistas anteriormente e outras em Barnett e Lewis(1994)], podem ser usadas em cada uma das CP consideradas individualmente. Outros testes combinam informação de várias componentes em vez de examinarem uma de cada vez.

Uma estatística sugerida por Rao(1964) é a soma dos quadrados dos valores das últimas q ($< p$) componentes:

$$d_{1i}^2 = \sum_{k=p-q+1}^{p} Z_{ik}^2 ,$$

onde Z_{ik} é o valor da k-ésima componente principal para a observação de ordem i, medido em relação à média para todas as observações. Se não existirem outliers, d_{1i}^2, i=1,2,...,n são aproximadamente observações independentes de uma distribuição Gama. Deste modo, os outliers podem ser facilmente reconhecidos,

recorrendo a um gráfico de probabilidades Gama com parâmetros adequados. Valores exageradamente elevados de d_{1i}^2 indicam que a observação i é possivelmente aberrante, ou que a observação tem um fraco ajustamento ao espaço de $(p-q)$ dimensões. Pode obter-se uma estimativa adequada do parâmetro de forma com base num conjunto dos menores valores observados de d_{1i}^2.

Uma das questões a considerar é o valor a escolher para q. Hawkins(1974), considera vários métodos para determinar o valor adequado para q. Este é um problema diferente do visto anteriormente aquando da escolha do número de componentes a utilizar em análises posteriores. Agora pretende-se conhecer o número de componentes que devem ser utilizadas, começando pela última, em vez de pela primeira. Sugerem-se várias possibilidades para escolher q, incluindo o "oposto" da regra de Kaiser, ou seja, reter as componentes com valores próprios inferiores à unidade. Jolliffe(1986) considera o ponto crítico de 1 muito elevado e utiliza o valor de 0,7.

A estatística d_{1i}^2 dá um peso insuficiente às ultimas CP, especialmente se q, o número de componentes que contribuem para d_{1i}^2 é muito próximo de p. Os valores de Z_{1k}^2 tornar-se-ão mais pequenos à medida que k aumenta, uma vez que as componentes principais estão ordenadas de forma decrescente dos valores das suas variâncias (isto significa que o seu peso ou contribuição para d_{1i}^2 vai ser muito reduzido), o que é grave uma vez que são precisamente essas componentes (as últimas, de menor variância) que são mais eficazes na detecção de certo tipo de outlier.

Para evitar esse efeito podem utilizar-se as componentes estandardizadas, ou seja,

$$Z_{ik}^{*} = Z_{ik} / \lambda_k^{1/2},$$

uma vez que desta forma todas as CP têm peso igual já que têm variâncias unitárias.

Quando $q=p$ a estatística d_{1i}^2 transforma-se em

$$d_{2i}^2 = \sum_{k=1}^{p} Z_{ik}^2 / \lambda_k$$

que é exactamente a distância de Mahalanobis entre a observação i e a média amostral (considerada como sendo a origem). Hawkins (1974) prefere usar d_{2i}^2 com ($q<p$) em vez de $q=p$, de modo a dar maior importância às componentes de menor variância. No caso de serem consideradas todas as componentes no cálculo de d_{2i}^2, então esta estatística tem aproximadamente distribuição χ_p^2.

Pode também ser considerada a estatística

$$d_{3i}^2 = \sum_{k=p-q+1}^{p} \lambda_k Z_{ik}^2$$

Se $p=q$ dá-se enfase às observações que têm um grande efeito nas primeiras componentes principais.

Hawkins(1974) mostra que podem ser detectados outliers utilizando a estatística

$$d_{4i} = \max_{p-q+1 \le k \le p} \left| Z_{ik}^* \right|$$

e também sugere métodos para escolher o valor de q mais adequado. Poderão ser obtidos ganhos, na detecção de outliers, se as últimas q componentes principais forem rodadas segundo o critério varimax antes de calcular a estatística d_{4i}. Então, esse teste estatístico para a observação de ordem i é o máximo valor absoluto das últimas q CP rodadas, avaliadas para aquela observação.

Segundo Jolliffe(1986), se não existirem outliers e os dados forem aproximadamente Normais multidimensionais, então os valores de d_{4i} são aproximadamente valores absolutos de uma variável aleatória com distribuição Normal N(0;1).

Tanto d_{3i}^2 e d_{2i}^2, quando $q=p$, como d_{1i}^2 terão aproximadamente distribuição Gama se não existirem outliers e se a normalidade aproximada das

observações puder ser assumida, de modo que os gráficos de probabilidade Gama de d_{2i}^2 (com $p=q$) e d_{3i}^2 podem ser utilizados na identificação de outliers.

Contudo, como geralmente na prática μ e Σ são desconhecidos e os dados não têm distribuição Normal multivariada, os resultados obtidos sob suposições restritivas só poderão ser considerados como aproximações. Elas devem ser particularmente exactas uma vez que a detecção de outliers se preocupa com a procura de observações que sejam bastante diferentes do resto, correspondendo a um muito pequeno nível de significância nos testes estatísticos.

Jackson & Hearne (1979) indicam que o oposto de d_{2i}^2, no qual é calculado o somatório dos quadrados das primeiras em vez das últimas componentes principais normalizadas, pode ser útil no controle de qualidade, quando o objectivo é procurar grupos de autocontrole, ou observações que são outliers.

A sua estatística básica é decomposta de modo a dar informação separada sobre a variação dentro da amostra de potenciais outliers e acerca da diferença entre a média amostral e um valor padrão conhecido.

2.4 – Estimação robusta de CP

Uma das críticas às técnicas de inferência para componentes principais é que elas dependem de uma hipótese irrealista que é a normalidade multivariada de **x**. Se o que se pretende é fazer a descrição da amostra através de CP ou o uso das componentes da amostra como estimativa das componentes da população, então a forma da distribuição de **x** não é de grande importância. A única excepção verifica-se quando existem outliers. Se os outliers são de facto observações influentes, então os resultados podem ser muito influenciados por essas observações.

O método clássico e habitualmemte usado para obtenção das componentes principais tem merecido algumas críticas por ter falta de robustez. A obtenção das

componentes baseia-se na determinação dos valores próprios e vectores próprios da matriz de covariâncias ou de correlações da amostra. O facto de se usarem como estimadores das matrizes de covariâncias ou de correlações as correspondentes matrizes amostrais leva à falta de robustez nas componentes principais. De facto, aquelas matrizes, baseadas na amostra, são especialmente sensíveis a observações outliers de tal forma que apenas uma observação, desde que suficientemente afastada do grupo formado pelas restantes, pode alterar significativamente os resultados.

Tem sido crescente o interesse na procura de estimadores robustos das matrizes de covariâncias e de correlações a serem usadas na obtenção das componentes principais. São referências importantes Devlin *et al.*(1975), Devlin *et al.*(1981), Jolliffe(1986) e Li e Chen(1985).

II PARTE - IDENTIFICAÇÃO DAS EMPRESAS PORTUGUESAS "OUTLIERS"

Ao analisar um conjunto de dados deve ter-se a preocupação de averiguar se existem algumas observações que podem ser consideradas outliers. Esta análise preliminar reveste-se de extrema importância dado que tais observações podem influenciar e distorcer os resultados.

Vamos analisar um conjunto de dados em relação à existência de observações outliers. Os dados são apresentados assim como as variáveis em análise e procede-se à análise exploratória e transformação dos dados de modo a serem aplicados alguns métodos de identificação referidos nos capítulos anteriores.

A detecção de observações outliers efectua-se em vários contextos. Faz-se a detecção de outliers supondo a normalidade do modelo básico, sendo também usada a análise de componentes principais com o mesmo objectivo. O caso não paramétrico também será analisado.

Os dados que são utilizados referem-se às maiores empresas com actividade em Portugal e foram retirados da edição especial da revista "EXAME" de Novembro de 1993. A ordenação das empresas foi feita tendo em conta o volume de vendas líquidas em 1992.

De entre um grande conjunto de informação à disposição na referida publicação foram seleccionados quatro rácios considerados como importantes para caracterizar a situação económica e financeira das empresas.

Estes rácios são:

Rotação do activo

Solvabilidade

Produtividade do trabalho

Rentabilidade dos capitais próprios

Os valores de alguns rácios estão em percentagem por ser usual neste tipo de análise.

Definição dos rácios:

Rotação do Activo (RA):

A rotação do activo é definida da seguinte forma:

$$RA = \frac{\text{vendas líquidas}}{\text{activo líquido}},$$

e mede o grau de eficiência na utilização dos recursos à disposição da empresa.

Solvabilidade (SV):

A solvabilidade mede a capacidade da empresa para satisfazer os compromissos de longo prazo e é definida como a relação entre os capitais próprios e o passivo, em percentagem.

$$SV = \frac{\text{capitais próprios}}{\text{passivo}} \times 100$$

Quanto maior o valor do rácio, melhor a empresa responde aos seus compromissos, mantendo-se uma certa autonomia financeira. Se a relação for inferior a 100, para se manter solvente, a empresa tem de ser capaz de gerar lucros para satisfazer as suas obrigações para com terceiros nos prazos previstos ou, em alternativa, os seus accionistas têm de injectar capitais na sociedade.

Produtividade do trabalho (PT):

Este rácio é calculado através da seguinte relação:

$$PT = \frac{\text{valor acrescentado bruto}}{\text{nº de trabalhadores}}.$$

A produtividade do trabalho mede a eficiência das empresas na utilização dos recursos humanos. O VAB(valor acrescentado bruto) pode ser calculado de duas formas. Aqui foi utilizada a seguinte fórmula:

VAB = vendas líquidas + trabalhos para a própria empresa + variação da produção + subsídios destinados à exploração + receitas suplementares – consumos intermédios.

Os consumos intermédios representam a soma dos custos das existências vendidas e consumidas, dos subcontratos, dos fornecimentos e serviços externos e dos impostos indirectos.

Rentabilidade do capital próprio (RCP):

Este indicador mede a taxa de retorno dos capitais investidos pelos accionistas. Pela sua leitura pode-se concluir se os capitais estão ou não a ser bem aplicados.

$$RCP = \frac{\text{resultado líquido}}{\text{capital próprio}} \times 100$$

3.1 – Análise exploratória dos dados

Na amostra que vai ser estudada incluem-se 400 empresas. Por ser uma amostra de dimensão elevada existem empresas bastante diferentes quer em tamanho (medido através do volume de vendas líquidas) quer noutras características mais difíceis de observar. Seguidamente, vai proceder-se à análise de cada uma das variáveis em relação à simetria, distribuição aproximada e existência de outliers.

Rotação do activo:

Pela análise do histograma desta variável (figura 1), pode constatar-se que ela não é simétrica. A sua assimetria é positiva verificando-se também uma frequência muito elevada na classe central.

O seu afastamento da distribuição Normal é notório. Tal facto pode ser comprovado pelo histograma, onde se representa a distribuição Normal a tracejado, como pelo gráfico de probabilidade normal ("*normal probability plot*"). Neste gráfico (ver figura 2), os pontos devem estar sobre uma linha recta para se poder afirmar que a distribuição está próxima da distribuição Normal. Tal não acontece. É inútil testar a normalidade da distribuição através de testes apropriados uma vez que empiricamente ela é rejeitada.

Como pode ser visto no diagrama "caixa de bigodes", na figura 3, as observações que poderão vir a ser declaradas como outliers são em grande número e situam-se todas na aba direita. Duas empresas apresentam valores muito afastados dos restantes: Cardol, S.A. e Reagro, S.A. (esta com menor afastamento). Este tipo de representação gráfica revela as principais características dos dados: localização, dispersão, (as)simetria, comprimento das caudas e outliers. Para construir a "caixa de bigodes" desenha-se primeiro uma caixa com extremidades nos quartos inferior e superior e uma linha vertical na mediana. São ainda traçadas duas linhas (designadas por bigodes), desde os quartos até aos pontos mais afastados, e que não são outliers, em cada uma das abas.

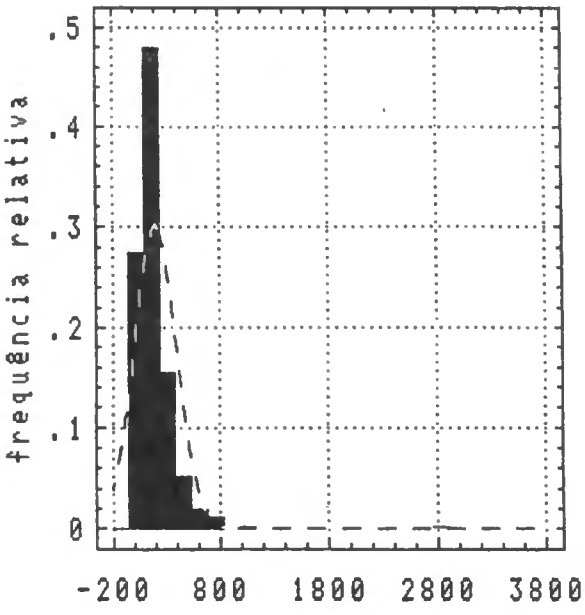

Fig.1-histograma de rotação do activo

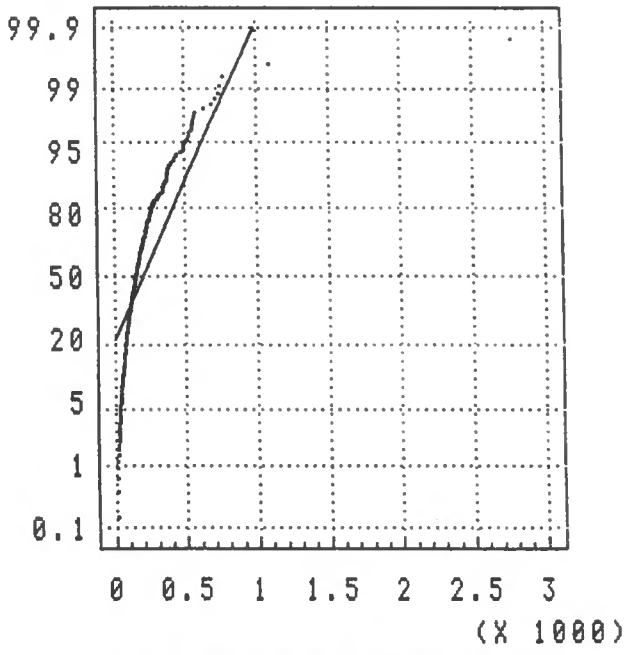

Fig.2-gráfico de probabilidade normal de rotação do activo

Os outliers moderados são representados individualmente por um pequeno ponto ou rectângulo enquanto que os outliers severos representam-se por uma cruz. Tais observações situam-se para além das barreiras de outliers. As barreiras de outliers são definidas por $F_L - \frac{3}{2}d_F$ e $F_U + \frac{3}{2}d_F$ onde F_L e F_U são os quartos e $d_F = F_U - F_L$ representa a dispersão quartal.

