ESTABILIDADE E PERTURBAÇÕES NOS COMPONENTES PRINCIPAIS: UMA QUESTÃO DE ANÁLISE DE SENSIBILIDADE

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida pela Srª Ruth Marilda Fricke e aprovada pela Comissão Julgadora.

Campinas, 2 de julho de 1990.

Prof. Dr. Belmer Garcia Negrillo
Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciências da Computação, UNICAMP, como requisito parcial para a obtenção do Título de Mestre em Estatística.

ESTABILIDADE E PERTURBAÇÕES NOS COMPONENTES PRINCIPAIS: UMA QUESTÃO DE ANÁLISE DE SENSIBILIDADE

DISSERTAÇÃO DE MESTRADO
PROFª RUTH MARILDA FRICKE
Mestranda
DEPARTAMENTO DE MATEMÁTICA
ICEN - UNIJUÍ - IJUÍ (RS)

ESTABILIDADE E PERTURBAÇÕES NOS COMPONENTES PRINCIPAIS: UMA QUESTÃO DE ANÁLISE DE SENSIBILIDADE

PROF. Dr. BELMER GARCIA NEGRILLO
Orientador
DEPARTAMENTO DE ESTATÍSTICA
IMECC - UNICAMP - CAMPINAS(SP)
1990

certas mãos,
muitas,
hábeis, rudes,
nem de longe
sabiam o esforço, prá onde ia?
prá essas mãos
(muitos nem as lembrariam)
este trabalho
que nem em sonhos
nem de longe haveria
sem elas

Ana Cris
28/2/1990

agradecimentos

A Belmer Garcia Negrillo, professor e amigo, que aceitou o desafio de orientar este trabalho;

A Elza Maria Fonseca Falkembach, companheira de trabalho que mudou o rumo de minha história, da Pedagogia para a Estatística;

A Irene Ferreira e Regina Moran, que me auxiliaram com apoio e amizade, na tentativa de superar minhas limitações;

Aos colegas da Estatística, em especial, Luis Aparecido Milan, Paulo Nakamura, Vera Lúcia Damasceno, Vilma Tachibana, Paulo Jacóe, José Luis LLanos Carillo, José Roberto Zorzatto, grandes colegas, na luta acadêmica, na luta de classe;

Aos professores e funcionários do IMECC, em especial aos professores do departamento de Estatística, à Maria Elisa Fini, Ricardo Bassi, Boldrini, Zezé, Cida, Isabel, Iara, Lurdinha, Tavares, ... ;

Aos professores e funcionários da UNIJUÍ, em especial Rosana Callai, José Crippa, Telmo Uriarte, Maria Cristina Pansera, Pedro Borges, Maria Helena, Marivana, Eder, Basso, ... ;

A Miriam Seibel, Evandro Bohrer e Ivanor Duranti, ex-alunos e companheiros de trabalho, que muito contribuíram em termos de Estatística;

A Teodorico, Carlos Renato e Nair Fricke, meu pai, irmão e cunhada e Engenheiro Mauri Diretores da Firma Fricke pelo apoio em termos computacionais;

A Corinta e Vanderlei Geraldi com seu apoio e ajuda prestimosa;

A Noili Demamann, Cláudio Boeira Garcia, Olívio Lopes Vicentini, grandes amigos e companheiros, cujas palavras e ações foram estrela-guia;

A estas pessoas fantásticas, João, Ana, Iara, Ivana, Sílvia, Cecília e Iúna que fazem a minha identidade;

A Sonny, minha mãe, cuja força de luta e perseverança é, ainda e sempre, meu exemplo, rumo e direção;

A cada um de nós, povo brasileiro, que cria espaço para este e outros estudos.

SUMÁRIO

PREFÁCIO

De um certo ponto de vista, o avanço do conhecimento se justifica à medida em que é colocado a serviço do bem estar da humanidade, na busca de continuidade e evolução dos seres vivos e do ambiente. A ESTATÍSTICA, desenvolvida pelos homens a partir da

procura de soluções para suas necessidades, busca o conhecimento em favor da melhoria coletiva. Grandes massas de informações são coletadas a partir de quaisquer perguntas que sejam formuladas no interesse de estudar comportamentos específicos de uma população. Este fato tanto é devido ao considerável aumento populacional da humanidade como à ampliação dos mecanismos de coleta, armazenagem e tratamento de dados. A complexidade crescente das relações no mundo, devida ao avanço da própria ciência, impõe a utilização simultânea de um grande número de variáveis em cada fenômeno social ou natural em estudo. A própria complexificação do saber exige uma maior quantidade de direções observadas em cada contexto. Por isso, também, seu estudo demanda a aplicação de métodos e técnicas cada vez mais abrangentes e complexas. Através deles deve ser possível fazer uma depuração nestas variáveis, auxiliando na compreensão do que é fundamental e do que é secundário ou irrelevante no estudo que está sendo realizado. Precisa-se de métodos e técnicas que desenvolvam uma análise multivariada capaz de operar, simultaneamente, um grande número de variáveis. Sua utilização torna-se mais acessível pela evolução da informática e subseqüente difusão do uso dos micro-computadores. Os programas estatísticos utilizados transformam os mais complicados cálculos em simples operações, liberando o estatístico para a busca de avanço na análise. De posse desta capacidade de entendimento é possível auxiliar a definir e traçar diretrizes de ações-econômicas que beneficiem os cidadãos. De um modo geral, as ações são desenvolvidas no sentido de melhorar a compreensão do mundo de modo a poder interferir na sua trajetória.

Ao se falar em métodos e técnicas mais abrangentes e com capacidade de analisar conjuntos p-dimensionais, justifica-se expor um tratamento estatístico de Análise Multivariada. **ANÁLISE DE COMPONENTES PRINCIPAIS (ACP)** é um destes métodos e possibilita a análise de um conjunto de dados independente do conhecimento de sua estrutura probabilística permitindo a redução da dimensionalidade dos dados e também o estudo das relações entre as variáveis. Consiste numa mudança de base, uma rotação no espaço que garante variáveis não correlacionadas. Basicamente trata-se de uma transformação dos dados

reescritos como uma função destas variáveis originais. Pretende-se escrever $Y=c'X$, onde:

Y: vetor p-dimensional de COMPONENTES PRINCIPAIS,
c: vetor p-dimensional de coeficientes
X: matriz nxp de dados originais.

Esta transformação conserva as distâncias do espaço original utilizando-se de uma transformação ortogonal sujeita a duas restrições:

i) $c'c = 1$
ii) $c_i'c_j = 0, \quad i \neq j$

Este método, iniciado por Pearson em 1901, é bastante conhecido e vários autores, como HOTELLING(1933), ANDERSON(1958), RAO(1964), MARDIA(1979), CHATFIELD(1980), JACKSON(1980), já detalharam sua teoria. Algumas das formas mais comuns de aplicações de CP são para:

* - observar e dimensionar os principais fatores explicativos do comportamento de um conjunto de dados. Com esta aplicação é possível conhecer as relações que existem entre as variáveis presentes no estudo e observar como os registros se agrupam.

* - possibilitar o estudo de um grande número de variáveis quantitativas das quais não se tem, a priori, conhecimento da estrutura de dados, sendo possível diminuir a dimensionalidade do conjunto quando muitas das direções observadas são repetitivas nos aspectos comportamentais que devem explicitar. É preciso, então, fazer uma seleção das variáveis de interesse. Esta pesquisa das famílias de variáveis bem como da triagem no interior de cada família pode ser realizada com a aplicação de ACP. Uma das vantagens da realização desta seleção é a redução da dimensionalidade dos aspectos mensurados em estudos posteriores.

* - reagrupar observações segundo características observadas durante sua aplicação. É a formação de blocos ou de estratos pela própria participação de cada variável no resultado multidimensional de cada observação. A ACP permite a formação destes blocos pela disposição gráfica dos indivíduos analisados ou pela observação das semelhanças e diferenças dos indivíduos através dos coeficientes, c_i, apresentados.

Em resumo, aplica-se ACP afim de saber como estão estruturadas as variáveis, quais são as variáveis que estão correlacionadas, quais são não correlacionadas, quais são as que apresentam aspectos de contrastes entre os níveis de participação, em que aspectos os indivíduos se assemelham, em que eles se desassemelham, como se pode agrupá-los em função de suas semelhanças e de suas diferenças. Definida como uma transformação linear, $Y=c'X$, é uma rotação ortogonal, tal que o novo conjunto de variáveis satisfaz:

* as distâncias do espaço original serão mantidas;

* as novas variáveis são não correlacionadas;

* os Componentes Principais são definidos de tal forma que Y_1, o primeiro componente tem variância máxima, e assim por diante os outros componentes terão variância decrescente até a de Y_p que será a mínima;

* a dimensão p do espaço original pode ser reduzida a um espaço r-dimensional tal que $r < p$.

