IDENTIFICAÇÃO DE OUTLIERS*

MARIA MANUELA CARIA FIGUEIRA**

* Este artigo baseia-se na tese do Mestrado em Matemática Aplicada à Economia e à Gestão.

** Professora Adjunta da E.S.T.G. do Instituto Politécnico da Guarda.

1-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. 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.

A preocupação com observações outliers é antiga e data das primeiras tentativas de analisar um 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.

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.

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.

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.

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 anormais em qualquer uma das variáveis.

No estudo de outliers multivariados, 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.

2-Métodos de identificação de outliers

2.1-Modelos de discordância

Uma abordagem muito utilizada 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, segundo a qual a amostra foi retirada de uma população com distribuição específica que pode ou não ser conhecida e ser especificada completamente ou não, e onde não existem observações "anormais". Em oposição, a hipótese alternativa 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. As hipóteses alternativas podem ser determinísticas, inerentes, de "mistura" ou por contaminação, por deslizamento(slippage) ou permutáveis.

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

- aqueles em que as observações discordantes da amostra são identificadas como sendo outliers, e

- aqueles que testam a presença de outliers mas não identificam observações particulares como outliers.

Os primeiros têm a forma típica de $T=máx_{1\leq i\leq h}\ h_i(X_i,U)$, onde U é uma estatística baseada na amostra e h_i uma função dessa estatística e das observações. Rosado(1984), na sua tese de doutoramento, dedica grande atenção a este grupo de testes. Os últimos podem assumir uma de entre as seis formas básicas de testes estatísticos: estatística de excesso de dispersão (propagação), estatísticas de amplitude/dispersão, estatística de desvio/dispersão, estatísticas de "soma de quadrados", estatísticas dos momentos de ordem superior e estatísticas de localização/extremos.

2.2-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 **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 **0**, a média da amostra $\bar{\mathbf{x}}$ ou a média da população, μ , e Γ poderá ser a matriz de covariâncias (V) ou a matriz de covariâncias da amostra (S), dependendo do facto de μ e 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 (se

$$Z_i = a_{1i}X_1+a_{2i}X_2+...+a_{pi}X_p$$

$= \mathbf{a}_i^T \mathbf{x}$

onde $\mathbf{a}_i^T = [\, a_{1i}\, a_{2i} ... a_{pi} \,]$ é um vector de constantes. Deve impor-se a condição de normalização

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

Determinação da primeira componente: Z_1

Para tal, utiliza-se o método dos multiplicadores de Lagrange com a condição necessária de primeira ordem (derivada nula) e a restrição de normalização ($\mathbf{a}_1^T \mathbf{a}_1 = 1$), obtendo- se de tal forma o vector próprio $\mathbf{a}_1$ associado ao valor próprio λ da matriz $\sum$ **.**

A solução para $\mathbf{a}_1$ é diferente de zero se ($\sum - \lambda \mathbf{I}$) for uma matriz singular, logo λ deve ser tal que $| \sum - \lambda I | = 0$. Então, existe uma solução diferente de zero se e só se λ é um valor próprio de $\sum$. Mas $\sum$ terá, geralmente, p valores próprios todos não negativos visto $\sum$ 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 $\sum$.

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 ($\sum - \lambda I$) $\mathbf{a}_2 = \mathbf{0}$ e escolhe-se λ como sendo o segundo maior valor próprio de $\sum$ e $\mathbf{a}_2$ o correspondente valor próprio. Continuando com este processo obtêm-se todas as componentes principais, $Z_3, ..., Z_p$, que têm variância decrescente e não estão correlacionadas.

Não há dificuldade no caso de existirem alguns valores próprios de $\sum$ iguais, assim 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 CPs 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 $\sum$, utiliza-se a matriz de correlações **P**, o que significa que se pretendem determinar as CPs 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 $\sum$. 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 $\sum$ 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 $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p$.

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

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

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

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 ACPs como uma técnica matemática sem nenhum modelo estatístico subjacente. As CPs obtidas através de **S** são vistas como as componentes principais e não como estimativas do que seria obtido com $\sum$, por isso geralmente omitem-se os chapéus em $\hat{\lambda}_i$ e $\hat{a}_i$.

Se algumas das variáveis iniciais são linearmente dependentes, alguns dos valores próprios de $\sum$ serão nulos. A dimensão do espaço contendo as observações é igual à característica de $\sum$, 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 CPs. 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 CPs 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 CPs que devem (podem) ser desprezadas por não representarem acréscimo significativo de informação.

A questão, é então de saber quantas componentes são significativas, e 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 CPs é significativo. Pretende-se testar se os últimos q de p valores próprios são iguais, se isso acontecer as componentes principais correspondentes podem ser eliminadas.

4.2-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 inflaccionar 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 CPs mesmo quando as variáveis originais não têm essa distribuição.

4.3-Tipos de testes

Estatísticas para testar a presença de outliers numa amostra univariada **[** ver por exemplo Barnett e Lewis(1994) **]** , podem ser usadas em cada uma das CPs 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 fácilmente 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 CPs, especialmente se q, o número de componentes que contribuem para d_{1i}^2 é muito próximo de p. Os valores de Z_{ik}^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 CPs 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 varimáx 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 CPs 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.

4.4-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).

Bibliografia

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.

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.

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.

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.

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

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

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. (1981); "Influential observations and outliers in regression". Technometrics, **23**, 21-26.

Figueira,M.M.C.(1995); "Identificação de outliers:uma aplicação ao conjunto das maiores empresas com actividade em Portugal" Tese de Mestrado-Instituto Superior de Economia e Gestão.

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.

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

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

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.

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

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

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.

Morrison, D.F.(1990); "Multivariate Statistical Methods",3rd edition, McGraw-Hill, N.Y.

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.

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

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