A localização dos dados é representada pela linha no interior da caixa, ou seja, a mediana. Se ela se situar a meio, entre o quarto inferior e o quarto superior e se os "bigodes" tiverem aproximadamente o mesmo comprimento, a distribuição dos dados poderá considerar-se aproximadamente simétrica.

O comprimento da caixa mostra a dispersão dos dados, em termos da dispersão quartal e o comprimento dos "bigodes" representa o comprimento das caudas. Como, na sua construção, são utilizadas medidas resistentes à existência de alguns valores perturbadores, este diagrama também é resistente à influência desse tipo de observação.

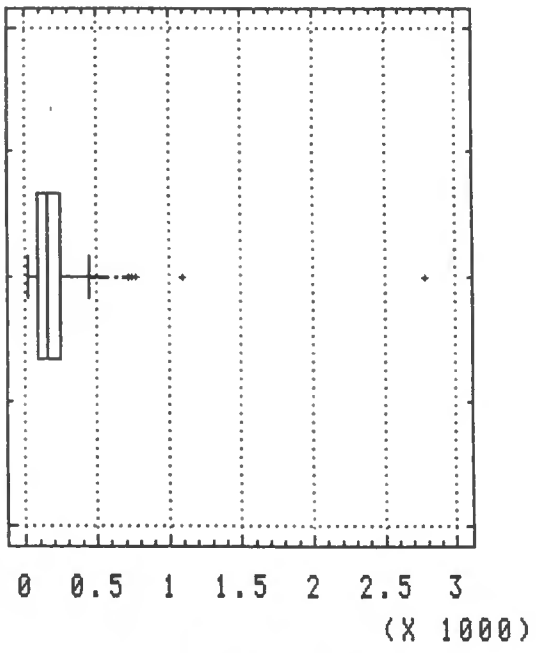

Fig. 3-caixa de bigodes de rotação do activo

Em relação à variável em análise, e através do diagrama anteriormente referido, é visível a sua assimetria positiva uma vez que a mediana se afasta do quarto superior. A cauda direita é relativamente mais comprida que a esquerda dado que o "bigode" esquerdo é bastante curto. Algumas destas características podem ser observadas no histograma da distribuição. No entanto, o diagrama "caixa de bigodes" vem fornecer informação sobre a existência de outliers potenciais que não são visíveis no histograma.

A seguir apresentam-se algumas estatísticas sumárias da variável rotação do activo (quadro 1). Mais uma vez pode ser constatada a assimetria positiva, ou enviesamento à esquerda, uma vez que se verifica que média>mediana>moda.

Para estudar a variabilidade da variável podem utilizar-se medidas de dispersão absoluta. Por exemplo, o desvio padrão, o desvio médio, a dispersão quartal e o pseudo–desvio–padrão–F, entre outras. As medidas anteriores dizem-se "absolutas" por serem expressas na mesma unidade da variável a que se referem.

Estatística	valor
média	198
mediana	152
moda	151
desvio padrão	196
mínimo	17
máximo	2786
1° quartil	89
3° quartil	242
amplitude inter-quartis	153
coeficiente de variação	99 %

quadro 1: estatísticas sumárias de rotação do activo

Há conveniência em utilizar medidas de dispersão independentes da unidade da variável de forma a caracterizar a dispersão relativamente à ordem de grandeza dos valores. Tais medidas designam-se por medidas de dispersão relativa e devem ser usadas sempre que se pretendem comparar diversos conjuntos de dados. A medida de dispersão relativa mais utilizada é o coeficiente de variação, sendo definido pela relação entre desvio padrão e média, em percentagem.

A variável rotação do activo apresenta uma dispersão relativamente elevada uma vez que o coeficiente de variação é de 99%.

Solvabilidade:

A distribuição desta variável não é simétrica. Como se tem média>mediana≥moda, a assimetria é positiva, tal como pode ser constatado no gráfico caixa de bigodes (figura 4), no quadro 2 e no histograma (figura 5). Neste último gráfico é também visível o excesso ou kurtosis assim como o afastamento em relação à normalidade. Para confirmação deste facto observe-se o gráfico de probabilidades normal (ver figura 6). Neste gráfico os pontos que representam as

observações afastam-se visivelmente da recta constituída pelos pontos que são consistentes com a distribuição Normal.

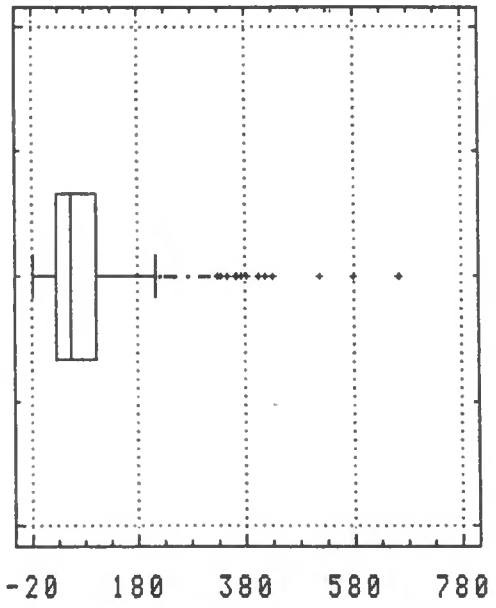

Fig. 4-caixa de bigodes de solvabilidade

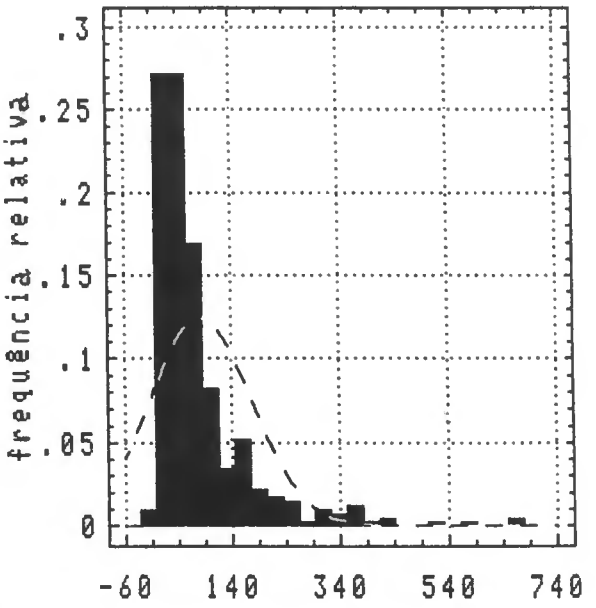

Fig. 5-histograma de solvabilidade

Quanto a outliers potenciais, eles são em grande número. As empresas que, mais tarde, poderão ser consideradas como outliers apresentam valores muito elevados neste rácio. Com efeito, todas as observações potencialmente aberrantes situam-se na aba direita da distribuição. Se existirem outliers eles corresponderão a empresas com uma grande capacidade para responder aos compromissos de longo prazo. As empresas que mais se afastam do conjunto formado pelas restantes são: Inlan, S.A.; Secil, S.A. e Sacor Marítima, S.A.. Estas três empresas caracterizam-se por terem um passivo muito reduzido em relação aos seus capitais próprios.

Estatística	valor
média	82
mediana	53
moda	53
desvio padrão	95
mínimo	–19
máximo	666
1° quartil	25
3° quartil	101
amplitude inter-quartis	76

quadro 2: estatísticas sumárias de solvabilidade

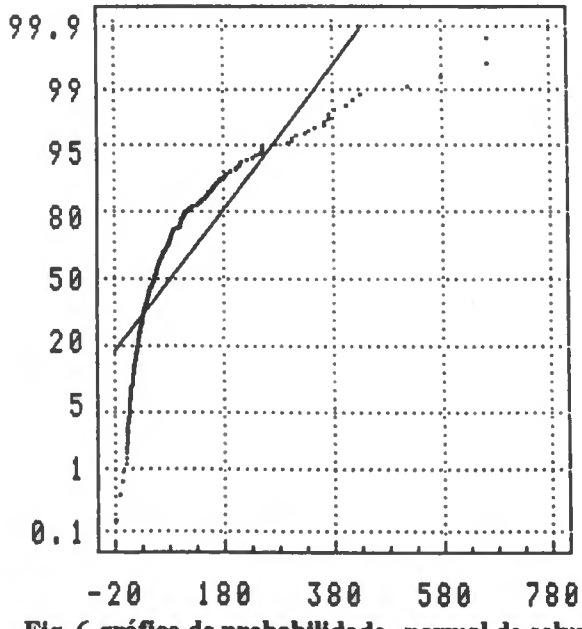

Fig. 6-gráfico de probabilidade normal de solvabilidade

Produtividade do trabalho:

A variável que é definida como a porção de VAB por trabalhador ao serviço de cada empresa em 31 de Dezembro de 1992 apresenta uma distribuição completamente diferente da distribuição Normal. O gráfico de probabilidade normal mostra-o (ver figura 7). A figura constituída pelos pontos referentes às 400

empresas afasta-se completamente da recta que define a distribuição normal, neste tipo de gráfico. É inútil o teste formal à hipótese da normalidade uma vez que tal hipótese, certamente será rejeitada.

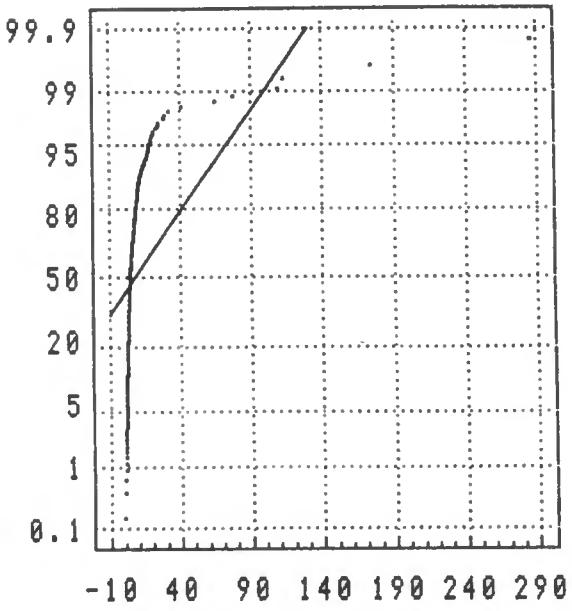

Fig.7-gráfico de probabilidade normal de produtividade do trabalho

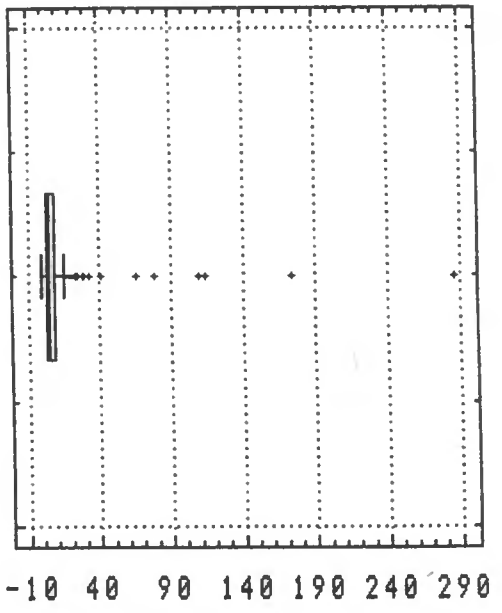

Fig.8- caixa de bigodes de produtividade do trabalho

A assimetria positiva é evidente: média aritmética=8 e mediana=moda=5 (ver quadro 3). Em relação a esta variável existem diversas observações "anormais" na aba direita da sua distribuição.

O diagrama caixa de bigodes, figura 8, identifica um grande número de observações que poderão vir a ser consideradas como outliers. Essas observações situam-se todas na aba direita, portanto apresentam uma produtividade da mão-de-obra muito elevada. As empresas onde esse facto é mais evidente são: United Destillers & Companhia Velha, Lda e Oleocom, SA (ver quadro 4).

Estatística	valor
média	8
mediana	5
moda	5
desvio padrão	19
mínimo	−1
máximo	286
1° quartil	3
3° quartil	8
amplitude inter-quartis	5

quadro 3: estatísticas sumárias de produtividade do trabalho

Empresa	valor do rácio
United Destillers & Companhia Velha, Lda	286,5
Oleocom, SA	174,2
Cepsa, SA	113,9
Transcomércio, SA	109,5
média aritmética	8

quadro 4: empresas com produtividade do trabalho mais elevada

Rentabilidade do capital próprio:

Com esta variável pretende-se conhecer a eficiência da empresa, ou seja, saber se os capitais estão a ser bem aplicados, de uma forma rentável.