Hotelling(1933) diz que: " No sentido de ir tão rápido quanto é razoavelmente possível em um dado caso expressando os escores dos testes x_i em um pequeno número de componentes, um procedimento ordenado é requerido para selecionar os componentes no sentido de sua existência , ou de sua importância para nossos propósitos, e rejeitando qualquer um que prove ser de pequena

importância, ou os quais não são claramente definidos pelos dados. Uma situação análoga surge no ajustamento de curvas empíricas. Uma série da forma

$$y = a + bx + cx^2 + \ldots$$

pode ser ajustada, o número de termos usado é limitado pela probabilidade crescente dos erros dos coeficientes de ordem superior, e também pela diminuição das contribuições para a variância total de y por estes termos de ordem mais alta. Se uma série é modificada para consistir de funções ortogonais, os coeficientes sucessivos têm intercorrelação zero. Somente os termos que são significantes podem ser retidos. Uma outra analogia é o uso de equações de regressão envolvendo mais e mais variáveis x_1, x_2, x_3,... para explicar ou predizer y, estas são escolhidas de acordo com suas contribuições para a variância de y.

Estas analogias sugerem que, escolhendo entre a infinidade de possíveis modelos de resolução de nossas variáveis em componentes, nós começamos com um componente Y_1 cuja contribuição para a variância residual é a maior possível; e que nós procedemos neste caminho determinando os componentes, não excedendo n em número, e talvez negligenciando aqueles cujas contribuições para a variância total seja pequena. Isto nós queremos chamar de "o método de Componentes Principais". "

Hotelling(1933) procurou resolver o problema do número de fatores que deveriam participar de um processo quando do estudo de algum fenômeno de interesse. Assim ele avalia o método sob duplo aspecto: o gerenciamento dos significados dos fatores como elementos interpretativos e o gerenciamento do número de variáveis que serão utilizadas considerando-se a sua participação na variância total. Por outro lado, é significativo observar a influência da Análise de Regressão Linear para o próprio desenvolvimento de CP. Fica evidenciado que a preocupação em termos de correlação entre as p variáveis e a análise dos resíduos encaminharam os estudos e, até mesmo a sua formulação.

Para CHATTFIELD(1980) " Na prática não é sempre fácil de lhe achar um significado tanto que o seu uso fica mais em reduzir a dimensionalidade dos dados no sentido de simplificar análises posteriores. Por exemplo, graficar os escores dos dois primeiros componentes para cada indivíduo, é um caminho útil para tentar encontrar os "clusters" nos dados (...) quando efetivamente reduzimos a dimensionalidade para dois".

Chattfield(1980) acentua a dificuldade de interpretação dos CP concordando com seu uso para a redução de dimensionalidade de que fala Hotelling. No seu trabalho admite que cada observação pode ser classificada num grupo, para o caso r=2, onde r é o nº reduzido de CP. A base desta concepção está na observação da dispersão dos dados. Quanto maior a variância entre as observações, maior será o poder de discriminar os possíveis grupos formados no interior do conjunto de observações. As observações redimensionadas irão explicitar quais os elementos formadores de cada "cluster", bem como esclarecerão quanto ao número que pode ser formado. Este procedimento é útil principalmente naqueles casos em que não se está de posse do conhecimento "a priori" de uma estrutura dos dados, impossibilitando a elaboração de estratos com antecedência. Assim, como os CP são resultado de uma metodologia que garante uma variância decrescente, estaria assegurado um poder de definir os "clusters" e, a partir dos coeficientes - c_i, de classificar um indivíduo quanto ao grupo ao qual pertence. Os coeficientes aqui referidos são os autovetores - c_i - associados aos autovalores - λ_i - calculados a partir da Matriz de Covariância Σ. Estes coeficientes informam a contribuição de cada uma das p-variáveis originais no CP, após ter passado pela transformação de ACP. Em termos de análise o que se pode observar é que eles evidenciam as relações entre as variáveis numa espécie de mensuração da participação de cada uma. Um CP tanto pode demonstrar a existência de um contraste entre as p-variáveis como pode representar um peso de participação num mesmo sentido ou direção. Quando estes pesos se assemelham podem representar um valor médio semelhante a média aritmética. O grau de complexidade destes contrastes pode dificultar sua interpretação. No entanto é possível avaliar seu significado

buscando-se formas mais simplificadas de expressão destes coeficientes. Uma destas formas é o uso de aproximações cuja utilização foi estudada por Green(1977) e por Bibby(1980).

Segundo ANDERSON(1958) "Do ponto de vista da teoria estatística, o conjunto de Componentes Principais leva a um conveniente conjunto de coordenadas e o acompanhamento das variâncias dos componentes caracteriza suas propriedades estatísticas. Nas aplicações estatísticas, o método de Componentes Principais é utilizado para encontrar as combinações lineares com a maior variância. Em muitos estudos exploratórios o número de variáveis em consideração é muito grande. Desde que são os desvios, nestes estudos que mais interessam , um caminho para reduzir o número de variáveis a serem tratadas é descartar as combinações lineares que têm as menores variâncias e estudar somente aquelas com grandes variâncias."

Se X, a matriz de dados originais, tem uma distribuição normal multivariada então os contornos de iguais densidade são elipsóides. Neste caso, se os componentes podem ser referidos como um conjunto de coordenadas em "p" dimensões do espaço, requer-se que cada ponto represente um indivíduo posicionado neste espaço segundo estas "p" direções. ANDERSON(1958) define como principal finalidade da aplicação de CP o descarte de variáveis. São conservadas as combinações lineares que apresentam variância máxima desprezando aquelas que apresentarem as menores variâncias, atitude já altamente questionada pelo valor que os menores, principalmente o último CP tem em termos de de estabilidade e para detectar "outlier". Segundo Anderson(1958) CP tem sua aplicação ao nível da análise exploratória de dados. ACP faz uma análise do conjunto de dados de uma forma mais explícita interrelacionando a informação contida nos dados sem no entanto ficar limitado pela correlação apresentada pelas variáveis originais em estudo.

Também PERES(snt) diz sobre o assunto que "A Análise de Componentes Principais e a ANÁLISE FATORIAL são usadas quando o conjunto de variáveis não pode ser dividido em conjunto de variáveis

dependentes e outro de variáveis independentes. Transforma-se então o conjunto de variáveis em outro de interpretação conjunta mais fácil (...). A Análise de Componentes Principais constitui-se num método para a análise da estrutura de interrelação de um conjunto de variáveis aleatórias."

Peres considera um vetor-resposta onde as direções observadas são a própria definição da observação. É o seu comprimento que interessa: o CP capta o máximo de variação dos dados, numa relação de ordem decrescente. Esta propriedade traria condições de diferenciação dos indivíduos. Cada estrato pode ser definido pelo valor dos coeficientes. Deste modo, Peres, define o uso de CP em dois aspectos além dos que usualmente são dados: formação dos "clusters" e seu emprego na construção de índices. Idéia que, em parte, está presente também no pensamento de Chattfield(1980) e é defendida por Mardia(1975). A definição dos "clusters" tem por base a capacidade que a estrutura de interrelação entre as variáveis, captada pela variância, tem de separar o grupo de dados segundo os desvios em relação ao ponto central. Já a construção de índices a partir da transformação CP tem por base a definição dos autovetores como o nível de participação de cada variável no vetor-resposta de cada indivíduo.

Para Ben(1985) "A primeira Componente Principal resume a informação das I variáveis em uma única variável que é uma combinação linear das variáveis originais, ou seja que para a unidade j a componente é da forma $a'X_j$. O vetor será construído de tal modo que a primeira componente seja um resumo da totalidade dos dados, no sentido de se minimizar os erros que se comete ao pretender reconstruir - predizer - cada uma das variáveis originais como função linear da primeira componente."

Esta observação de Ben tem por base a escolha do autovalor λ_1, definida como sendo aquele que maximiza a variância do primeiro CP e a conseqüente escolha do autovetor c_1, associado a ele. Isto garante ao primeiro componente a possibilidade de captar o

máximo de diferenciação existente entre os indivíduos observados. Então, ao reduzir a dimensionalidade encontra-se uma expressão preditiva baseada num menor número de variáveis, ampliando seu poder preditivo e diminuindo a margem de erro.

TOMASSONE(1987) diz "A Análise de Componentes Principais (....) é um método utilizado para descrever uma tabela de dados; cada observação está definida por p variáveis. É um método com base em análise de dados cujo conhecimento é indispensável porque vai descrever grandes tabelas de dados, em particular aquelas que provém de enquetes. Ela pertence ao grupo de métodos conhecido pelo nome de Análise Fatorial cujo objetivo é determinar a função de p variáveis ou fatores. Estes fatores servirão para representar as observações de uma maneira geralmente mais simples."

A Escola Francesa coloca o emprego de ACP ao nível da análise descritiva multivariada de um conjunto de dados , análise esta que permite conhecer a estrutura dos dados e que fornece as primeiras relações que podem ser observadas entre os indivíduos pela participação de cada direção avaliada no espaço multidimensional no vetor resposta. Este enfoque pode, também, ser encontrado em trabalho de Phillipeau(1988). Estas estatísticas têm um cunho descritivo pois demonstram as direções dos pontos nos planos do tipo 1, 2 além de que possibilitam visualizar os agrupamentos dos pontos, enumerando-os de forma a identificar os indivíduos que estão nos mesmos sub-espaços. Estudos posteriores descrevendo a composição do grupo pelo nível de participação de cada variável levam a formação de critérios de separação e alocamento de novas observações.