Analisar a distribuição deste rácio não parece tarefa fácil. Existe assimetria negativa dado que a média aritmética é inferior à mediana. O afastamento em relação à distribuição normal é evidente, como pode ser visto no histograma (figura 9) ou no gráfico de probabilidades normal (figura 10). Neste último, pode verificar-se a disparidade em relação àquela distribuição. As abas são quase inexistentes e a classe central tem uma frequência muito elevada, ela ronda os

80%. Essa classe inclui o valor zero e valores muito próximos deste. Pode concluir-se que uma grande parte das empresas tem muito fraca rentabilidade dos capitais próprios. Tal facto é consequência de os resultados líquidos serem diminutos em relação à situação líquida. O valor mais observado é zero, uma vez que a moda=0 (quadro 5). Deve ter-se em conta que o ano a que respeitam os dados foi um ano de crise na economia portuguesa.

Estatística	valor
média	3
mediana	6
moda	0
desvio padrão	73
mínimo	−1310
máximo	100
1° quartil	1
3° quartil	15
amplitude inter-quartis	14

quadro 5: estatísticas sumárias de rentabilidade dos capitais próprios

No diagrama "caixa de bigodes"(figura 11), podem ser confirmados alguns dos comentários anteriores, assim como também se pode caracterizar esta variável em termos de observações potencialmente outliers. Assim, é visível a existência de um grande número desse tipo de observação, sendo preocupante apenas uma (quando muito duas) pelo seu grande afastamento em relação ao grupo formado pelas restantes. A empresa Philip Morris, Lda que se dedica ao comércio, apresenta uma rentabilidade do capital próprio no valor de −1310. Este valor, pouco comum, é negativo e extremamente elevado pelo facto de os resultados líquidos serem negativos e o capital próprio ser muito reduzido. A empresa com o segundo valor mais extremo, neste rácio, é Keller Marítima, Lda. Pertence ao sector de serviços e apresenta o valor de −367 como rentabilidade dos capitais

próprios. As razões são as já apresentadas anteriormente, para a empresa com o valor mais extremo.

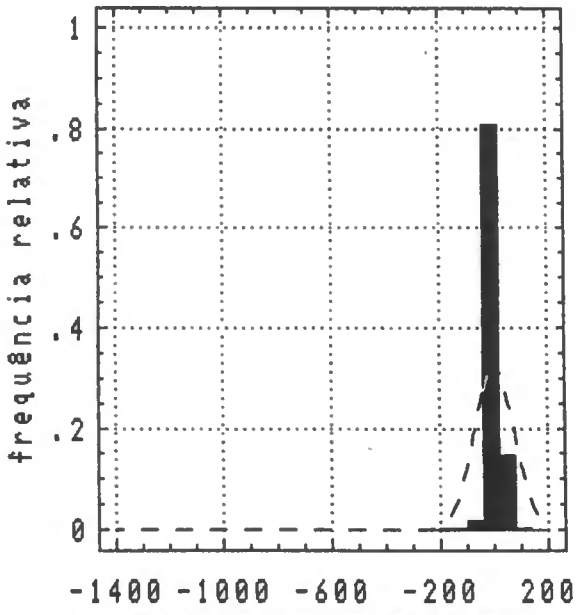

Fig.9-histograma de rentabilidade dos capitais próprios

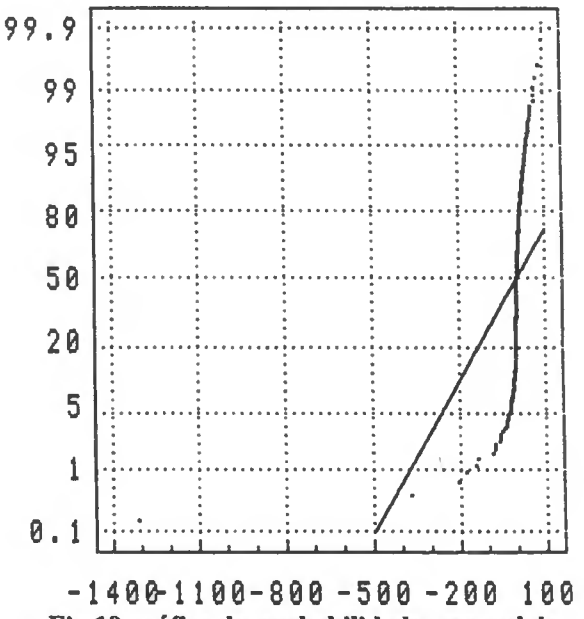

Fig.10-gráfico de probabilidade normal de rentabilidade dos capitais próprios

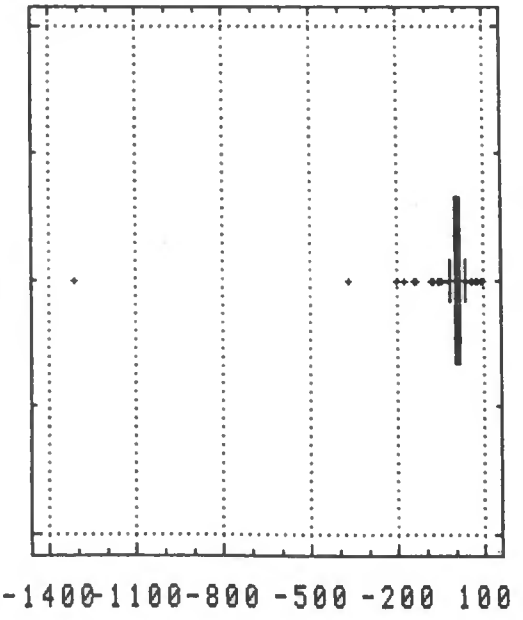

Fig.11-caixa de bigodes de rentabilidade dos capitais próprios

A rentabilidade dos capitais próprios é de entre as quatro variáveis estudadas aquela que apresenta a dispersão mais reduzida.

3.2 – Transformações

É nosso propósito identificar empresas outliers na amostra em estudo. Dado que as variáveis têm uma distribuição difícil de trabalhar, opta-se por operar algumas tranformações. As transformações têm como objectivo aproximar a distribuição das variáveis da distribuição Normal. Esta distribuição, pelas suas características, é de fácil modelização. Relativamente ao estudo de outliers, a distribuição Normal é a que até agora foi mais estudada. Em grande parte dos estudos a suposição básica é a de os dados pertencerem a populações Normais. Também a maioria dos testes de discordância se destinam a dados Normais.

Analisando individualmente cada uma das variáveis e estudando as suas características não é muito difícil determinar o tipo de transformação mais adequado para o objectivo em vista. Uma transformação muito utilizada para eliminar a assimetria é a transformação logarítmica natural (ou decimal, cujos efeitos são idênticos).

No entanto, tal transformação não é cura universal para a assimetria. Se a distribuição for assimétrica negativa a transformação logarítmica só vai agravar essa situação, produzindo uma distribuição ainda mais assimétrica. No caso de a assimetria ser positiva, mas pouco acentuada, a transformação logarítmica poderá ser demasiado potente. A assimetria positiva da distribuição original poderá ser transformada em assimetria negativa.

No caso de a variável original apresentar assimetria positiva muito forte, a transformação logarítmica poderá ser insuficiente para eliminar por completo esse tipo de assimetria. Em tais casos, é necessária a utilização de outras transformações.

As transformações da classe potência são de uso frequente quer pela facilidade de utilização quer pelos bons resultados que proporcionam. A designação de transformações potência resulta do facto de estar envolvida a elevação dos valores da variável a uma potência λ.

Os efeitos destas transformações dependem do valor de λ. Assim, para $\lambda=1$ os dados não sofrem alteração. Para as situações em que existe assimetria negativa, e se pretende eliminá-la ou reduzi-la, devem ser utilizados valores de λ superiores à unidade. Quanto maior for λ, maior será o efeito.

Para eliminar ou reduzir a assimetria positiva os dados devem ser elevados a uma potência $\lambda<1$. Quanto menor for a potência mais forte é o efeito. A raíz quadrada, $\lambda=1/2$, é uma transformação muito usada para reduzir a assimetria positiva se o logarítmo se revela excessivo. Ela "encurta" menos a aba direita que a transformação logarítmica.

A transformação logarítmica é um caso particular da transformação potência quando $\lambda=0$. Quando a assimetria positiva é muito acentuada de modo que a logaritmização é insuficiente, pode ser aplicada a transformação definida como o simétrico do inverso da raiz quadrada.Tal transformação consiste em elevar todos os valores a $\lambda=-1/2$ e depois multiplicá-los por -1.

Há necessidade de mudar o sinal dos valores transformados sempre que se utiliza uma transformação do tipo potência com expoente negativo. Elevando os valores a um expoente negativo os menores ao serem transformados ficam maiores e os maiores são reduzidos. A multiplicação por (-1) serve para manter a ordem inicial dos dados.

Perante uma distribuição assimétrica, e depois de identificado o tipo de assimetria, devem fazer-se várias tentativas na procura da transformação que proporcione o efeito desejado. A análise gráfica é um instrumento de grande utilidade, nesta fase. Testes formais deverão ser utilizados em caso de dúvida ou para obter confirmação de algumas características da distribuição. Experiência e um pouco de sorte têm aqui papel fundamental.

Na prossecução dos nossos objectivos, estamos neste momento preocupados com a determinação da transformação que aproxime a distribuição das variáveis em análise da distribuição normal. Como foi descrito no ponto anterior, um dos grandes problemas das variáveis em estudo consiste na ausência de simetria nas

respectivas distribuições. Como a assimetria das variáveis parece ser a principal causa para o distanciamento em relação à distribuição Normal vamos, através de transformações adequadas, tentar eliminar esse problema. O método utilizado baseia-se no seguinte:

--análise pormenorizada da distribuição de cada uma das variáveis

--detecção do tipo de assimetria existente

--aplicação da transformação adequada para o tipo de assimetria em causa

--análise da distribuição transformada quanto à normalidade

--repetição do processo até ser obtida uma distribuição relativamente próxima da Normal.

No terceiro passo do processo descrito anteriormente, a escolha do valor de λ, inferior ou superior a 1, depende da assimetria da variável ser, respectivamente, positiva ou negativa. O valor de λ a utilizar na transformação das variáveis foi escolhido de entre os que pertencem ao intervalo que contém o valor "óptimo", ou seja, aquele que normaliza a distribuição da variável. Não existe certeza de que o valor usado seja o óptimo, no entanto se não o for ele estará bem próximo do óptimo.

Depois de eliminada a assimetria não existem garantias de que as distribuições resultantes sejam aproximadamente normais. Resta a certeza de que uma distribuição simétrica se identifica mais com a distribuição Normal do que se não existir simetria.

O rácio rotação do activo é transformado elevando todos os valores observados ao expoente λ=0,1. Tal transformação é escolhida pelo facto de inicialmente existir assimetria positiva pouco acentuada. A distribuição resultante é simétrica, uma vez que se obtém igualdade entre média, mediana e moda (ver quadro 6 e figura 12). O seu valor é de 1,65. Naturalmente a dispersão dos dados diminuíu significativamente sendo o coeficiente de variação de apenas 8%, aproximadamente.

Estatística	valor
média	1,65
mediana	1,65
moda	1,65
desvio padrão	0,13
mínimo	1,33
máximo	2,21
1° quartil	1,57
3° quartil	1,73
amplitude inter-quartis	0,16
coeficiente de variação	7,57 %

quadro 6: estatísticas sumárias de RA transformado

A redução, em número, de outliers potenciais é evidente passando agora a cifrar-se em apenas dois (figura 13). Esta redução parece-nos natural pelo facto de a transformação operada sobre os dados reduzir a sua escala. As empresas que poderão ser consideradas como outliers são Cardol, SA e Reagro, SA, as duas empresas atrás indicadas como as mais "aberrantes".

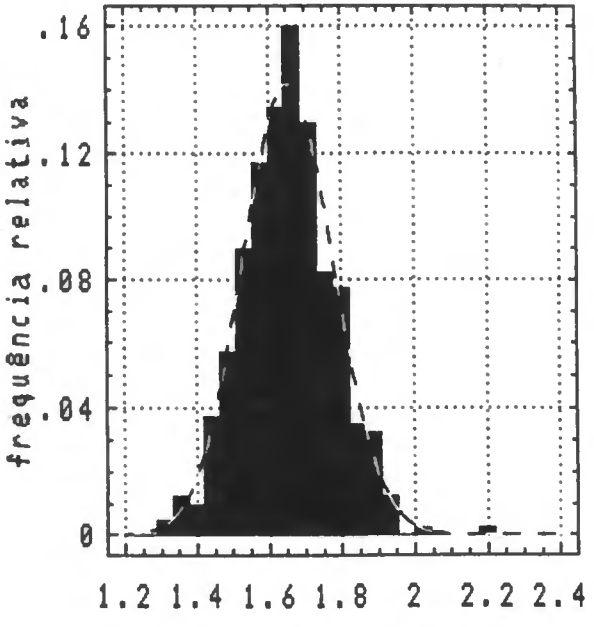

Fig. 12-histograma de RA tranformada

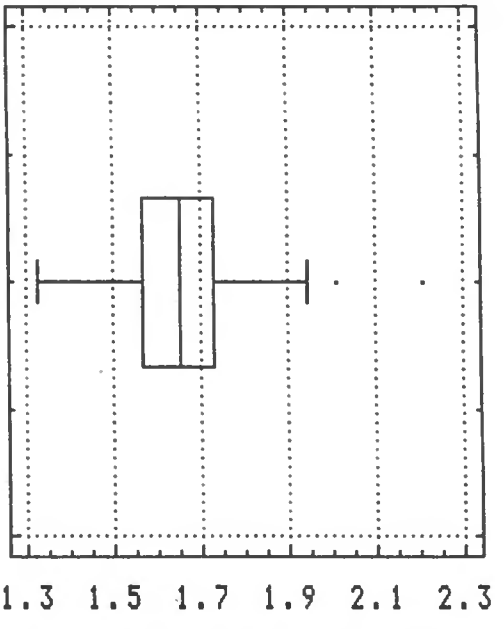

Fig. 13-caixa de bigodes de RA transformada

Estas empresas apresentam um valor muito elevado neste rácio pelo facto de os seus activos líquidos serem muito pequenos.

No gráfico de probabilidade normal (ver figura 14), os pontos representativos das observações estão dispostos sobre uma linha recta correspondente à normalidade da variável em causa. Apenas um ponto se afasta dessa disposição e situa-se na aba direita. Mais à frente serão efectuados testes formais quanto à normalidade da distribuição de rotação do activo.

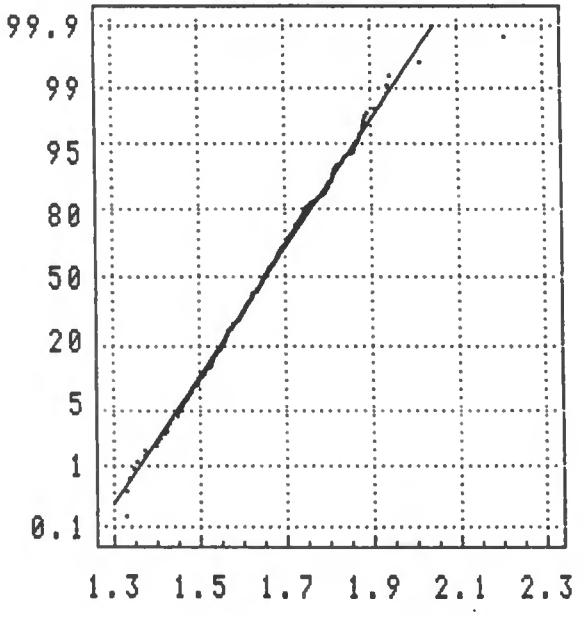

Fig.14-gráfico de probabilidade normal de RA transformada

A variável que representa a solvabilidade das empresas em estudo apresenta assimetria positiva relativamente acentuada. Para tornar esta variável simétrica há necessidade de utilizar uma transformação potência de valor negativo. Para além disso é necessário, antes de mais, que todos os valores sejam positivos. Assim adiciona-se uma quantidade positiva a todas as observações, neste caso o seu valor é 21. Depois de aplicada a transformação, e por forma a manter a ordenação inicial dos dados, terá que ser trocado o sinal das observações.

A distribuição do rácio da solvabilidade fica simétrica depois de os seus valores serem elevados a (−1/5).Mais especificamente, a transformação a utilizar é

$$-[(SV+21)^{(-1/5)}].$$

Deste modo, a distribuição resultante tem média=mediana=moda=−0,42. A dispersão foi reduzida, o coeficiente de variação é de 15%. Tal redução provocou o desaparecimento de alguns outliers potenciais. Com efeito, depois de efectuada a transformação permanecem apenas três observações como potencialmente aberrantes. Tais observações representam as empresas com menor solvabilidade. Os valores do rácio para essas empresas são negativos. Não existe capacidade para satisfazer os compromissos a longo prazo através dos capitais próprios. Nesta situação encontram-se as sociedades Henkel Hibérica, SA , Sol-Aves, SA e Portugália, SA.

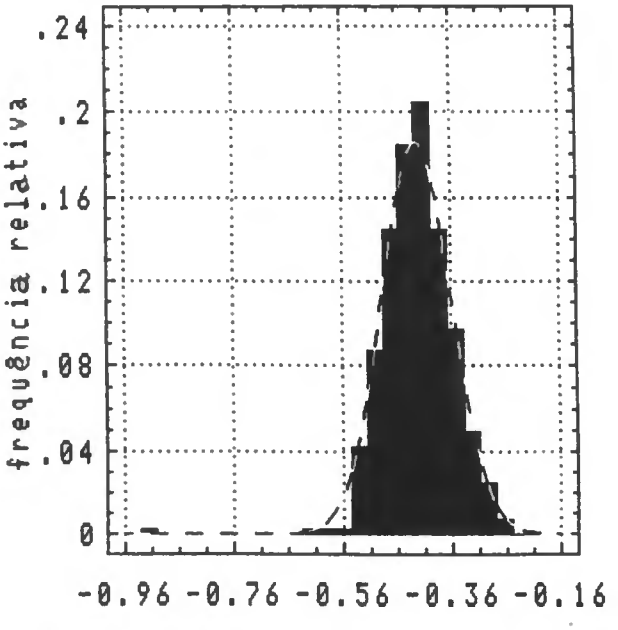

Fig. 15-histograma de solvabilidade transformada

A produtividade do trabalho também precisa ser transformada por forma a ser obtida uma distribuição simétrica. Este rácio apresenta assimetria positiva e um afastamento claro em relação à distribuição normal.

Dada a existência de valores negativos para este rácio, adicionamos uma unidade a todas as observações de modo a que todas apresentem valores positivos. Estamos assim em condições de aplicar qualquer transformação.

A simetria obtém-se depois de ser aplicada a seguinte transformação:

$$-[(PT+1)^{(-0,4)}].$$

A distribuição transformada tem média = mediana = moda = −0,494. O efeito da transformação é visível por comparação dos histogramas correspondentes à situação anterior (figura 16) e posterior à tranformação (figura 17).

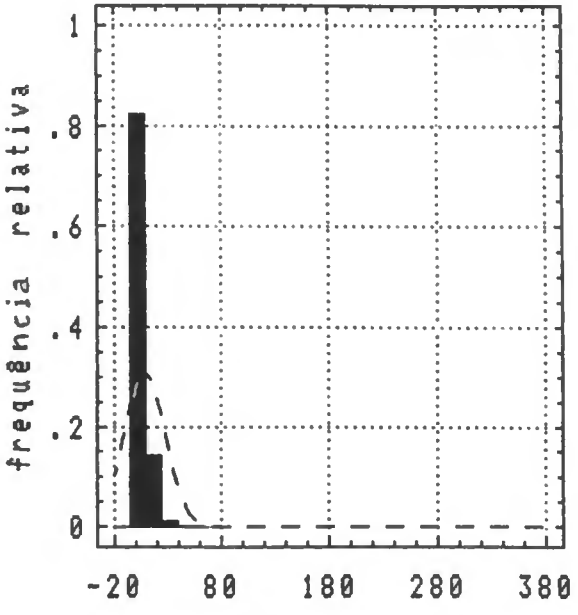

Fig. 16-histograma de produtividade do trabalho

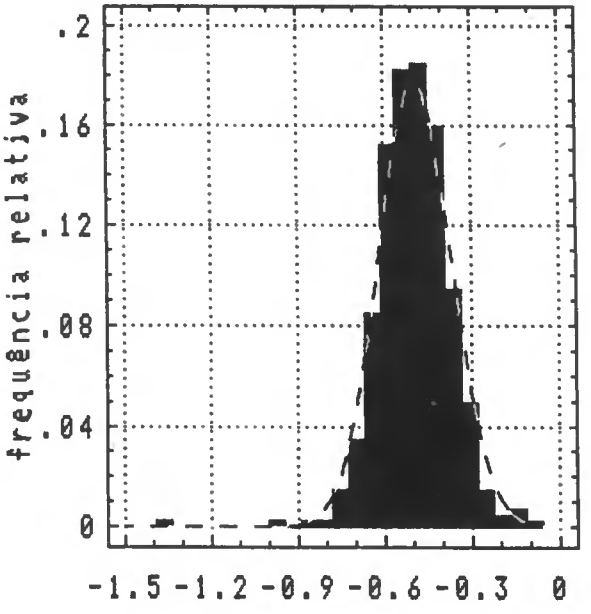

Fig. 17-histograma de PT transformada

Depois de os dados sofrerem a transformação permanecem ainda algumas dificuldades nas abas da distribuição que certamente não serão de grande gravidade. Quanto a candidatos a outliers, inicialmente eles eram muitos mas todos eles se situavam na parte direita da distribuição. Depois da transformação eles ficaram reduzidos a nove.

As quatro seguintes empresas serão, possivelmente, outliers inferiores:

.Ticket Restaurante de Portugal, SA

.Farsul, CRL

.Farbeira, CRL

.Alicoop, CRL.

Todas estas empresas se distanciam das restantes por terem uma produtividade da mão-de-obra muito reduzida. Como este rácio é a relação entre VAB e número de trabalhadores ao serviço, ele poderá apresentar valores muito reduzidos pelo facto de o VAB ser reduzido ou por existirem muitos empregados ao serviço da empresa.

Em relação à empresa Ticket Restaurante de Portugal, SA a razão para o aparecimento da produtividade do trabalho muito reduzida prende-se com o facto de o VAB apresentar um valor negativo. Isto é consequência de os consumos intermédios terem um valor muito elevado, ficando muito perto do valor das vendas. O valor acrescentado bruto representa a contribuição de cada empresa para a economia do país, ora sendo esse valor negativo, a empresa não contribui para o produto da economia.

Quanto às restantes três empresas, elas têm em comum o facto de serem grossistas organizadas em forma de cooperativas. Farsul e Farbeira do sector farmacêutico e Alicoop trabalha com produtos alimentares. A produtividade do trabalho é reduzida, nestas três empresas, porque os VABs respectivos apresentam valores muito pequenos, mas positivos. Quanto a trabalhadores, Farsul empregava 183, Farbeira 64 e Alicoop 146.

Para outliers superiores existem cinco candidatas:

.United Destillers & Companhia Velha, Lda

.Oleocom,SA

.Cepsa, SA

.Transcomércio, SA

.Cruz & Cª,Lda

A razão de as empresas anteriores terem, neste rácio, valores muito elevados relaciona-se com o reduzido número de trabalhadores ao seu serviço. As duas primeiras têm apenas três trabalhadores e as duas últimas têm, respectivamente, sete e trinta e três trabalhadores. Isto está relacionado com o facto de aquelas empresas, à excepção da Cepsa, serem grossistas e de este tipo de empresa empregar pouca mão-de-obra. United Destillers & Companhia Velha e Cruz & Cª são grossistas de vinhos e bebidas alcoólicas, Transcomércio é grossista de produtos alimentares enquanto que Oleocom é grossista de cereais, comercializando oleaginosas.

Em relação à Cepsa, o valor elevado do rácio advém do facto de ter um VAB bastante elevado, uma vez que em termos de pessoal utilizado ele ronda os 100 trabalhadores.

Comparações entre a produtividade de empresas de diferentes sectores devem ser feitas com cuidado. Empresas de capital intensivo terão, em princípio, uma produtividade do trabalho superior à de empresas de mão-de-obra intensiva. É o que acontece com as empresas anteriores. À excepção da distribuidora de combustíveis Cepsa, as restantes pertencem ao sector de distribuição alimentar, sector esse que utiliza muito capital e poucos recursos humanos.

Ao analisar os resultados anteriores deve ter-se em conta que as empresas que apresentam produtividade do trabalho elevada não são obrigatoriamente muito eficientes na utilização dos recursos humanos, uma vez que pertencem a um sector, que por si só, apresenta produtividades do trabalho elevadas pelo facto de necessitar de pouco pessoal no seu funcionamento.

O último rácio em estudo, a rentabilidade dos capitais próprios, é em nosso ver, o de análise mais difícil. Com efeito, e como foi visto no ponto anterior, esta variável apresenta assimetria negativa pouco acentuada, mas a sua classe central apresenta uma frequência extremamente elevada. Tal facto vai complicar, de certo modo, o nosso trabalho. Na tentativa de simetrizar a distribuição e ao mesmo tempo aproximá-la da distribuição Normal, foram ensaiadas várias transformações. A distribuição da variável que representa a rentabilidade dos

capitais próprios das empresas, em análise, fica aproximadamente simétrica com a seguinte transformação:

$$(RCP + 1310)^{0.7}$$

Antes de aplicar a transformação foi adicionado o valor 1310 a cada uma das observações, por forma a que todos os valores sejam maiores ou iguais a zero.

A distribuição da variável transformada é simétrica, dado que em termos das medidas de localização, se passa para uma situação em que média, mediana e moda apresentam o mesmo valor, 152.

A dispersão da variável após a transformação ficou muito reduzida uma vez que apresenta um coeficiente de variação de cerca de 5%. As observações estão muito concentradas na classe central (ver figura 18). Tal classe tem uma frequência de cerca de 90%. O primeiro quartil assume o valor de 152,2 enquanto que o terceiro quartil tem o valor de 153,3. Quer isto dizer que metade das observações se situam entre aqueles dois valores.

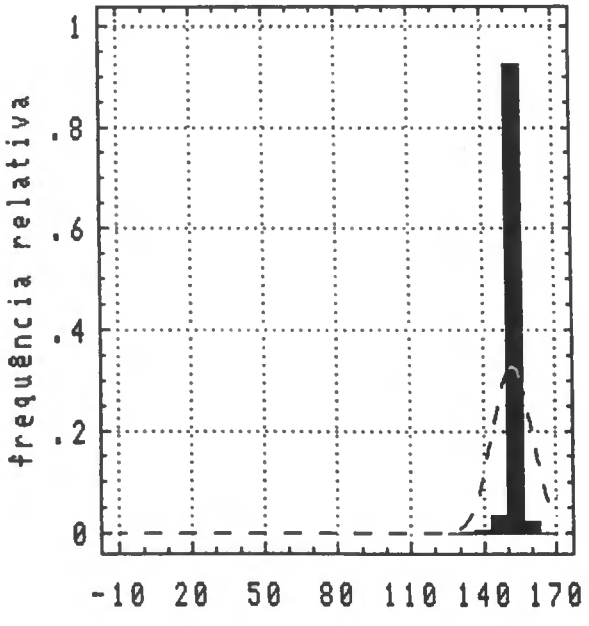

Fig.18-histograma de rent. cap. próprios transformada

Em relação a empresas que poderão vir a ser consideradas como outliers, depois de efectuada a transformação, não houve alteração significativa. Continua a existir um grande número de empresas potencialmente aberrantes e situam-se em

ambas as abas. Na aba esquerda elas são em maior número. As duas empresas com maior afastamento das restantes continuam a ser Philip Morris (Portugal), Lda e Keller Marítima, Lda.