Para MARDIA(1979) " (...) Como um primeiro objetivo Análise de Componentes Principais busca a SLC (combinação linear padronizada) das variáveis originais que tem variância máxima. (...) a variância máxima separaria os candidatos, facilitando considerações sobre diferenças entre eles. (...) Considerações preliminares aplicam em outras situações, tal como a construção de um índice de custos de vida. Mais genericamente, a Análise de Componentes Principais olha

para algumas combinações Lineares as quais podem ser usadas para sumarizar os dados, perdendo no processo tão pouca informação quanto possível. Esta atenção para reduzir a dimensionalidade pode ser descrita como uma "parcimoniosa sumarização" dos dados".

O grau de participação de cada variável nas respostas é mensurado e expresso através dos coeficientes, c_i, que deste modo representam um peso de cada variável na elaboração do índice. O sinal com que se apresenta este peso pode estar indicando contraste entre as variáveis mensuradas, e, como caso particular, se forem todos do mesmo tamanho, positivos, podem representar um valor médio do indivíduo segundo aqueles "p" aspectos.

Esta retrospectiva demonstra que os autores concordam em que o método serve para reduzir a dimensionalidade do conjunto original de variáveis sem que se perca a informação conquistada. Este procedimento sumariza toda informação original mas ao mesmo tempo leva a reduzir os "p" aspectos sob os quais os registros estão sendo observados. Para alguns autores a utilização prática dos CP é dificultada em muitos casos pela interpretação dos coeficientes, preferindo utilizá-lo como método auxiliar nos estudos exploratórios de dados para apoiar análises subseqüentes.

ACP é vista também como um método que permite a separação de uma grande massa de indivíduos em "clusters". No entanto, para que isso ocorra é necessário que a variância entre os grupos seja maior que a variância interna dos mesmos. Mardia(1979) define a construção de um escore geral, indicando inclusive o seu uso para a construção de índices que permitam a classificação de um indivíduo pela participação de cada variável original em seus r primeiros componentes. Neste sentido os coeficientes, c_i, revelam a contribuição de cada variável na formação dos "clusters" como verdadeiros índices. Portanto, com o uso de ACP, é possível conhecer o que varia de indivíduo para indivíduo.

Segundo Mardia(1979), uma das propriedades dos CP está expressa da seguinte maneira: " Os Componentes Principais de um vetor aleatório não são escalarmente invariantes." Devido à esta propriedade, quando não ocorrer unidade de medida única os CP serão calculados a partir da matriz de correlação ao invés da matriz de covariância . Segundo Wold(1978) " (...) é costume padronizar as variáveis antes da análise dando a todas as variáveis a média zero e variância um." Com este procedimento pretende-se estar operando com uma matriz Y apropriadamente escalada, cujos parâmetros e resíduos correspondem à mesma matriz de dados.

O descarte das variáveis pode ser realizado horizontal e verticalmente. É possível descartar CP e, baseados na não contribuição efetiva de variáveis originais, descartar r variáveis, r<p, entre as p variáveis originalmente pesquisadas. Se o máximo de variabilidade já está contido nos componentes ou variáveis que devem permanecer, os demais podem ser descartados pois nada acrescentam para ampliar o conhecimento da estrutura dos dados em termos de relações entre as variáveis enfocadas.

A redução da dimensionalidade ocorre em dois níveis:

a) REDUÇÃO DE COMPONENTES PRINCIPAIS - Y_i

a1 - Descartar os componentes com a menor participação na variância total explicada.

A redução de dimensionalidade tem por base algumas propriedades dos CP, conforme será relatado a seguir :

- a proporção da variação total explicada pelos primeiros "q" componentes principais é dada por :

$$\sum_{i=1}^{q} \lambda_i \;/\; \sum_{i=1}^{p} \lambda_i$$

pois a "variação total" é dada pelo tr Σ, uma vez que:

$$\sum_{i=1}^{p} V(y_i) = tr\ \Sigma \qquad e \qquad V(y_i) = \lambda_i,$$

$$V(y_1) \geq \ldots \geq V(y_p)$$

A razão apresentada dá uma idéia do montante de variação retida pelas q primeiras componentes.

-Se a matriz de covariância de X tem posto q < p , então a variação total de X pode ser inteiramente explicada pelos primeiros q componentes principais.

Este resultado é obtido pelo fato de que se o posto de Σ é q, os (p - q) restantes λ_i, autovalores de Σ, são identicamente nulos , isto é, $\lambda_{q+i} = \ldots = \lambda_p = 0$, e neste caso:

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \longrightarrow 1 \qquad e \qquad \frac{\sum_{i=q+1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \longrightarrow 0$$

A determinação da participação de cada variável na variação total explicada é definida a partir do coeficiente de determinação da variação explicada calculado com base em r^2_{ij}, coeficiente de correlação entre a variável original X_i e o CP Y_i. Dimensiona-se as perdas de informação ocorridas com cada variável através de r^2 acumulado.

a2 - Descartar aqueles autovalores que, na relação entre a soma parcial dos autovalores e a soma total, não sejam necessários para garantir uma razão de aproximadamente 90 %.

Cattel(1966) propôs um gráfico entre λ_j e j, sempre decrescente em termos de valor uma vez que os autovalores são escolhidos para maximizar a variância do componente nesta ordem até λ_p que será o menor deles. Com este gráfico é possível visualizar, nessa relação descendente, onde termina o grupo de maiores autovalores e onde iniciam os menores autovalores. A regra básica deste procedimento pode ser a de conservar autovalores que garantam 90% de variância total contida nos componentes a serem conservados. A visualização em termos de gráfico auxilia na determinação do número de componentes pois permite que se avalie os pontos segundo seu agrupamento e não apenas pela definição do percentual que, em alguns casos, pode não captar amplamente a questão da separação entre os dois grupos de autovalores.

a3 - Descartar os componentes que não atingem, em seu autovalor, o valor um, isto é, $\lambda_i < 1$.

A justificativa para utilizar este critério deve-se ao fato de que na matriz de correlação, R, a variância original é $\sigma_{ii} / \sigma_{ii} = 1$. Como λ_i é a variância do CP i, sendo menor do que 1 está explicando menos do que a variável original sozinha.

a4 - Excluir aqueles componentes cujos autovalores sejam menores que a sua média.

Na prática, ambos os critérios necessitam de ajuste para o caso em que $p \leq 20$, isto é, o critério a_4 tende a subestimar o número de componentes a serem conservados enquanto que o critério a_3 tende a superestimar o número de componentes, não correspondendo a necessidade real de dimensionalidade. Também estes critérios necessitam de uma metodologia auxiliar que consiga avaliar a validade do número indicado pelo critério utilizado. Esta pode ser a utilização do gráfico de Cattel.

as - VALIDAÇÃO CRUZADA

Qualquer um dos critérios citados acima é um tanto subjetivo, as decisões são tomadas como a melhor forma de não perder informação. O posto da matriz de dados tem sido utilizado para este fim. A tentativa de encontrar um critério mais objetivo na escollha de "q" resulta na definição do modelo de validação cruzada. Neste modelo o número de CP a serem retidos é aquele que garante uma melhor previsão da matriz Y.

Eastment e Krzanowski(1982,1983) realizaram estudos sobre o método de validação cruzada para a seleção de CP. Este método, desenvolvido por Wold(1976,1978), pressupõe a escolha daqueles componentes para os quais é possível realizar uma predição da matriz de dados.

Wold escreve a matriz $Y = [(Y_{ik})]$, com

$$Y_{ik} = \alpha_i + \sum_i \beta_{i\alpha} \theta_{\alpha k} + \varepsilon_{ik}$$

onde: α , β e θ expressam a parte sistemática dos dados Y_{ik} e ε_{ik} representam os resíduos, ou seja, o ruído , parte aleatória. Estes parâmetros são estimados para minimizar a variância dos resíduos.