3.3 – Normalidade

Nesta altura, as distribuições das variáveis em estudo são aproximadamente simétricas. Tal só foi possível depois de aplicadas transformações da classe potência aos dados originais. Neste momento estamos particularmente interessados em saber se tais distribuições se aproximam da normalidade. Para isso vão realizar-se testes à normalidade univariada de cada um dos rácios.

Para testar a normalidade de cada uma das variáveis são utilizados os testes do Qui-Quadrado e o de Kolmogorov-Smirnov. Em ambos os testes considera-se que, se o "p–value" é inferior a 5% a distribuição Normal não é indicada para descrever o comportamento da variável em causa. A análise gráfica é também aqui de grande utilidade pois proporciona uma visão clara da direcção em que a normalidade pode ser obtida.

Quanto à rotação do activo transformado, e por observação do respectivo histograma (figura 12), é visível a sua identificação com a distribuição Normal. Com efeito, qualquer um dos dois testes referidos anteriormente não rejeitam a hipótese da normalidade para aquela variável. O teste do Qui–Quadrado não rejeita a normalidade do rácio transformado com um "p–value" de cerca de 82% enquanto que esse valor sobe para perto de 93% se usado o teste de Kolmogorov–Smirnov.

A solvabilidade é uma variável que, pela análise gráfica, evidencia o afastamento em relação à normalidade. Em ambos os testes aqui utilizados a normalidade é rejeitada pois o valor do "p–value" é 0. Efectuada a transformação adequada e testada a hipótese da normalidade, esta não é rejeitada pelo teste do

Qui–Quadrado e também pelo teste K–S. No primeiro o "p–value" é de 48% e no segundo de 63%.

A produtividade do trabalho tem uma distribuição completamente distinta da distribuição normal. Tal facto já foi devidamente evidenciado anteriormente. A distribuição resultante depois de efectuada a transformação potência conveniente tem, como pode ser visto na figura 17, uma distribuição muito próxima da normal. Com efeito, os testes formais à hipótese da normalidade aproximada confirmam o resultado da análise gráfica. O teste do Qui-Quadrado apresenta um "p–value" de 51% enquanto que o valor correspondente para o teste de K–S é de 45%.

A variável que representa a rentabilidade dos capitais próprios de cada uma das empresas em análise, depois de transformada, é já simétrica, como se viu no ponto anterior. Porém, a sua distribuição apresenta um grande afastamento em relação à distribuição normal. Tal afastamento é devido à grande frequência da classe central das observações. De facto, a classe central engloba praticamente 90% das observações, sendo as restantes repartidas pelas outras classes. Como consequência, a distribuição tem abas muito curtas, o que a par do facto de a classe central ser muito elevada conduz ao distanciamento notório em relação à distribuição Gaussiana. Os testes formais à normalidade rejeitam essa hipótese visto apresentarem um "p–value" igual a zero. Não se conseguiu obter, neste caso, uma distribuição aproximadamente Normal.

Nesta altura já foi realizada a análise exploratória das variáveis consideradas. Dado o seu afastamento em relação à distribuição Gaussiana efectuaram-se transformações para aproximar a distribuição das variáveis daquela distribuição. Tal foi conseguido para três das quatro variáveis: rotação do activo, solvabilidade e produtividade do trabalho. O mesmo não foi conseguido para a rentabilidade dos capitais próprios. As observações candidatas a outliers foram também identificadas através da análise gráfica, nomeadamente do diagrama caixa de bigodes. Tais observações vão ser objecto de alguns testes formais para averiguar se podem ser consideradas como aberrantes.

CAPITULO 4 – IDENTIFICAÇÃO DAS EMPRESAS "OUTLIERS"

No capítulo anterior foram identificadas as empresas potencialmente outliers relativamente a cada uma das variáveis em análise. Inicialmente essas observações foram identificadas pelas variáveis originais e depois pelas variáveis transformadas. Assistiu-se, de uma maneira geral, a uma redução do número de outliers potenciais depois de efectuadas as transformações. É importante referir que a identificação de tal tipo de observação foi feita utilizando apenas instrumentos gráficos.

A análise gráfica é de grande interesse e utilidade neste, como noutros tipos de análises. Não pode, no entanto, deixar de ser acompanhada de procedimentos formais.

Pretende-se agora testar, formalmente, essas observações por forma a certificarmo-nos se elas são de facto empresas outliers.

No quadro seguinte(quadro 7) apresentam-se as empresas identificadas como outliers potenciais na análise exploratória dos dados. Tal identificação foi feita pelas variáveis depois de serem transformadas. O seu número de ordem, tendo em conta o volume de vendas, a variável que identificou cada empresa e o tipo de outlier, são informações incluídas nesse quadro. Querendo saber se as empresas anteriores são efectivamente outliers vamos, seguidamente, testar formalmente essa hipótese.

A hipótese nula consiste na afirmação de que os dados foram retirados de uma população normal e não existem observações discordantes na amostra. Em contraste, na hipótese alternativa uma ou mais observações são outliers. Para testar a hipótese nula são utilizadas algumas das estatísticas de teste descritas no primeiro capítulo e destinadas a populações normais.

Empresa(nº de ordem)	variável que a identificou	tipo de outlier
United Destillers & Comp.Velha, Lda(220)	PT	superior
Oleocom, SA(40)	PT	superior
Cepsa, SA(23)	PT	superior
Transcomércio, SA(50)	PT	superior
Cruz & Cª, Lda(176)	PT	superior
Ticket Restaurante de Portugal, SA(86)	PT	inferior
Farsul,CRL(59)	PT	inferior
Farbeira, CRL(320)	PT	inferior
Alicoop, CRL(187)	PT	inferior
Henkel Hibérica,SA(191)	SV	inferior
Sol-Aves(388)	SV	inferior
Portugália(370)	SV	inferior
Reagro, SA(28)	RA	superior
Cardol,SA(197)	RA	superior
Philip Morris,Lda(58)	RCP	inferior
Keller Marítima,Lda(337)	RCP	inferior

quadro 7: empresas identificadas como outliers potenciais na análise exploratória

4.1 – Detecção univariada de empresas outliers

A empresa United Destillers & Companhia Velha, Lda, com o número de ordem 220, foi identificada pelo rácio produtividade do trabalho, depois de transformado, como a mais extrema na aba direita da sua distribuição. Para testar a hipótese de esta observação ser efectivamente outlier utiliza-se o teste T_1(com ambos os parâmetros desconhecidos), que proporciona o seguinte resultado:

$$T_1 = (x_{(n)} - \bar{x}) / s = (-0{,}103894 + 0{,}494213) / 0{,}124708 = 3{,}13$$

Valores críticos para este teste encontram-se na tabela XIIIa de Barnett e Lewis(1994, pág. 485). Os valores aí tabelados destinam-se a amostras de dimensão inferior a 120. Surge aqui, e provavelmente mais tarde, uma dificuldade. A nossa amostra tem 400 observações e os valores críticos tabelados não contemplam esse valor. Diversas tabelas ficam muito aquém desse valor.

Neste caso, esta dificuldade é ultrapassável. Os valores críticos para uma amostra de dimensão 120 são 3,27 e 3,66 para 5% e 1%, respectivamente. Como o valor observado para T_1 é inferior a estes valores, e eles são crescentes à medida que aumenta a dimensão da amostra, e supondo que essa tendência não se altera, não podemos rejeitar a hipótese inicial. Tal empresa não é considerada como outlier.

Foram identificadas mais quatro empresas como outliers potenciais, na aba direita. Dado que tais empresas, apesar de distantes não se afastam tanto das restantes com a empresa United Destillers, considerada anteriormente como não sendo outlier, também aquelas não poderão ser consideradas como tal.

O teste

$$T_6 = \frac{n^{1/2} \sum_{j=1}^{n} (x_j - \bar{x})^3}{\left[\sum_{j=1}^{n} (x_j - \bar{x})^2 \right]^{3/2}}$$

representa a assimetria da amostra. Para mais que um outlier é aplicado sucessivamente. Com a sua aplicação é testada a maior ou menor observação, consoante o sinal de $\sum_{j=1}^{n} (x_j - \bar{x})^3$ seja positivo ou negativo. Quanto maior for o valor de T_6, em valor absoluto, maior é a tendência da observação em teste para ser outlier.

Nesta variável tem-se $\sum_{j=1}^{400} (x_j - \bar{x})^3 = -0{,}568$. Por ser negativo vai ser testada a menor observação, ou seja, a de ordem 86, correspondente à empresa

Ticket Restaurante de Portugal, SA. O valor observado para a estatística de teste é $T_6 = -0{,}734$. O valor do teste é maior, em termos absolutos, que os valores críticos correspondentes em Barnett e Lewis(1994, pág.499, tabela XXa), pelo que se considera aquela empresa como outlier.

Excluíndo a empresa considerada anteriormente como outlier, é novamente aplicado o teste. Como $\sum_{j=1}^{n} (x_j - \bar{x})^3$ é positivo, depois de excluir a empresa Ticket Restaurante de Portugal, SA, vai ser testada a empresa com o maior valor neste rácio tranformado, ou seja, United Destillers. Obtém-se $T_6 = 0{,}063$, que por ser um valor muito reduzido, em relação aos valores críticos, leva à não rejeição da hipótese de se estar perante uma observação "normal". Este resultado já tinha sido obtido aquando da aplicação de T_1.

Pode-se afirmar que, para o rácio produtividade do trabalho, apenas se considera existir uma observação discordante. Ela corresponde à empresa Ticket Restaurante de Portugal,SA.

a análise exploratória efectuada à variável solvabilidade, depois de transformada, foram identificadas três empresas como potencialmente discordantes. Essas empresas têm os números de ordem 191, 388 e 370 e são, respectivamente, Henkel Hibérica, Sol-Aves e Portugália. Elas têm os menores valores nesse rácio.

teste

$$T_7 = \frac{n \sum_{j=1}^{n} (x_j - \bar{x})^4}{\left[\sum_{j=1}^{n} (x_j - \bar{x})^2 \right]^2},$$

representa a kurtosis da amostra e ao ser utilizado é testada a observação que mais se afasta da média aritmética amostral. É robusto em relação a problemas de "masking", por isso é indicado para ser usado sucessivamente se existir mais que um outlier. Valores elevados da estatística indicam discordância da observação em análise.

O teste para a observação de ordem 191, a mais extrema, assume o valor de $T_7 = 11{,}555$. Tal valor é muito elevado e muito superior aos valores críticos tabelados em Barnett e Lewis(1994, pág 499 tabela XXb), situados entre três e quatro. Assim, aceita-se a hipótese de a empresa Henkel Hibérica ser discordante.

Este teste pode ser utilizado sucessivamente, bastando para tal a exclusão no seu cálculo da observação considerada como discordante. Obtem-se $T_7 = 2{,}95$ depois de excluída a empresa já considerada como discordante. Isso significa que a segunda empresa mais extrema, Sol-Aves não é discordante dado que o valor observado pelo teste é inferior aos valores críticos.

O teste T_6 é também utilizado para confirmar a discordância das observações identificadas como outliers potenciais pelo rácio da solvabilidade. Neste caso, o sinal de $\sum_{j=1}^{400} (x_j - \bar{x})^3$ é negativo, sendo portanto testada a observação de menor valor, ou seja, a que corresponde à empresa Henkel Hibérica, SA.

Obtém-se o seguinte valor: $T_6 = -1{,}116$. Este valor é muito elevado, em termos absolutos, em relação aos valores críticos, pelo que se aceita a discordância dessa observação.

Ao ser excluída a observação anterior, o valor de $\sum_{j=1}^{n} (x_j - \bar{x})^3$ é positivo, pelo que deverá ser testada a observação com o maior valor neste rácio. Tal observação não tinha sido identificada como candidata a outlier. Se aplicado o teste aceita-se a hipótese da maior observação não ser discordante.

Assim, para este rácio, a única empresa considerada como outlier é a sociedade Henkel Hibérica.

Em relação ao rácio rotação do activo, foram identificadas como outliers potenciais as empresas Cardol, SA e Reagro, SA, com os números de ordem 197 e 28, respectivamente. São as empresas que apresentam os valores mais elevados neste rácio.

Utilizando T_6 está-se a testar a maior observação, ou seja a empresa Cardol, SA. O valor do teste é $T_6 = 0{,}187$. Este valor é menor que os valor críticos para

n=500 a 1% de significância. Como os valores tabelados são decrescentes com o aumento da dimensão da amostra, conclui-se que para a dimensão da nossa amostra a relação continua a verificar-se. Logo, a empresa em teste não é considerada outlier. A outra empresa candidata a discordante não é efectivamente outlier por se afastar ainda menos das restantes empresas que a empresa Cardol,SA.

Em relação ao rácio de rentabilidade, as empresas mais extremas são Philip Morris, Lda e Keller Marítima, Lda, apresentando valores de rentabilidade muito reduzidos. Como a distribuição desta variável não é conhecida e se apresenta muito distinta das distribuições mais conhecidas e utilizadas neste tipo de análise, Normal, Gama ou Exponencial, estamos perante a impossibilidade de aplicar os testes que têm vindo a ser aplicados até aqui. Optamos então pela utilização de procedimentos não paramétricos para identificar observações outliers.

Pela análise gráfica, como já foi referido, identificam-se algumas observações com valores aberrantes mas apenas duas têm fortes possibilidades de virem a ser consideradas outliers. Utlizando o teste não paramétrico para grandes amostras definido por Walsh [ver por exemplo, Barnett e Lewis(1980), pág.284], rejeita-se a hipótese nula da não existência de outliers a um nível de significância $\alpha=1/B^2$ se $Z<0$, com

$$Z=x_{(k)}-(1+A)x_{(k+1)}+Ax_{(N)},$$

onde $N=k+(2n)^{1/2}$, n é o tamanho da amostra, k é o número de outliers préviamente definido e A é dado por:

$$A=\frac{1+B\sqrt{\left\{(\sqrt{2n}-B^2)/(\sqrt{2n}-1)\right\}}}{\sqrt{2n}-B^2-1}.$$

Fazendo k=2 outliers e usando um nível de significância de 1% tem-se A=0,20763 o que faz com que Z=−11,03, levando à rejeição da hipótese nula. Se for k=1 também se rejeita a hipótese nula, mas tal não acontece se forem considerados três outliers inferiores. Conclui-se pela aceitação das empresas Phillip Morris(58) e Keller Marítima(337) como observações outliers.