A estimação do posto de Y deve considerar o quanto dos dados é devido à parte sistemática e o quanto ao ruído. Assim a matriz de dados é particionada em g grupos sendo que estes, um a um, vão sendo deletados ao mesmo tempo em que são substituídos por valores preditos a partir do restante da matriz. Tem-se então a possibilidade de calcular as diferenças entre os valores verdadeiros de Y_i e os valores preditos de $\hat{y}_i$, no caso, obtém-se $(Y_i - \hat{y}_i)$; esta diferença constitui-se no ε_{ik}. Após esta comparação a matriz é recomposta e deleta-se outra linha de componentes, percorrendo todos os passos enunciados anteriormente. Quando todos os grupos de CP passaram pelo processo de deleção, previsão e comparação, estar-se-á apto a determinar para qual valor de "q" o processo desenvolvido

apresentou o melhor poder de previsão. Assim, serão retidos os q componentes que garantam o menor resíduo.

b - REDUÇÃO DAS VARIÁVEIS ORIGINAIS - X_i

O descarte de variáveis originais é aconselhável tendo em vista, principalmente, dois aspectos :

1º Crescimento da precisão dos estimadores nas variáveis retidas;

2º Redução do número de variáveis necessárias em estudos similares futuros.

b1 - Excluir as variáveis com a contribuição máxima nos menores CP

O método utilizado para descartar as variáveis originais consiste em observar seu grau de participação tanto a nível dos componentes definidos para serem conservados como a nível de sua participação naqueles componentes que formam o grupo de menores autovalores. Aquela variável que apresentar o maior índice de participação em Y_p, componente com o menor valor de λ, será descartada procedendo-se da mesma maneira em relação ao (p - 1)-ésimo componente, até que todos os (p - q) menores componentes tenham sido investigados e as variáveis originais tenham sido descartadas sempre que não tenham ainda sido descartadas em um dos componentes investigados. Este procedimento justifica-se pela própria definição dos CP. Eles devem conter o máximo de variância em seus primeiros q componentes e, nesse caso, a variação dos dados é dada pelas variáveis que aí apresentam os maiores índices de participação nos autovetores correspondentes.

b2 - Procedimento iterativo

Este método é uma variante do anterior. Após haver sido descartada a variável X que apresentar o maior coeficiente na CP com o menor autovalor, deve-se recalcular os CP. A repetição deste

processo será mantida até que permaneçam somente aqueles componentes com as mais altas variâncias. Segundo Mardia não existe diferença muito significativa entre os dois métodos, estudos recentes vêm comprovando a validade desta afirmação do autor. (Ver Jollife 1972,1973)

REDUÇÃO DA DIMENSIONALIDADE NA REGRESSÃO MÚLTIPLA

Na análise de regressão o interesse está na estrutura de dependência das variáveis. Variáveis altamente dependentes impõem muita imprecisão nas estimativas dos coeficientes de regressão. Com o objetivo de explicar a variável dependente toma-se aqueles componentes que apresentam as maiores correlações com ela. Na regressão múltipla, as correlações com cada variável dependente devem ser examinadas.

Seja a equação de regressão

$$Y = X\beta + \varepsilon ,$$

onde $\varepsilon \sim N (0 ; \sigma^2 H)$ e $H = I - n^{-1} 1 1'$. Este é um modelo recorrente, pois os erros são correlacionados.

Dado que os coeficientes nos CP têm um significado em termos de participação das variáveis originais X, pode-se expressar a regressão de CP como sendo:

$$y = W \alpha + \varepsilon$$

onde $W = XG$: transformação ortogonal dos CP
$\alpha = G'\beta$, $G\,G' = I$

Assim, dada a sua ortogonalidade, o descarte de alguns componentes, supondo que os últimos α_i são identicamente nulos, não altera os estimadores de mínimos quadrados $\hat{\alpha}_i$.

Breve referência ao comportamento assintótico dos CP

A base principal da construção dos componentes principais está na utilização dos autovalores e dos autovetores de Σ. A matriz de covariância Σ é uma matriz positiva definida com a informação da variância populacional e, como salienta Mardia(1979), sem a suposição de normalidade dificilmente conseguir-se-á encontrar a distribuição assintótica das raízes características e dos vetores associados a elas. Se os dados amostrais são provenientes de uma população com $N \longrightarrow \infty$, os autores preferem assumir os autovalores e os autovetores associados da matriz de covariância amostral S não como estimadores, $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, mas como os próprios autovalores $\lambda_1, \ldots, \lambda_p$.

Com base na suposição de que X é uma normal multivariada utilizar-se-á

* S como o estimador de Σ

* as raízes de $| S - dI | = 0$ são os estimadores de δ_i

A matriz de covariância Σ pode ser reescrita como $\Gamma' \Delta \Gamma$, onde Γ é uma matriz ortogonal com $\gamma_{ii} > 0$ e Δ é uma matriz diagonal com $\delta_1 \geq \ldots \geq \delta_p$. A representação é única se as raízes δ_i são todas diferentes e não existe γ_{ii} nulo. A matriz de covariância amostral S pode também ser representada por $C' D C$ e a representação é única com probabilidade 1. Se a representação populacional é assumida como única, C é o estimador de máxima verossimilhança de Γ e $(n/N)D$ é o estimador de máxima verossimilhança de Δ .

Tem-se, então, que se as raízes de Σ são diferentes entre si, os desvios $(d_i - \lambda_i)$ e $(c_i - \delta)$ são assintoticamente, normalmente distribuídos.

Se as q_r últimas raízes são asssumidas como iguais é outro caso em que se pode tirar uma posição com base nessa suposição.

Utilizando a maior das menores raízes para testar a hipótese de igualdade das raízes populacionais tem-se que a distribuição assintótica da maior das q_r menores raízes é normal e neste caso é possível querer determinar se estas raízes são pequenas o suficiente para que os componentes correspondentes sejam descartados . Um teste visando conhecer esta suficiência é dado pela razão da "variância não explicada" pelo total , isto é, quer-se detectar se ela não é maior que uma fração f,

$$\sum_{i=q+1}^{p} \delta_i \leq f \sum_{i=1}^{p} \delta_i$$

para algum q e que pode ser testada através do valor amostral, onde

$$\sum_{q+1}^{p} d_i - f \sum_{1}^{p} d_i =$$

$$= \sum_{q+1}^{p} d_i - f \sum_{1}^{q} di - f \sum_{q+1}^{p} d_i =$$

$$= - f \sum_{1}^{q} d_i + (1 - f) \sum_{q+1}^{p} d_i ,$$

o qual é assintóticamente normal.

Segundo Krzanowski(1985), em ACP não existe nenhuma estrutura imposta a priori na matriz de dados originais e ela é tratada simplesmente como uma dispersão de n pontos em v dimensões. Olha-se para uma rotação dos eixos tal que a variância total das projeções dos pontos no primeiro eixo é um máximo, a do segundo eixo é ortogonal para o primeiro e contém tanto quanto possível da variância remanescente, e assim por diante. Mesmo assim, diz Krzanowski, a estrutura não está sendo considerada. Hotelling(1933)

considerou apenas os sistemas de componentes normalmente distribuídos e os conjuntos de dados que tem o quarto momento nulo. Esta categoria de informações possibilita o estudo do comportamento assintótico dos autovalores e autovetores.

T.W.Anderson(1963) fez um estudo sobre a teoria assintótica de CP para dados calculados a partir de uma matriz de covariância amostral quando as observações provêm de uma distribuição normal multivariada cuja matriz de covariância tem autovalores de multiplicidade arbitrária. É possível que a dimensionalidade seja reduzida de p para r tal que r < p, se for provado que (p - r) raízes são iguais e aproximadamente nulas.

No caso normal, considera-se importante o uso de duas aproximações das quais David Tyler(1983) faz referências em seu artigo. James(1960) fez estudos que podem ser considerados precursores na questão da densidade conjunta exata das raízes amostrais e G.A.Anderson (1965) demonstrou que esta densidade tem função hipergeométrica. Estes serviram de base para a formulação dada por Muirhead(1978), ele verificou que a função hipergeométrica da densidade é assintoticamente normal.

Outra aproximação foi apresentada primeiramente por Girshick(1939) e T.W.Anderson(1963) que em resumo refere-se à expansão das raízes amostrais em torno da matriz de covariância populacional. Tyler(1983) propôs uma classe de estimadores denominados M que são estimadores "affine" invariantes, assintoticamente normais possuindo certas propriedades de invariância.

Para a "performance" desta classe de estimadores, Tyler propos matrizes aleatórias simétricas esfericamente invariantes.

$$Z_n = n^{1/2} \Gamma_n (S_n - \Sigma_n) \Gamma'_n$$

Estas matrizes, Z_n, satisfazem as seguintes restrições:

(1) A distribuição de Z_n é invariante sob a transformação $Z_n \longrightarrow Q\ Z_n\ Q'$ para qualquer Q ortogonal.

(2) $Z_n \longrightarrow Z$ em distribuição onde Z é normal multivariada.

Os estimadores M-affine invariantes foram definidos por Maronna(1976) como estimadores de locação e de dispersão.

Seja

$Y_1, \ldots, Y_n$ amostras aleatórias do vetor Y distribuídas elipticamente.

$$n^{-1} \sum_i u^1(d_i)(y_i - \mu_n) = 0$$

$$n^{-1} \sum_i u^2(d_i)(y_i - \mu_n)' = S_n$$

onde $d_i^2 = (y_i - \mu_n)' S_n^{-1} (y_i - \mu_n)$

u_1 e u_2 são funções que satisfazem as suposições gerais feitas por Maronna(1976).

Então

(μ_n, S_n) são estimadores dos parâmetros (μ, Σ).

Maronna(1976) mostrou que $n^{1/2}(S_n - \Sigma) \longrightarrow N$ em distribuição sendo que N tem uma distribuição normal multivariada com média zero.