Depois de terem sido utilizados alguns testes univariados para determinar a discordância das observações candidatas a outliers, as empresas constantes no quadro 8 são consideradas outliers.

Empresas	variável que a identificou	teste usado
Henkel Hibérica	SV	T_7 e T_8
Ticket Rest. de Portugal, SA	PT	T_7 e T_8
Philip Morris(Portugal)	RCP	teste de Walsh
Keller Marítima, Lda	RCP	teste de Walsh

quadro 8: empresas consideradas outliers

As restantes empresas não se consideram outliers. Para algumas foram realizados testes, para outras isso não foi necessário dados que essas empresas se afastavam menos das restantes do que aquelas já consideradas como não discordantes. Seria inútil testar a sua discordância, pois a sua rejeição é clara. Nos testes operados às empresas constantes do quadro 9, a sua discordância não foi aceite.

Empresa	variável que a identificou	teste usado
Cardol,SA	RA	T_7
Sol-Aves	SV	T_7 e T_8
United Destillers & comp.Velha,Lda	PT	T_1

quadro 9: empresas consideradas como não sendo outliers

Cada empresa foi observada em termos de quatro variáveis: rotação do activo, solvabilidade, produtividade do trabalho e rentabilidade dos capitais próprios. Essas variáveis já foram analisadas individualmente. Em relação a cada uma delas foram identificadas as empresas candidatas a outliers. Depois de

realizar alguns testes ficou claro quais as empresas que se consideram como discordantes em relação à distribuição Normal subjacente a algumas das variáveis.

Seguidamente, passa-se à análise das observações em termos multivariados. Isto significa que, em vez de considerarmos separadamente a informação dada por cada uma das variáveis em análise, vamos considerar cada empresa como sendo uma observação multivariada, de dimensão 4, e que por isso é caracterizada pela informação dada pelos quatro rácios, considerados em simultâneo.

4.2 – Detecção multivariada de empresas outliers

A análise de componentes principais é um método usado neste tipo de situações. Um dos principais objectivos com que este tipo de análise é usado diz respeito à redução da dimensão dos dados. Neste trabalho esse não é o nosso objectivo principal, apesar de não ser indesejado o seu alcance.

Pretendemos identificar empresas outliers. A análise de componentes principais parece ser um meio para atingir o nosso objectivo.

O primeiro passo para realizar a análise de componentes principais, consiste no cálculo e análise da matriz de correlações entre as variáveis. Se as correlações entre as variáveis forem reduzidas, não fará muito sentido a utilização deste tipo de análise se o principal objectivo for a redução da dimensão da amostra. Esse objectivo, certamente não será alcançado.

No nosso caso, mesmo que as correlações entre as variáveis sejam reduzidas, este tipo de análise faz sentido se pensarmos na sua função. A análise de componentes principais permite a identificação de outliers, daí a sua utilização neste trabalho. Seguidamente, apresenta-se a matriz de correlações entre as quatro variáveis em análise(depois de transformadas): rotação do activo(RA), solvabilidade(SV), produtividade do trabalho(PT) e rentabilidade dos capitais próprios(RCP).

	RA	SV	PT	RCP
RA	1,0000	–0,3225	–0,1241	–0,0208
SV	–0,3225	1,0000	0,1380	0,1340
PT	–0,1241	0,1380	1,0000	0,0677
RCP	–0,0208	0,0677	0,0677	1,0000

Tabela 1-Matriz de correlações

Na modelização com componentes principais obtêm-se os resultados apresentados no quadro 10.

	CP_1	CP_2	CP_3	CP_4
valor próprio	1,445	0,986	0,911	0,658
prop.explicada	0,361	0,247	0,228	0,164
prop.acumulada	0,361	0,608	0,836	1,000
peso das variáveis				
RA	0,5898	0,4144	0,2189	0,6577
SV	–0,6387	–0,0795	–0,2747	0,7143
PT	–0,4059	0,1016	0,9083	0,0023
RCP	–0,2819	0,9009	–0,2273	–0,2393

quadro 10: resultados obtidos na análise de componentes principais

Como pode ser visto no quadro anterior, apenas a primeira componente apresenta o valor próprio superior à unidade. No entanto, os valores próprios correspondentes à segunda e terceira componente não se afastam muito, assumindo os valores 0,98 e 0,91, respectivamente.

A primeira componente explica cerca de 36% da variabilidade contida nas quatro variáveis em estudo. Esse valor eleva-se para 61% se forem consideradas

duas primeiras componentes principais. As três primeiras componentes explicam cerca de 84% da variabilidade dos dados iniciais.

As três primeiras componentes seriam suficientes para caracterizar a situação das empresas em análise uma vez que, em relação às variáveis iniciais, se perde 16% da informação aí contida. Tal perda parece não ser muito importante uma vez que se ganha em termos de dimensão já que se passa de quatro variáveis para três.

Por outro lado, e seguindo de perto a regra de Kaiser, deveriam ser consideradas tantas componentes quantos valores próprios superiores a um. Tal parece insuficiente porque, neste caso, só seria considerada uma CP o que levaria a um grande desperdicio de informação contida nas variáveis iniciais. Mais razoável seria considerar as componentes com valores próprios superiores a 0,7 (seguindo Jolliffe,1986), que neste caso seriam três.

Porém, para o estudo que nos propomos fazer há necessidade de analisar todas as componentes principais não podendo haver nenhuma exclusão. O nosso objectivo não é o da redução da dimensão dos dados. Pretendemos apenas analisar um conjunto de dados multivariados, e a análise de componentes principais parece-nos uma das técnicas mais indicadas.

A primeira componente pode ser escrita na seguinte forma:

$$Z_1 = 0{,}6RA - 0{,}6SV - 0{,}4PT - 0{,}3RCP$$

Esta componente contrasta, essencialmente, rotação do activo com solvabilidade. Pode ser vista como uma relação entre a função comercial e a função financeira, ou de outra forma, como uma relação entre a política comercial e o financiamento. Em termos comerciais, esta componente dá informação sobre a eficiência das empresas, dado que a variável rotação do activo tem um peso elevado. Neste contexto, vender muito e rápidamente é sinónimo de eficiência.

A distribuição da primeira componente é aproximadamente Normal. O teste do Qui-Quadrado não rejeita essa hipótese para um "p-value" de 36%.

A utilização das componentes principais permite-nos testar a discordância das observações em termos multivariados. No entanto, dada a normalidade

aproximada desta componente, podem ser utilizadas as estatísticas de teste já aqui usadas para determinar se existe ou não discordância em relação ao modelo básico, que se supõe ser Normal.

Como as variáveis apresentam correlações reduzidas (ver matriz de correlações, tabela 1), cada componente contém, aproximadamente, a mesma informação de uma variável considerada isoladamente. Deste modo são identificadas as mesmas empresas outliers usando quer as variáveis iniciais ou as componentes principais consideradas individualmente. Por esta razão, não apresentamos esses resultados dado que eles são idênticos aos obtidos anteriormente.

A segunda componente, $Z_2 = 0{,}4RA - 0{,}1SV + 0{,}1PT + 0{,}9RCP$, é definida principalmente pela rentabilidade dos capitais próprios e pela rotação do activo. Pode ser vista como um indicador da eficiência das empresas. Essa eficiência diz respeito tanto ao factor comercial, vender muito e rápidamente, como ao factor financeiro, fazer uma boa utilização dos capitais.

A seguinte expressão define a terceira componente:

$$Z_3 = 0{,}2RA - 0{,}3SV + 0{,}9PT - 0{,}2RCP$$

Nesta componente é evidente o peso elevado da variável produtividade do trabalho, com um coeficiente de 0,9, em relação às outras variáveis cujo peso é apenas de 0,2 e 0,3. Pode ser caracterizada como um indicador da eficiência das empresas, em particular na utilização dos recursos humanos. Esta componente tem distribuição relativamente próxima da distribuição normal. O teste do Qui-Quadrado não rejeita essa hipótese para um "p-value" de 7%.

Pela análise gráfica das características desta componente, são visíveis as observações identificadas como outliers potenciais. Existem várias observações deste tipo e em ambas as abas da distribuição, no entanto, as observações relativas às empresas Ticket restaurante de Portugal, SA e Philip Morris, Lda afastam-se das restantes estando a primeira na parte inferior e a segunda na parte superior da distribuição.

$$Z_4 = 0{,}7RA + 0{,}7SV - 0{,}2\ RCP,$$

é a expressão que define a última componente. Contribuem para esta variável, e com peso idêntico os rácios rotação do activo e solvabilidade. O peso dos restantes rácios é reduzido, a rentabilidade dos capitais próprios é 0,2 e a produtividade do trabalho tem um peso próximo de zero, por isso pode ser omitido.

A sua distribuição está relativamente próxima da distribuição normal, o " p-value" com que é aceite esta hipótese é de cerca de 8%. Existem várias empresas candidatas a outliers e em ambas as abas da distribuição.

Apesar de úteis na identificação e teste a outliers potenciais, as componentes principais consideradas individualmente detectam outliers que também são detectados pelas variáveis iniciais, no caso de existirem correlações reduzidas entre as variáveis, como é o caso. Não existe ganho aparente na utilização das componentes individuais em vez das variáveis iníciais, a não ser o da redução da dimensão.

O interesse principal das componentes principais reside no facto de elas permitirem a detecção de outliers em termos multivariados. São utilizadas estatísticas de teste que combinam a informação de várias componentes.

A seguinte estatística de teste foi sugerida por Rao(1964), e representa a soma dos quadrados das últimas $q(<p)$ componentes. Formalmente esse teste é representado por :

$$d_{1i}^2 = \sum_{k=p-q+1}^{p} Z_{ik}^2 ,$$

onde Z_{ik} representa o valor da k-ésima componente para a observação de ordem i.

Um problema a considerar é a escolha do valor de q, ou seja, quantas devem ser as componentes a considerar, começando pela última. A determinação desse valor parece-nos um pouco subjectiva, dado existirem vários métodos que podem ser seguidos.

O "oposto" da regra de Kaiser permite reter as componentes com valores próprios inferiores à unidade. No entanto, o valor 1 utilizado como crítico neste critério pode ser considerado muito severo. O valor 0,7 foi indicado por

Jolliffe(1986). No nosso caso, existem três componentes com valores próprios inferiores à unidade mas apenas um deles é inferior a 0,7. Assim, se utilizado como critério o oposto da regra de Kaiser, o teste incluiria os valores relativos às três últimas componentes, ou apenas a última seguindo o critério indicado por Jolliffe.

No quadro seguinte (quadro 11) mostram-se os resultados obtidos: o número de componentes usadas, o valor observado para a estatística de teste e as observações que mais contribuiram para o valor do teste, ordenadas em forma decrescente. Essas observações apresentam valores elevados nesta estatística, por isso são potencialmente discordantes.

Componentes utilizadas	**valor do teste**	**observações com valores mais elevados**
CP_2, CP_3 e CP_4	1019,61	Philip Morris (Portugal) (58); Henkel Hibérica, SA(191); Ticket Restaurante de Portugal,SA(86).
CP_4	262,45	Henkel Hibérica,SA(191); Ticket Restaurante de Portugal,SA(58).

quadro 11: resultados da aplicação de d_{1i}^2

Ao ser analisada apenas a última componente principal obtêm-se como principais candidatos a outliers as empresas Henkel Hibérica(191) e Philip Morris(58), como pode ser visto no gráfico da figura 19. O afastamento daquelas observações em relação às restantes é notório pelo que poderão ser consideradas como outliers. Em relação às empresas que apresentam os valores elevados, nesta estatística, e a seguir às empresas anteriores, pensamos que o afastamento que apresentam em relação ao "grosso" das restantes empresas não é suficiente para poderem ser consideradas como empresas outliers. Ao serem consideradas as três últimas componentes, no cálculo de d_{1i}^2, existem três observações como

potencialmente aberrantes, como pode ser visto na figura 20. Em conjunto, as três empresas apresentam uma contribuição de cerca de 38% para o valor do teste. Pensamos ser possível declarar aquelas três empresas, Henkel Hibérica, Philip Morris e Ticket Restaurante, como sendo outliers. Tal decisão baseia-se, sobretudo, na análise gráfica e vem confirmar os resultados de alguns testes realizados anteriormente.

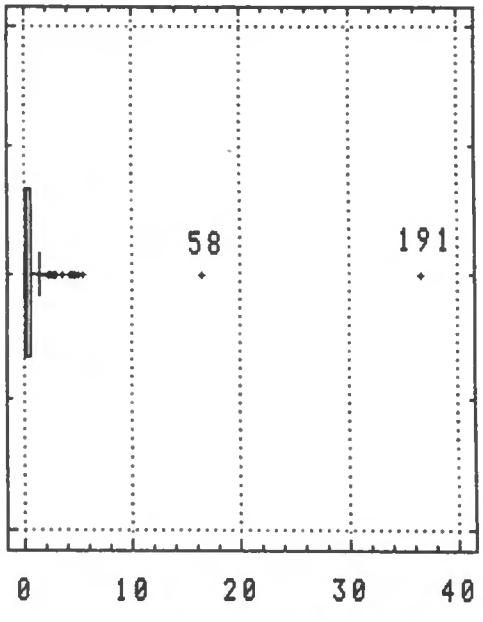

Fig.19-caixa de bigodes de d_{1i}^2 incluindo CP_4

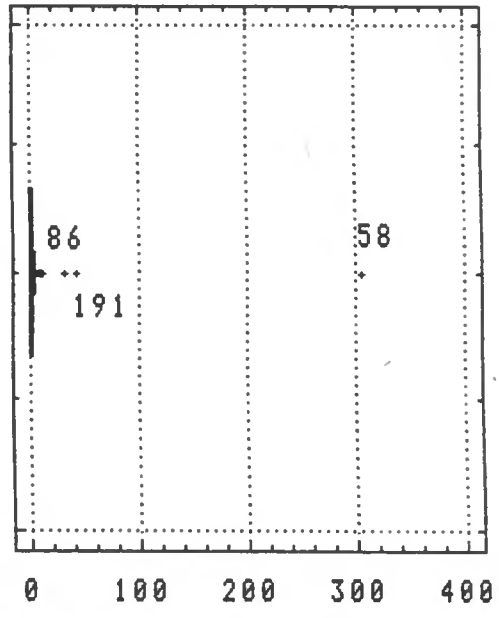

Fig.20-caixa de bigodes de d_{1i}^2 incluindo CP_2, CP_3 e CP_4

Uma grande limitação da estatística anterior refere-se ao facto de ser dada pouca importância às últimas componentes. Os valores de Z_{ik}^2 são mais pequenos à medida que aumenta o número da componente, k, dado que elas estão ordenadas segundo a sua variância. Isso significa que ao serem utilizadas várias componentes no cálculo de d_{1i}^2, e se o número de componentes utilizadas se aproxima do número total de componentes, como é o caso, então a estatística atribui um peso insuficiente às últimas componentes. Esse efeito é indesejado uma vez que as últimas componentes têm um papel importante na detecção de certo tipo de outlier, que por vezes não é identificado pelas variáveis originais nem pelas primeiras componentes. Para evitar este problema, e permitir que todas as componentes

tenham igual peso na construção dos testes, utilizam-se as componentes estandardizadas:

$$Z^{*}_{ik} = Z_{ik} / \lambda_k^{1/2}.$$

Z^{*}_{ik} representa o valor da componente estandardizada k para a observação i e é obtido pela divisão do valor da componente k, para a observação i, pelo desvio padrão, $\sqrt{\lambda_k}$, dessa componente. Ao serem divididas pela respectiva variância, as componentes elevadas ao quadrado, apresentam igual contribuição no cálculo das estatísticas de teste. Os pesos das componentes são inversamente proporcionais às suas variâncias. A nova versão de d^2_{1i}, onde são utilizadas as componentes estandardizadas, tem o seguinte aspecto:

$$d^2_{2i} = \sum_{k=p-q+1}^{p} Z^2_{ik} / \lambda_k$$

Esta estatística foi estudada de modo análogo ao já efectuado para d^2_{1i}, e os resultados obtidos encontram-se resumidos no seguinte quadro:

Componentes utilizadas	**observações com valores mais elevados**
CP_2, CP_3 e CP_4	Philip Morris(58); Mague(30)
CP_4	Henkel Hibérica(191); Philip Morris(58)

quadro 12: resultados da aplicação d^2_{2i}

Ao ser utilizada apenas a última componente, naturalmente, não existe alteração em relação aos resultados obtidos aquando da utilização da primeira estatística de teste. Tanto a empresa Henkel Hibérica como a Philip Morris são outliers, como foi visto na altura.

Ao serem utilizadas as três componentes com valores próprios inferiores à unidade, depois de serem estandardizadas, a observação mais extrema diz respeito à empresa Philip Morris(o mesmo aconteceu com d^2_{1i}). Como pode ser constatado

na figura 21, existe um grande número de observações potencialmente aberrantes, não estando, no entanto, tão afastadas em relação à localização central dos dados como aquela empresa. Deste modo, pensamos poder considerar como outliers as empresas Philip Morris (58) e Mague(30).

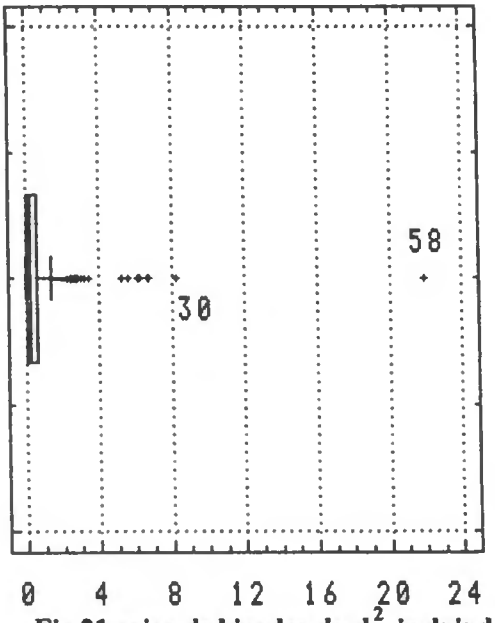

Fig 21-caixa de bigodes de d^2_{2i} incluindo CP_2, CP_3 e CP_4

Outra estatística de interesse na detecção de outliers utilizando componentes principais foi considerada por Gnanadesikan & Kattenring num trabalho de 1972 e referido por Jolliffe(1986) e é definida pela seguinte expressão:

$$d^2_{3i} = \sum_{k=p-q+1}^{p} \lambda_k Z^2_{ik}$$

Na utilização deste teste entram $q(<p)$ componentes, sendo o valor de q definido pelos mesmos métodos usados para as estatísticas anteriores. Se $p=q$ é dada grande importância às observações que têm um grande efeito nas primeiras componentes.

Para identificar empresas outliers usamos a estatística anterior para $q=1$, $q=2$, $q=3$ e $q=p=4$(quadro 13).

Componentes utilizadas	observações com valores mais elevados
CP_4	Henkel Hibérica(191); Philip Morris(58); Mague(30)
CP_3 e CP_4	Philip Morris(58); Henkel Hibérica(191); Ticket Restaurante(86)
CP_2, CP_3 e CP_4	Philip Morris(58)
CP_1,CP_2, CP_3 e CP_4	Philip Morris(58)

quadro 13: resultados da utilização de d^2_{3I}

Utilizando três ou quatro componentes aparece apenas uma empresa como sendo outlier potencial. É a única observação detectada como extrema no diagrama caixa de bigodes. Não restam dúvidas: esta empresa, Philip Morris(58), é uma empresa outlier.

Se no cálculo de d^2_{3i} entrar apenas a última ou as duas últimas componentes, é grande o número de empresas com valores suspeitos. Apesar de a ordem porque aparecem ser diferente, as observações com valores mais estranhos são as mesmas. Aquelas empresas já foram declaradas como outliers depois da aplicação de outros testes.

Não podemos deixar de referir que estes resultados, apesar de confirmarem outros obtidos anteriormente, devem ser aceites com algumas reservas. Com efeito, a aplicação das estatísticas de teste anteriores requer a normalidade das componentes principais. Isto significa que a observação multivariada **x** devia ter aproximadamente distribuição Normal multivariada. Não se pode assumir essa hipótese uma vez que nem todas as variáveis são normalmente distribuidas, como é o caso da variável rentabilidade dos capitais próprios.

Socorrendo-nos, mais uma vez, da análise gráfica e analisando diagramas de dispersão das componentes duas a duas, acabamos por aceitar como outliers as empresas de ordem 58, 86 e 191 que correspondem respectivamente a Philip Morris, Ticket Restaurante de Portugal e Henkel Hibérica. De facto, todos os diagramas apresentam a empresa Philip Morris muita afastada das restantes. Também a empresa Ticket Restaurante se afasta muito das restantes se se considerarem os diagramas que incluem CP_3. Os gráficos de CP_4 com outra componente qualquer evidenciam a inconsistência da observação relativa à empresa Henkel Hibérica. Veja-se, a título de exemplo, os gráficos seguintes.

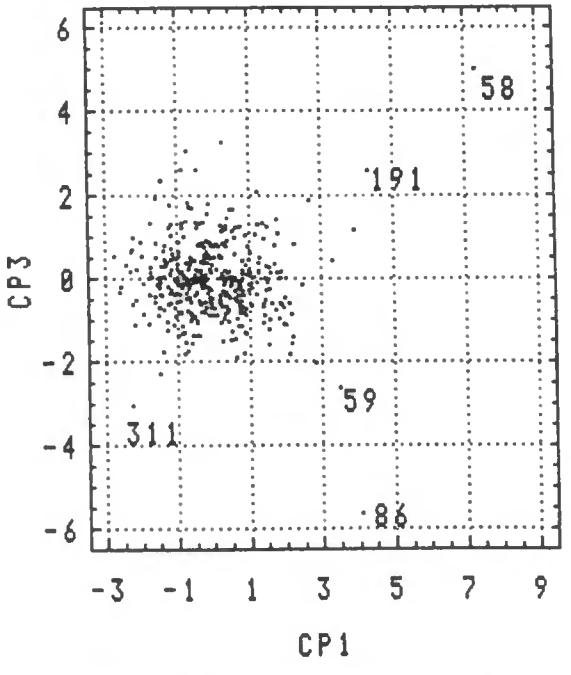

Fig.22-diagrama de dispersão de CP_1 e CP_3

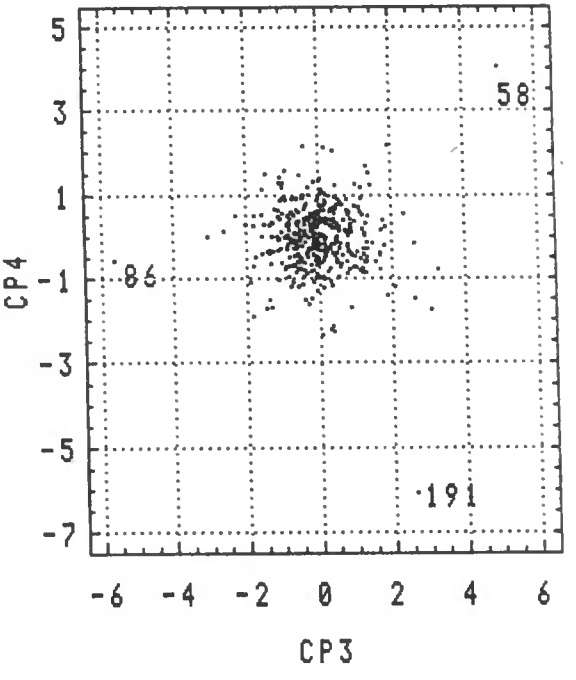

Fig.23-diagrama de dispersão de CP_3 e CP_4

CONCLUSÕES

Pretendeu-se com esta dissertação contribuir para a discussão do estudo de observações "anormais" ou aberrantes designadas por outliers. O objectivo principal era o de detectar e testar outliers num conjunto de dados reais aplicando diversos métodos sem, no entanto, esquecer o enquadramento teórico deste tema.

Assim, na primeira parte abordamos vários aspectos teóricos e metodológicos relativos a outliers. No primeiro capítulo foram indicadas as formas mais usuais de tratar este tipo de observação: modelos de discordância e acomodação.

Os modelos de discordância constituem a abordagem mais conhecida e mais frequente de testar observações aberrantes. Nesta abordagem utilizam-se testes para verificar a discordância de algumas observações em relação ao modelo básico. Essas observações são préviamente identificadas como candidatas a outliers, por exemplo através da análise gráfica ou, como justifica Rosado(1984), pelo próprio teste. Foram apresentadas algumas estatísticas de teste para o caso de o modelo básico subjacente ser Normal ou Gama (ou Exponencial).

Com a acomodação de outliers pretende-se garantir uma certa protecção contra a influência nos resultados deste tipo de observação. A utilização de métodos robustos é a via mais adequada para a obtenção desse objectivo. Foram vistos alguns tipos de estimadores robustos em relação à localização e dispersão dos dados.

No segundo capítulo descrevemos a análise de componentes principais e justificamos a sua utilização na identificação de outliers. Foram também apresentadas as estatísticas de teste para outliers mais importantes no contexto das componentes principais.

A segunda parte foi dedicada a questões práticas. No terceiro capítulo foram apresentados os dados que serviram de base à análise empírica. Os dados, representando 400 das maiores empresas com actividade em Portugal, foram analisados em termos de quatro rácios financeiros: *rotação do activo*,

solvabilidade, produtividade do trabalho e *rentabilidade dos capitais próprios.* Cada rácio foi analisado relativamente à sua simetria, distribuição aproximada e outliers potenciais. Foram aplicadas transformações aos dados originais com o objectivo de aproximá-los da distribuição Normal por forma a poderem ser aplicados testes adequados com valores críticos e/ou distribuição conhecida. Obtiveram-se distribuições muito próximas da distribuição Normal para os rácios *rotação do activo, solvabilidade* e *produtividade do trabalho.* Tal não aconteceu em relação à *rentabilidade dos capitais próprios*, no entanto a sua distribuição é aproximadamente simétrica.

No quarto capítulo testaram-se alguns outliers potenciais quanto à sua discordância em relação ao modelo básico. Utilizaram-se alguns dos testes univariados indicados para o modelo básico Normal para os rácios que seguem, aproximadamente, esta distribuição.

Testamos ainda alguns outliers multivariados recorrendo à análise de componentes principais.

Verificou-se que, para os rácios *rotação do activo* e *rentabilidade dos capitais próprios*, as empresas com maiores possibilidades de serem declaradas como outliers são as mesmas antes e depois da transformação dos dados. Este facto é consequência de a transformação potência operada ser de expoente positivo. Neste caso, só a ordem das observações é mantida como a redução da escala é feita com a mesma intensidade tanto para valores elevados como para valores reduzidos.