Tyler(1983) demonstrou que "a distribuição assintótica das raízes de S_n é encontrada sob a seguinte seqüência de alternativas locais de raízes múltiplas.

$$\frac{n^{1/2}\,(\,\lambda_i\,(\Sigma_n) - \Gamma_{m,n}\,)}{\Gamma_{m,n}} \longrightarrow \begin{cases} \infty, & i \in \varphi_r\ ,\ r < m \\ d_i, & i \in \varphi_m \\ -\infty, & i \in \varphi_r\ ,\ r > m \end{cases}$$

onde $\varphi_1, \ldots, \varphi_k$ é uma partição do conjunto $(1, \ldots, p)$ com $\varphi_m = (i_{(m)}, i_{(m)+1}, \ldots, i_{(m)+q(m)-1})$ e $\Gamma_{m,n}$ é a média de $\lambda_i(\Sigma_n)$ sobre $i \in \varphi_m$, isto implica que $\frac{\lambda_i(\Sigma_n)}{\lambda_j(\Sigma_n)} \longrightarrow 1$ se $i \in \varphi_m$ e $j \in \varphi_m$ e que $\sum_{i \in \varphi_m} d_i = 0$.

Jonsson(1982) realizou estudos sobre teoremas limites para autovalores da matriz de covariância amostral quando a dimensão da matriz assim como o tamanho da amostra tende a infinito. Foi utilizado o método dos momentos como limite da função de distribuição acumulada dos autovalores amostrais. Através desse método concluiu que a soma dos autovalores, até, o k-ésimo posto $k = 1, \ldots, m$ é assintoticamente normal.

Dauxois, Pousse e Romain(1982) através de um estudo sobre a teoria assintótica da função de um vetor aleatório para a ACP colocam que: "A ACP de um processo é definida num espaço linear infinito-dimensional, portanto o uso da teoria de matrizes é impossível. Conseqüentemente duas dificuldades aparecem: a primeira é devida à dimensão e a segunda para o fato que o processo pode não ser um único escalar ". Neste trabalho foi utilizade o "esquema de dualidade" onde qualquer espaço Hilbert e seu dual é identificado. Através dele, foram feitas algumas aplicações para a Inferência Estatística: estimação no ponto; intervalos de confiança para o principal valor, para a variância total.

* Com estes resultados foram construídos testes para a razão da variância explicada e para o principal fator.

* Tyler(1983) apresenta a hipótese de que um conjunto de vetores cai no subespaço expandido por um prescrito subconjunto de vetores CP para uma população normal. Os testes são derivados para o subespaço expandido por um conjunto de vetores CP que a matriz dos valores do estatístico tem uma distribuição assintótica Wishart. A suposição de normalidade neste caso não é tão rigorosamente observada quando do uso da matriz de covariância amostral. Novamente Tyler busca a generalização para qualquer estimador de dispersão M-affine invariante para populações elípticas.

Silverstein(1984) comprovou que M_n converge para uma N(0,1) mesmo em casos em que a matriz não é um caso de matriz Wishart. Ele fez estudos dos teoremas limites para autovetores de matrizes de covariância de grande dimensão .

Boente(1987) demonstra que " A distribuição assintótica dos autovalores da matriz de dispersão robusta proposta por Maronna em 1976 é dada quando as observações provém de uma distribuição elipsóide. Os elementos de cada vetor característico são os coeficientes de uma versão robustificada dos Componentes Principais. Dá uma definição para a eficiência assintótica destes estimadores e avalia sua influência na curva. O problema de maximizar a eficiência da curva é resolvido. No entanto os estimadores são ótimos sob a suposição de distribuição normal multivariada".

Ruymgaart(1981) apresenta uma ACP robusta para o caso de função de distribuição bivariada. O ponto de partida são os estimadores robustos para a dispersão no caso univariado estendido para a estrutura bivariada. "Ao lado da continuidade funcional definindo a direção de uma aceitável modificação no eixo principal, prova a consistência da seqüência correspondente de estimadores"

Novamente estes resultados são estabelecidos sob uma

normalidade assintótica atingida sob restrições adicionais onde a sequência é composta de elementos aleatórios identicamente distribuídos em alguma função no espaço.

Para os casos de não-normalidade, diversos estudos foram realizados com a finalidade de conhecer o comportamento dos CP.

Sugiura(1976) derivou uma aproximação subassintótica para a distribuição das raízes amostrais a qual foi utilizada também em casos para os quais o quarto momento é não nulo.

Waternaux(1976) e David(1976), pesquisando a distribuição assintótica das raízes características da matriz de covariância amostral em populações não normais, utilizaram esta aproximação. Mais recentemente, Fujikoski(1980) encontrou uma distribuição subassistótica para o caso não normal, sendo que este estudo considera tão somente os casos de raízes distintas. Isto é, se l_i : raíz característica de S então $0 < l_i < \ldots < l_p$. Waternaux (1976) realiza um estudo e comprova que retiradas de normalidade afetam a distribuição das raízes características de S. A teoria assintótica e os resultados amostrais indicam que esta distribuição apresenta alterações considerando-se os quartos cumulantes não nulos da população parente. Esta é consistente com outros estudos sobre robustez de estatísticos multivariados Mardia(1971) e Layard(1972,1974) para o teste de igualdade das matrizes de covariância.

Em aplicações tal como ACP e Análise Fatorial, são realizados testes usando os momentos assintóticos derivados para a população normal multivariada. Estes testes são incorretamente empregados quando se trata de populações não normais. Existem estudos para correções que podem ser aplicadas para o caso de quarto cumulamente não nulo.

Waternaux(1976) desenvolveu o estatístico W_q para testar hipótese de igualdade de "q" raízes características da matriz

de covariância. O que foi observado neste estudo é que W_q não é robusto para retiradas de normalidade e não pode ser utilizado em distribuições de caudas longas. Outro estatístico W^*_q foi então, desenvolvido e estudado, para o caso de distribuições elípticas em espaços de dimensões mais reduzidas.

UMA APLICAÇÃO DE ACP EM ANÁLISE DE DADOS DE SOBREVIVÊNCIA

Segundo RAO(1965) " O método de Componentes Principais têm se mostrado um útil instrumento para resolver o problema de identificação de fatores essenciais afetando os resultados de um tratamento médico." Com este pressuposto Danielyan, Zharinov e Osipova (1986) empregaram CP para a identificação e interpretação das variáveis segundo a participação das mesmas através do grau dos coeficientes associados a cada uma delas. Pretende-se delinear os principais fatores relacionados com o tempo de sobrevivência dos pacientes portadores de câncer cervical. O tratamento estatístico adotado deve servir para selecionar os fatores que exercem as maiores influências pelo índice de participação nos componentes.

A matriz de dados originais é composta de 12 variáveis. Este conjunto de variáveis contemplam três aspectos: a caracterização do paciente, do tumor e do método de tratamento. Ao todo, 603 pacientes com câncer cervical foram observados segundo estas 12 variáveis, pacientes estes submetidos a tratamentos de terapia de radiação no "Central Research Institute of Roentgenology and Radiology ".

Com a aplicação de ACP o conjunto de variáveis passou por uma redução de sua dimensionalidade e 4 Componentes Principais foram selecionados, tendo sido considerados suficientes com sua participação de 63% na variação total.

A análise dos coeficientes, autovetor c_1 associado à λ_1, autovalor de Y_1 o primeiro CP demonstrou resultados um tanto

surpreendentes: os fatores com os maiores coeficientes em Y_1 são aqueles relacionados com a própria doença enquanto que em Y_2 o segundo CP evidencia principalmente os fatores relacionados com o tratamento. Uma vez que Y_1 concentra o máximo de variabilidade pela própria definição de ACP, pode-se atribuir neste o maior poder discriminatório para os pacientes aos fatores biológicos que caracterizam a situação do organismo do paciente quanto ao tumor.

Segundo o Danielyan et allii(1986) "Após estabelecer quais fatores iniciais entram linearmente neste Componente Principal e com grandes pesos, o método permite interpretar o componente como um fator generalizado retirado por grupos de características as quais são mais significantes para a previsão dos resultados da terapia de radiação.

Observe-se o quadro com os resultados dos quatro Componentes Principais selecionados quando da redução da dimensionalidade de p = 12 para q = 4:

Tabela nº 1 Matriz de fatores Característicos Componentes Principais (63% da variação)

	COMPONENTES PRINCIPAIS			
	I	II	III	IV
	26 %	18 %	10 %	9 %
1. Idade dos pacientes	0.05	-0.20	-0.79	-0.20
2. Tipo histológico do tumor	-0.01	0.06	-0.24	0.07
3. Estágio da doença	0.71	-0.12	0.37	-0.15
4. Volume do tumor	0.73	-0.25	0.36	-0.15
5. Tempo doubling do tumor	-0.60	0.13	-0.05	0.19
6. Dose de irradiação intracavitária no ponto A	-0.15	0.82	0.03	-0.19
7. Dose combinada de irradiação no ponto A	-0.13	0.80	0.07	-0.42
8. Dose de irradiação remota no ponto B	0.079	0.35	0.15	-0.51
9. Dose integral absorvida	0.022	0.04	-0.06	-0.82
10. Razão de atividade endocervical/Endovaginal	-0.02	0.67	-0.10	0.13
11. Tempo de falha	-0.87	-0.05	0.28	-0.20
12. Tempo de recorrência	-0.89	-0.05	0.27	-0.17

As conclusões do trabalho demonstram uma importância fundamental das características biológicas em primeiro lugar e do tipo de tratamento, dose e local de irradiação realizada em segundo lugar, para a sobrevivência dos pacientes com esse tipo de tumor.