Em relação aos rácios *solvabilidade* e *produtividade do trabalho* a realidade é bem diferente. Com efeito, estas variáveis só apresentam uma distribuição bastante próxima da distribuição normal depois de operada uma transformação potência de valor negativo. Aos dados é adicionada uma constante positiva de forma a que todos os valores sejam positivos para possibilitar a sua elevação a um valor negativo. A aplicação deste tipo de transformação traz como consequência a redução com maior intensidade da escala dos valores elevados do

que dos valores menores. Assim, e como se multiplicou cada observação por (−1) para manter a ordem inicial dos dados, as observações extremas (máximo e mínimo), correspondem às mesmas empresas, antes e depois da transformação. No entanto, depois da transformação as empresas com os menores valores afastam-se mais do grupo formado pelas restantes do que as empresas com os valores mais elevados, ao contrário do que acontecia inicialmente. Assim somos confrontados com uma situação em que as observações com maiores possibilidades de serem consideradas outliers deixam de o ser depois de efectuada a transformação, passando outras observações a ocupar o seu lugar.

Em termos univariados foram testadas algumas das observações com valores mais afastados dos restantes em cada um dos rácios, depois de efectuada a tansformação. Foram aplicados alguns dos testes descritos no capítulo 1 para populações Normais, aos rácios que seguem aproximadamente a distribuição Normal. Relativamente à *rotação do activo* não foi encontrada nenhuma observação como discordante. Quanto à *solvabilidade*, apenas a empresa Henkel Hibérica foi considerada como outlier sendo a empresa Ticket Restaurante de Portugal discordante em relação ao rácio *produtividade do trabalho.*

As empresas potencialmente outliers relativamente ao rácio *rentabilidade dos capitais próprios* foram testadas através de métodos não paramétricos dado esta variável não seguir uma distribuição conhecida. Duas empresas são consideradas outliers em relação a este rácio: Phillip Morris e Keller Marítima.

Em termos multivariados pretendeu-se detectar empresas outliers, não em relação a qualquer uma das variáveis, mas em relação às variáveis consideradas conjuntamente.

Depois de determinadas e caracterizadas, as componentes principais são utilizadas para testar outliers através do uso de estatísticas de teste que consideram a informação de várias componentes no seu cálculo.

Com a utilização de estatísticas de teste que incluem informação de várias componentes principais existe um relativo acréscimo de informação. Assim, ao ser

usada a estatística d_{2i}^2 é identificada como outlier a empresa Mague, SA, com o número de ordem 30, que até aqui ainda não tinha sido identificada como tal. Também com a utilização de d_{3i}^2, aquando da utilização de apenas a quarta componente, aquela empresa é identificada como outlier. A justificação para tal é dada pelo facto de esta empresa ter sido considerada a melhor no sector de metalomecânica seguida a grande distância pela segunda.

Quanto às outras observações identificadas como aberrantes por estes testes elas correspondem a empresas já anteriormente identificadas como tal pelos métodos univariados.

BIBLIOGRAFIA

Anderson, T.W.(1984); "*An Introduction to Multivariate Statistical Analysis*", 2nd ed. Wiley, New York.

Andrews, D. F. e **Pregibon**, D.(1978); "Finding outliers that matter" *J.R.Statist. Soc.* B, **40**, 85-93.

Anscombe, F.J.(1960); "Rejection of outliers".*Technometrics*, **2**, 123-147.

Atkinson, A. C. (1981); "Two graphical displays for outlying and influential observations in regression". *Biometrika*, **68**, 13-20.

Atkinson, A. C. (1986); "Masking unmasked". *Biometrika*, **73**, 533-541.

Bacon-Shone, J. e **Fung**, W.K.(1987); "A new graphical method for detecting single and multiple outliers in univariate and multivariate data". *Applied Statistics*, **36**, 153-162.

Barnett,V.(1978);"The study of outliers: purpose and model". *Applied Statistics*, **27**, 242-250.

Barnett, V. e **Lewis,** T. (1978); "*Outliers in statistical data*", John Viley & Sons, New York.

Barnett, V. e **Lewis**, T. (1994); "*Outliers in statistical data*", John Viley & Sons, New York.

Beckman, R. J. e **Cook**,R. D.(1983);"Outlier.....s".*Technometrics*, **25**,119-163.

Bendre, S. M. e **Kale**, B. K. (1987); "Masking effects on tests for outliers in normal samples". *Biometrika*, **74**, 891-896.

Brant, R. (1990); "Comparing classical and resistant outlier rules". *Journal of the American Statistical Association*, **85**, 1083-1090.

Campbell, N. A.(1978); "The influence function as an aid in outlier detection in discriminant analysis". *Applied Statistics*, **27**, 251-258.

Caroni, C. e **Prescott**, P.(1992); "Sequential application of Wilk´s multivariate outlier test", *Applied Statistics*, **41**, 355-364.

Carroll, R. J.(1982); "Two examples of transformations when there are possible outliers". *Applied Statistics*, **31**, 149-152.

Chambers, R. L. e **Heathcote**, C. R. (1981); "On the estimation of slope and the identification of outliers in linear regression". *Biometrika*, **68**, 21-33.

Chatfield, C. e **Collins**, A.J.(1980); "*Introduction to Multivariate Analysis*", Chapman and Hall, Londres.

Collett, D. e **Lewis**, T.(1976); "The subjective nature of outlier rejection procedures". *Applied Statistics*, **25**, 228-237.

Cook, R. D. (1977); "Detection of influential observation in linear regression". *Technometrics*, **19**, 15-18.

Cook, R. D. (1979); "Influential observations in linear regression". *Journal of the American Statistical Association*, **74**, 169-174.

Cook, R. D. e **Prescott**, P. (1981); "On the accuracy of Bonferroni significance levels detecting outliers in linear models". *Technometrics*, **23**, 59-63.

Cook, R. D. e **Weisberg**, S. (1982); "*Residuals and influence in regression*", Chapman & Hall.

Critchley, Frank (1985); "Influence in principal components analysis". *Biometrika*, **72**, 627-636.

Dagnelie, P.(1977); "*Analyse Statistique à Plusieurs Variables*" Les Presses Agronomiques de Gembloux, Bélgica.

Daniel, Cuthbert (1960); "Locating outliers in factorial experiments". *Technometrics*, **2**, 149-166.

David, H. A. e **Paulson**, A. S. (1965); "The performance of several tests for outliers". *Biometrika*, **52**, 429-436.

Davies, Laurie e **Gather**, Ursula (1993); "The identification of multiple outliers". *Journal of the American Statistical Association*, **88**, 782-801.

Devlin, S. L., **Gnanadesikan**, R. e **Kettenring**, J. R. (1975); "Robust estimation and outlier detection with correlation coeficients". *Biometrika*, **62**, 531-545.

Devlin, S. L., **Gnanadesikan**, R. e **Kettenring**, J. R. (1981); "Robust estimation of dispersion matrices and principal components". *Journal of the American Statistical Association*, **76**, 354-362.

Draper, N. R. e **John**, J. A. (1980); "Testing for three or fewer outliers in two--way tables". *Technometrics*, **22**, 9-15.

Draper, N. R. e **John**, J. A. (1981); "Influential observations and outliers in regression". *Technometrics*, **23**, 21-26.

Eastment, H. T. e **Krzanowski**, W. J. (1982); "Cross-validatory choice of the number of components from a principal component analysis". *Technometrics*, **24**,73-77.

Galpin, J. S. e **Hawkins**, D. M. (1981); "Rejection of a single outlier in two - or three - way layouts". *Technometrics*, **23**,65-70.

Gentleman, J. F. (1980); "Finding the k most likely outliers in two-way tables". *Technometrics*, **22**, 591-600.

Gnanadesikan, R. (1977); "*Methods for statistical data analysis of multivariate observations*", New York: Wiley.

Green, R. F. (1976); "Outlier-prone and outlier-resistant distributions". *Journal of the American Statistical Association*, **71**, 502- 505.

Grubbs, F. E. (1969); "Procedures for detecting outlying observations in samples". *Technometrics*, **11**, 1-21.

Guttman, I. e **Smith**, D. E. (1971); "Investigation of rules for dealing with outliers in small samples from the normal distribution II: estimation of the variance". *Technometrics*, **13**, 101-111.

Hadi, A.S.(1992); "Identifying multiple outliers in multivariate data" *J. R.Statist. Soc.* B, **54**, 761-771.

Hamilton, L. C. (1990); "*Modern data analysis*". Chapman and Hall,Londres.

Hampel, F. R. (1981); "The breakdown points of the mean combined with some rejection rules". *Technometrics*, **27**, 95-107.

Hawkins, D. M. (1974); "The detection of errors in multivariate data using principal components". *Journal of the American Statistical Society*, **69**, 340-344.

Hawkins, D. M. (1980); "*Identification of outliers*", Chapman & Hall, Londres.

Hoaglin, D. C., **Iglewicz**, B. e **Tukey**, J. W. (1986); "Performance of some resistant rules for outlier labeling". *Journal of the American Statistical Association*, **81**,991-999.

Hoaglin, D.C., **Mosteller**, F. e **Tukey**, J.W.(1992); "*Análise Exploratória de Dados. Técnicas Robustas: um Guia*", Edições Salamandra, Lisboa. (Tradução Portuguesa de *Understanding Robust and Exploratory data Analysis*).

Jackson, J. E. e **Hearne**, F. T. (1979); "Hotelling's T_M^2 for PC's - what about absolute values". *Technometrics*, **21**, 253-255.

Jackson, J. E. e **Mudholkar**, G. S.(1979); "Control procedures for residuals associated with principal component analysis". *Technometrics*,**21**, 341-349.

Jain, R. B. (1981); "Percentage points of many-outlier detection procedures" *Technometrics*, **23**, 71-75.

Jobson, J. D. (1991); "*Applied multivariate data analysis*", vol. I: *Regression and Experimental Design*, Springer-Verlag, New York.

Jobson, J. D. (1992); "*Applied multivariate data analysis*", vol. II: *Categorical and Multivariate Methods*, Springer-Verlag, New York.

Jolliffe,J. T.(1986); "*Principal component analysis*".SpringerVerlag,New York.

Johnson, D. E., **McGuire**, S. A. e **Milliken**, G. A. (1978); "Estimating σ^2 in the presence of outliers". *Technometrics*, **20**, 441-456.

Kimber,A.C.(1982); "Tests for many outliers in an exponential sample". *Applied Statistics*, **31**, 263-271.

Kitagawa, G. (1979); "On the use of AIC for the detection of outliers". *Technometrics*, **21**, 193-199.

Krazanowski, W.J.(1984); "Sensivity of principal components". *J. R. Statist. Soc.* B, **46**, 558-563.

Lewis, T. e **Fieller**, N. R. J. (1979); "A recursive algorithm for null distributions for outliers: I. Gamma samples". *Technometrics*, **21**, 371-376.

Li, G. e **Chen**, Z. (1985); "Projection - Persuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo". *Journal of the American Statistical Association*, **80**, 759-766.

Marasinghe, Mervyn G. (1985); "A multistage procedure for detecting several outliers in linear regression". *Technometrics*, **27**, 395-399.

Mardia, K.V. *et al.* (1979); "*Multivariate Analysis*", Academic Press.

Mason, Robert L. e **Gunst**, Richard F. (1985); "Outlier - induced collinearities". *Technometrics*, **27**, 401-407.

Martinsek, A. T. (1988); "Negative regret, optional stopping and the elimination of outliers". *Journal of the American Statistical Association*, **83**, 160-163.

Massy, W. F. (1965); "Principal components regression in exploratory statistical research". *Journal of the American Statistical Association*, **60**, 234-256.

McMillan, R. G. e **David**, H. A. (1971); "Tests for one of two outliers in normal samples with known variance". *Technometrics*, **13**, 75-85.

McMillan, R. G. e **David**, H. A. (1971); "Tests for one of two outliers in normal samples with unknown variance". *Technometrics*, **13**, 87-100.

Morrison,D.F.(1990); "*Multivariate Statistical Methods*",3rd edition, McGraw-Hill, New York.

Muñoz-Garcia, J. ; **Moreno-Rebollo**, J. L. e **Pascual-Acosta**, A.(1990); "Outliers:a formal approach".*International Statistical Review*, **58**,215-226

Murteira, B. J. F. (1993); "*Análise exploratória de dados - estatística descritiva*".McGraw-Hill, Lisboa.

Pires, Ana M. e **Branco**, J. A. (1991); "Importância da estimação robusta em CP's". *3ª Conferência de MAEG.*

Prescott, P.(1978); "Examination of the behaviour of tests for outliers when more than one outlier is present". *Applied Statistics*, **27**, 10-25.

Prescott, P. (1979); "Critical values for a sequential test for many outliers". *Applied statistics*, **28**, 36-39.

Rao, C. R.(1964); "The use and interpretation of principal component analysis in applied research", *Sankhyã A*, **26**, 329-358.

Rosado, F. M. F. (1984); "*Existência e detecção de outliers, uma abordagem metodológica*" - Tese de Doutoramento, Faculdade de Ciências de Lisboa.

Rosner, Bernard (1975); "On the detection of many outliers".*Technometrics*, **17**, 221-227.

Rosner, Bernard (1983); "Percentage points for a generalized ESD many-outlier procedure". *Technometrics*, **25**, 165-172.

Rousseuw, P. J. e **Zomeren**, B. C.(1990); "Unmasking multivariate outliers and leverage points". *Journal of the American Statistical Association*, **85**, 633-651.

Schweder, T. (1976); "Some "optimal" methods to detect structural shift or outliers in regression". *Journal of the American Statistical Association*, **71**, 491-501.

Tatsuoka, M. M.(1971); "*Multivariate Analysis: Techniques for Educational and Psychological Research*", John Wiley & sons, Inc., New York.

Tietjen, G. L. and Moore, R. H. (1972); "Some Grubbs-type statistics for detection of several outliers". *Technometrics*, **14**, 583-597.

Watson, C. J.(1990); "Multivariate distributional properties, outliers, and transformations of financial ratios". *The accounting review*, **65**, 682-695.

Wold, S. (1978); "Cross-validatory estimation of the number of components in factor and principal components models". *Technometrics*, **20**, 397-405.