Assim estes autores utilizaram ACP para detectar quais os fatores determinantes, em primeira instância, na previsão do tempo de sobrevivência dos pacientes, sendo o grau de importância mensurado pelo peso dos coeficientes nos primeiros CP. Neste caso, ACP além de fornecer elementos para a redução da dimensionalidade original ($p = 12$ para $q = 4$, garantindo 63 % de variação total explicada) foram utilizados os coeficientes como elementos determinantes na seleção de variáveis para compor a análise previsiva em relação a doença relatada.

INTRODUÇÃO

A aplicação dos **Componentes Principais - CP** implica em acompanhar e avaliar como determinadas mudanças nas observações provocam alterações nos resultados do método, isto é alteram os autovetores que em última instância serão utilizados na análise.

Efetivamente, não interessa só conhecer a sensibilidade do próprio método de Análise de Componentes Principais ante a presença destes elementos perturbadores, bem como interessa saber como ela pode ser utilizada no processo de identificação dos elementos perturbadores. Estes elementos perturbadores decorrem possivelmente da mobilidade na população-alvo, do uso de amostragem gerar o conhecimento da estrutura de variação dos dados e a estimação dos próprios resultados de ACP a partir de uma Matriz de Covariância amostral, ocasionam mudanças significativas na variabilidade. A própria seleção e descarte de variáveis de interesse que podem flutuar à medida em que o tempo avança, apóiam a decisão de buscar um aprofundamento no estudo das pertu bações que apresentam os coeficientes quando sujeitos a estas situações.

Qual a sua estabilidade?
Qual é a sua estabilidade nestes casos?
Como detectar se foram ou não estáveis?

Entender o comportamento dos Componentes Principais é um tema que gerou diversos estudos buscando, principalmente, avaliar o efeito na variância, de arredondamentos dos coeficientes e de zerá-los. Green(1977) e de Bibby(1980) pesquisaram as distribuições limites para a variância nos casos acima. Mas, acompanhar o comportamento dos coeficientes em função de pequenas retiradas de otimalidade na variância, de arredondamentos dos dados originais, da ocorrência de "outliers" preocupa muito mais pois na maioria dos casos, mesmo se tratando de amostras provenientes de populações com $N \longrightarrow \infty$, os resultados dos Componentes Principais serão inferidos a partir de estimadores, e a perda de otimalidade na variância é um fator a ser muito considerado. As interpretações serão feitas a partir destes coeficientes, portanto, é necessário conhecer a sua estabilidade. Krzanowski(1984) propôs uma medida dessa estabilidade, chamada Análise de Sensibilidade e que deverá acompanhar a Análise de Componentes Principais.

Avaliar-se-á as formas de mensurar a estabilidade, observando o comportamento destas técnicas e sua capacidade de analisar a estabilidade dos autovetores em dados provenientes de populações com distribuições desconhecidas mas com um tamanho bastante grande, tendendo a infinito. Outros estudos que estão sendo apresentados envolvem a presença de "Outliers" na amostra e sua influência nos coeficientes e a questão dos arredondamentos.

Para avaliar a solidez destes esclarecimentos serão empregadas técnicas de simulação utilizando os diferentes tipos de alterações propostas neste trabalho. Após a geração de dados com os diferentes casos de interesse, será aplicado o método de ACP para estudar o comportamento de seus coeficientes frente à estas perturbações. Algumas conclusões poderão ser tiradas quanto à estabilidade dos coeficientes, c_i, quando se verifica qualquer perturbação, seja na variância com pequenas retiradas de otimalidade na matriz de covariância, seja pela presença de "Outliers", seja pelo uso de Arredondamentos nos dados originais. Entre as situações observadas, poderá ser percebido, em especial, o comportamento do primeiro Componente Principal quanto à sua estabilidade, quaisquer que sejam as alterações que forem procedidas. O último Componente Principal parece ser um elemento chave na Análise de Sensibilidade, na identificação de "Outliers", e principalmente na observação de sua estabilidade quando ocorreram perturbações na variância de um tamanho máximo ϵ.

Apresentar-se-á um estudo com dados coletados numa indústria de produção de balas comestíveis. Trata-se de uma aplicação de Componentes Principais cujo objetivo é analisar os fatores que interferem na qualidade do produto final para reduzir perdas no processo produtivo. Dois aspectos serão enfocados neste estudo: a análise de fatores ambientais interagindo com a umidade da bala e o estudo laboratorial das condições da massa em termos de doçura, acidez e umidade. Uma aplicação em termos de Análise de Componentes Principais-Regressão pode ampliar a análise dos fatores em discussão. Estas aplicações devem ser acompanhadas de uma Análise de

Sensibilidade. Verifica-se, assim, na prática a aplicabilidade da técnica de análise de Sensibilidade. Os componentes mais sensíveis não são sempre os mesmos apesar de haver uma forte tendência em apresentar o 1º e o último Componente Principal como os mais estáveis e os componentes intermediários como os mais perturbados.

CAPÍTULO 1

O ESTUDO DA SENSIBILIDADE DOS COMPONENTES PRINCIPAIS

Vários métodos de Análise Multivariada, Análise Fatorial, Análise Discriminante, Análise de Componentes Principais, utilizam uma transformação linear do tipo $Y_i = c_i' X$. A semelhança entre eles está na escolha de uma função V, a partir da qual serão

calculados os autovalores e os autovetores. Na sua grande maioria estas funções são quadráticas em c. Segundo Krzanowski(1971) "Sua otimização leva para a solução de uma equação de autovalor/autovetor, onde cada autovalor leva o valor de V à um ponto estacionário e o correspondente autovetor providencia os coeficientes apropriados."

Um modelo estatístico para o estudo de um processo envolvendo p variáveis é ACP que possibilita a redução da dimensionalidade do conjunto de variáveis que estão sendo estudadas e sua recomposição, apresentando o grau e a forma de participação de cada variável no vetor resposta, permite que seja interpretado o tipo de informação contida em cada CP. Daí a responsabilidade do estatístico em avaliar se pequenas alterações na estrutura de dados correspondem à alterações nos coeficientes afetando a interpretação dos mesmos. As decisões serão tomadas a partir desses resultados. Muitas interpretações são realizadas sem avaliar com rigor a sensibilidade do método no caso de ocorrerem perturbações nas condições ideais de sua aplicação.

Alguns conceitos básicos são definidos em função de favorecer seu emprego no decorrer do texto.

O método que serve de base para este estudo é ACP. Segundo Anderson(1984) " **Componentes Principais são combinações linearizadas de variáveis aleatórias que têm propriedades especiais em termos de variância.**"

Trata-se, basicamente, de um estudo sobre a estabilidade e perturbações em ACP. As perturbações são elementos da estrutura de dados, conjunturais ou impostos, que modificam as condições ideais de aplicação do método. Neste contexto, sensibilidade é o reflexo que os componentes apresentam frente à ocorrência de elementos perturbadores. Assim um componente estável será aquele que não se alterar com a ocorrência de perturbações, ou, então, apresentar mudanças estatisticamente insignificantes.

Entre os elementos que podem afetar os resultados de ACP, foram selecionados os seguintes:

* ARREDONDAMENTOS

Algumas vezes para favorecer a análise dos dados é necessário fazer uso de aproximações. Esta atitude produz uma modificação na estrutura de variância. Qual a modificação que esta mudança provoca nos CP? A perturbação pode alterar significativamente seus coeficientes? Em que casos isto vai ocorrer?

* PRESENÇA DE "OUTLIERS"

A presença de "outliers" provoca alterações na variabilidade dos dados. Como os CP se comportam neste caso se sua estrutura está definida em função da análise da concentração e dispersão das informações ? Se os CP são sensíveis à presença de "outliers", quais os mais sensíveis. Quais os que apresentam maior estabilidade? Como esta sensibilidade aos "outliers" pode ser utilizada em ACP?

* MUDANÇAS NA VARIÂNCIA

Quais as mudanças que apresentam os CP quando a variância apresenta perdas de tamanho máximo e em seu valor real, implicando numa perda de otimalidade? Quais as alterações nos coeficientes quando as perturbações são do tipo aumento ou diminuição da variância? Existe uma forma de mensuração destas alterações para avaliar seus efeitos? Quais os CP que apresentam, de um modo em geral, maiores alterações nesse caso? Quais os componentes que são menos sensíveis à perturbações desse gênero? Como realizar a Análise de Sensibilidade em cada caso, simultaneamente, com a ACP?

1.1 REFERÊNCIAS BIBLIOGRÁFICAS SOBRE APLICAÇÕES E EVOLUÇÃO DE CP

ACP têm sido aplicado com eficiência em muitos campos e com as mais diversas finalidades . Alguns artigos foram selecionados com a finalidade de acompanhar a contribuição que está sendo dada com vistas à evolução da aplicação de ACP. A Análise de Sensibilidade de ACP deve considerar seu emprego e sua evolução, uma vez que alguns procedimentos clássicos podem já estar superados.

1.1.1 ACP EM MÉTODOS DE CONTROLE DE QUALIDADE

Jackson(1959) utilizou a ACP para melhorar a determinação dos níveis de Controle de Qualidade de um processo de produção. Desta forma pretende ele ampliar os níveis de aplicação da estatística T^2 de Hotelling. A interpretação dos coeficientes, c_i, apresentados pelos CP encaminha a um tratamento estatístico posterior e à uma análise da representatividade dos mesmos.

Jackson(1959) " O método de Componentes Principais é introduzido tanto como um método de caracterizar um processo multivariado assim como um instrumento de controle associado com os procedimentos de controle." A questão principal que se coloca é no processo de produção. "O processo está sob controle?" Esta questão não corresponde na maioria dos casos a avaliar o comportamento univariado dentro das linhas do controle. Na verdade, quase sempre duas ou mais variáveis precisam simultaneamente estar sob controle. O que comumente se entende por isso é a observação simultânea do comportamento das mesmas em dois ou mais gráficos de controle, construídos a partir das observações feitas. Isto é, observa-se comparativamente comportamentos univariados. Como coincidir a definição de "fora do controle" especificado pelas linhas de controle, individualizadas por gráfico? E, não só isso, como construir o "fora de controle" conjunto do elemento caracterizado considerando simultaneamente o mesmo nas p direções? Então poder-se-á considerar um ponto como fora do controle quando o processo está

exatamente no padrão estabelecido. O tipo de erro que se estabelece, neste caso, caracteriza uma falha onde ela realmente não existe. É um erro do tipo I, cujo valor máximo pode ser pré-fixado num tamanho α. Considerando para cada variável do processo p-dimensional, $\alpha = 0.05$, o tratamento univariado, sendo p=2, faz com que a probabilidade de que o ponto esteja sob controle em ambos tenha efeito multiplicativo. Pode-se considerar a independência das observações neste caso, por isso $P[\ (X_1 \in I_1) \cap (X_2 \in I_2)\ ] = P(X_1 \in I_1).P(X_2 \in I_2)$, onde X_i: variável e I_i:região de controle, i=1,2 e como tal obter-se-á $(0.95)(0.95)=0.9025$ elevando dessa maneira a probabilidade de Erro Tipo I para aproximadamente 10 %. Este tipo de erro, portanto, cresce com o crescimento da dimensionalidade levando as conclusões obtidas ao descrédito.

Jackson(1959) propõe o estabelecimento de uma região elipsóide definida a partir da rotação dos eixos e que deixa qualquer ponto em seu perímetro equiprovável. A rotação dos eixos é na realidade uma rotação ortogonal que permite o tratamento de variáveis, ainda que altamente correlacionadas pois esta rotação leva à variáveis não correlacionadas. " O eixo maior é chamado de linha de regressão ortogonal em que é minimizada a soma de quadrados perpendiculares a esta linha. O comprimento do semi-eixo maior é igual a $\sqrt{\lambda_1 T^2_\alpha}$ e o comprimento do eixo menor é igual à $\sqrt{\lambda_2 T^2_\alpha}$ onde λ_1 e λ_2 são ambos raízes da equação :

$$\lambda = \frac{[\ (s^2_x + s^2_y) \pm \sqrt{(s^2_x + s^2_y)^2 - 4(s^2_x s^2_y - s^2_{xy})}}{2} \qquad (1)$$

e

$$T^2_\alpha = \frac{2(N-1)}{N-2} F_\alpha \qquad (2)$$

onde F tem $n_1 = 2$ e $n_2 = N - 2$ graus de liberdade, $s^2_x s^2_y$ e s_{xy} tem sido obtidos de um período base amostral de tamanho N. " Isto se estiver sendo considerado p = 2, um espaço bidimensional.

A região sob controle definida no gráfico de controle T^2 de Hotelling tem identidade com a região definida pela elipsóide. Coincidem na definição dos pontos que estão "sob controle" e dos que estão "fora do controle". A rotação que dá origem a esta elipsóide resulta da aplicação de uma transformação nos dados originais denominada ACP.

Três métodos alternativos de controle são assim considerados: Registro gráfico das observações originais, Gráfico T^2 para a quantidade $x\ s^{-1}\ x'$ e Gráfico T^2 para a quantidade $y\ y'$. Com o primeiro método visualiza-se graficamente o ponto fora de controle. O segundo e o terceiro métodos, além da escala de tempo que é preservada, também expressam, por meio de um número, a condição do processo multidimensional. O procedimento de T^2 é baseado na distribuição T^2 de Hotelling e requer que os dados sejam provenientes de uma distribuição normal multivariada com uma matriz de covariância conhecida.

Como ilustração, Jackson apresenta um estudo de Controle de Qualidade Multivariada em teste balístico de míssil. Este trabalho prático proposto como aplicação de ACP no Controle de Qualidade Multivariada baseia-se no estudo da componente desempenho de um míssil balístico segundo o impulso produzido durante o disparo.

Tabela nº 1.1 - Conjunto de autovetores do estudo de Jackson

		COMPONENTES PRINCIPAIS			
		I	II	III	IV
V					
A	X_1	0.0256	-0.0897	-0.1055	-0.0642
R.	X_2	0.0332	-0.0258	0.1403	-0.0361
D	X_3	0.0251	0.0200	-0.0310	0.2127
R	X_4	0.0254	0.1081	-0.0483	-0.1015

Como se pode observar, o primeiro CP apresenta quase que uma igualdade de pesos assumidos pelos coeficientes de cada variável original configurando uma média aritmética dos dados. Seja

$$Y_1 = 0.0256\ X_1 + 0.0332\ X_2 + 0.0251\ X_3 + 0.0254\ X_4$$

Quanto à Y_2, o segundo CP trabalha com a diferença entre as duas formas de mensuração adotadas em cada medidor, representando um contraste entre os dois. Quanto a Y_3 e Y_4 a relação entre as variáveis originais pode ser interpretada como privilegiando a diferença entre a variável que está associada ao mais alto coeficiente e as demais. Estas, se for pensado em termos de descarte de variáveis originais, seriam as indicadas para serem descartadas conforme o sistema apresentado por Mardia (1979). Olhando, então, para os componentes que apresentam as menores variâncias (Y_3 e Y_4) vê-se que as variáveis originais X_3 com peso $c_{34} = 0.2127$ referente a leitura do integrador com o medidor 2 e a variável X_2 com peso $c_{23} = 0.1403$ referente a medida do planímetro pelo medidor 1 poderiam ser descartadas. Esta indicação poderia estar indicando que a leitura dos dois medidores poderia ser reduzida a cada um apresentar um dos aspectos, um contra o outro, suficiente para fazer a checagem pretendida.

Uma transformação é, então, indicada como conseqüência da interpretação dos CP:

$$u_1 = \sum_{i=1}^{4} X_i \qquad u_2 = (X_1 + X_2) - (X_3 + X_4)$$

$$u_3 = X_1 - X_2 \qquad u_4 = X_3 - X_4$$

Para estas novas bases são requeridas três restrições:

- Ortogonalidade Com a ortogonalidade garante-se que a soma dos produtos cruzados dos coeficientes de quaisquer linhas devem ser iguais a zero.

- Correlação entre Y_i e u_i Os CP Y_i e as transformações definidas a partir de sua interpretação u_i devem ser altamente correlacionadas indicando que a transformação é uma boa aproximação da informação contida nos mesmos. A definição de u_i como função dos X_i's uma vez

que é função dos CP que por sua vez também são função dos X_i`s deve garantir alta correlação entre eles.

- Independência Como as Y_i`s são definidas garantindo que os CP são independentes, $Cov(Y_1,Y_2)=0$ , requer-se que as $u_i$`s também sejam independentes. Porém como é difícil que essa independência seja garantida, é sugerido que a restrição se atenha a garantir covariância mínima.

Dado que

$Y\ Y' \sim T^2$ de Hotelling ,

se as u_i`s são independentes, então:

$$\sum_{i=1}^{p} \left[\frac{(u_i - \bar{u}_i)^2}{S_{ui}^2} \right] \sim T_u^2 \qquad (3)$$

Então a aproximação é vista como a melhor pois apresenta uma distribuição T^2. Também existe uma correlação direta quase perfeita, isto é, tendendo a 1 entre os valores apresentados por T^2 e T_u^2. Pode ser demonstrado que se a aproximação a ser utilizada é um instrumento para o controle de qualidade multivariado então, este tipo de aproximação, principalmente a aproximação T^2, é a que melhor se ajusta ao caso.

1.1.2 ACP POR REGRESSÃO NA PESQUISA EXPLORATÓRIA ESTATÍSTICA

Conforme Draper(1981) afirma em seu livro "Applied Regression Analysis", os clássicos modelos preditivos de Análise de Regressão baseados no uso de estimadores de mínimos quadrados calculados utilizando apenas as informações contidas nas variáveis originais esbarram, muitas vezes, em problemas como o da

autocorrelação dos resíduos, problemas de multicolinearidade para os quais a "Ridge regression" é uma técnica alternativa de superação. As dificuldades residem principalmente no desconhecimento da variabilidade - σ^2 - e na altísssima intercorrelação entre as variáveis X, uma questão de multicolinearidade. Isto impede que as contribuições individuais sejam avaliadas isoladamente, prejudicando o próprio modelo preditivo ao superpor informações. Necessita-se obviamente de uma outra solução que pode vir na forma da contribuição de métodos de Análise Multivariada construídos com estas perspectivas. Em se tratando de um caso em que a estrutura de correlação é imprescindível para o próprio sucesso do modelo preditivo, uma alternativa que se apresenta neste caso específico é o método de ACP cujos pressupostos incluem Cov (X_i, X_j) = 0 , $i \neq j$.

Este método permite analisar a estrutura de correlação em alguns detalhes e neste caso é possível buscar uma solução conjunta dos CP e da Regressão. O processo da aproximação de ACP-R determina que inicialmente seja aplicada a transformação CP, com a redução da dimensionalidade do conjunto de variáveis originais, se assim for o caso, para depois aplicar o modelo preditivo definido para Regressão com as variáveis já transformadas.

Massy(1965), partindo de idéias de Kendall e de alguns resultados de Stone, combinou as aplicações de ACP com métodos de Regressão tentando resolver problemas com a multicolinearidade. Massy, após realizar uma aplicação também para desenvolver uma comparação entre o tratamento de Regressão clássica e a ACP-R concluiu que esta aproximação é muito boa e é um bom ajuste no tratamento exploratório de relações complexas entre variáveis especialmente quando é o caso de colinearidade. O cálculo da regressão, após a aplicação de CP facilita a Análise Exploratória. A transformação CP vai permitir que seja simplificado o manuseio deste tipo de variáveis.

Massy(1965) afirma que o uso de Regressão Multivariada clássica não é indicado nos seguintes casos: " (...) i) quando as

variáveis independentes são colineares com alguma outra, tornando impossível a inversão da matriz de correlação e os elementos beta indeterminados; ii) quando, por causa da alta (mas não completa) colinearidade ou por alguma outra razão é desejável o colapso do espaço das variáveis independentes pela deleção de um ou mais Componentes Principais da relação de Regressão."

No primeiro caso, trata-se do posto da matriz , se alguns dos vetores são dependentes , ou então, podem ser considerados como tal, o posto da matriz é reduzido de p para q de tal modo que q<p. No entanto, a análise de Regressão Clássica tem solução indeterminada sob estas condições, estimulando o uso de CP pelas suas propriedades quanto a redução da dimensionalidade. Então, " (...) é possível estimar os parâmetros de Regressão sobre as projeções das variáveis originais no m-plano do espaço E^n expandido pelas linhas de Z ." A deleção dos CP quando os λ_js são não nulos pode afetar o cálculo de Regressão no sentido de que ocorre uma redução na informação a ser trabalhada. No entanto, o caso clássico de Regressão Multivariada apresenta problemas mais sérios . Assim após o uso de ACP, a deleção, se este for o caso, deve ocorrer considerando dois níveis:

i- Predição das variáveis originais independentes X

Deleta-se as variáveis aleatórias que têm os menores autovalores, as que têm o menor poder explicativo em termos de variação total, o que as coloca como dispensáveis em termos de previsão.

ii- Predição das variáveis dependentes Y

Deleta-se aquelas que apresentam menor correlação entre os CP e Y. Este critério apresenta uma relação maior com os estudos para análise exploratória de dados pois se privilegia a necessidade de prever as variáveis dependentes.

No estudo realizado por Massy os focos centrais em termos de variáveis dependentes são posse de televisão, posse de refrigerador, calefação central e Superpopulação. Estas variáveis qualitativas são tratadas em termos de proporção afim de se adaptarem à aplicação de ACP.

Com o propósito de comparar o desenvolvimento da análise de regressão clássica , que faz uso do método dos mínimos quadrados como modelo preditivo da Regressão Multivariada Clássica, com o emprego da ACP-R é necessário conhecer o modelo delineado para o estudo da Regressão nos dois casos.

- Análise de Componentes Principais - Regressão(ACP-R)

Os resultados são obtidos por região e por intervalo, assim o modelo preditivo para tratamento ACP-R foi padronizado pela média de saturação e é um modelo aditivo. Senão, observe:

- Relativo ao rendimento:

$$Y_{Ii} = \sum_{j=1}^{14} b_{Ij} (P_{Iij} - \overline{P}_{Ij}) + \overline{Y}_i + e_i \qquad (4)$$

- Relativo à educação

$$Y_{Ei} = \sum_{j=1}^{9} b_{Ej} (P_{Eij} - \overline{P}_{Ej}) + \overline{Y}_i + u_i \qquad (5)$$

Considerando conjuntamente rendimento e educação num modelo preditivo, faz-se (4) mais (5) obtendo-se $Y_i = Y_{Ii} + Y_{Ei}$, onde Y_i é a reta de regressão observada na região i, i = 1, ... , n. Então:

$$Y_{Ii}+Y_{Ei} = \sum_{j=1}^{14} b_{Ij}(P_{Iij} - \overline{P}_{Ij}) + \overline{Y}_i + e_i + \sum_{j=1}^{9} b_{Ej}(P_{Eij} - \overline{P}_{Ej}) + \overline{Y}_i + u_i$$

$$2\,Y_i = \sum_{j=1}^{14} b_{Ij}(P_{Iij} - \bar{P}_{Ij}) + \sum_{j=1}^{9} b_{Ej}(P_{Eij} - \bar{P}_{Ej}) + 2\,\bar{Y}_i + e_i + u_i$$

$$Y_i = \sum_{j=1}^{14} (1/2\, b_{Ij})(P_{Iij} - \bar{P}_{Ij}) + \sum_{j=1}^{9} (1/2 b_{Ej})(P_{Eij} - \bar{P}_{Ej}) + \bar{Y}_i + v_i \qquad (6)$$

onde $v_i = \frac{e_i + u_i}{2}$.

Os coeficientes de Regressão - β_i - representam o efeito dos desvios nas distribuições de rendimento e de educação de uma determinada região. Os coeficientes de Regressão conforme foram avaliados a partir de (6) são estimados por

$$\beta_i = (1/2\, b_i)(\sigma_X / \sigma_Y) \qquad (7)$$

O modelo adotado permite:

i) através do perfil dos betas avaliar os efeitos da saturação-renda e da saturação-educação, isto é, foi possível uma base "a priori";

ii) os resultados de ACP-R podem ser comparados com os resultados da Regressão Clássica pois são construídos na mesma base;

iii) uma vez obtidos os perfis dos betas, é possível avaliar a sensibilidade dos CP-Regressão através de retiradas no valor esperado de sua forma alisada.

O sucesso na aplicação de um modelo preditivo, no entanto, não está inteiramente dependente da capacidade preditiva intrínseca ao modelo adotado, mas absolutamente ligada à capacidade preditiva dos dados analisados baseadas em suas propriedades em termos de correlação. Se Y pode ser inteiramente ou, então, altamente explicado pelas variações de X e admitindo a concentração da dispersão dos dados nos primeiros CP, proporcionada pela própria

definição de ACP, pode-se ter qualidades superiores em termos de predição com o uso de ACP-R.

- Regressão Multivariada Clássica

A Regressão Multivariada Clássica pode ser avaliada através do seguinte modelo, diferente do modelo utilizado na ACP-R porque os conjuntos rendimento e educação - são colineares , cada uma das somas parciais em termos de proporção devem somar um. Assim,

$$Y_i = c_0 + c_1\bar{I}_i + c_2\bar{I}_i^2 + c_3\bar{E}_i + c_4\bar{E}_i^2 + e_i \quad (8)$$

Conclusões

a) A maioria das conclusões obtidas através de ACP-R sobre os efeitos dos rendimentos e da educação são semelhantes aos obtidos com o método clássico;

b) ACP-R apresenta R^2 maior, isto é, concentra maior poder explicativo pois utiliza menos regressores através de seu poder para descartar as possíveis duplicações de informações em variáveis altamente correlacionadas;

c) ACP-R permite uma seleção dos CP quando os perfis dos betas são utilizados para o estudo da estrutura. Quando avaliado em relação à pesquisa estatística exploratória pode-se concluir que o uso de ACP-R agiliza a percepção da participação das variáveis dependentes nos CP, o que pela própria natureza, implica em avaliar a participação dos mesmos na variabilidade, além de relacionar os resultados da Regressão como projeção das variáveis originais.

Em geral, o maior problema no uso de ACP-R provém da dificuldade de interpretar seus coeficientes que neste caso refere-se também à interpretação dos coeficientes de Regressão - os betas. Entretanto o que se vem observando em relação ao uso de CP é que a

dificuldade de interpretação está mais relacionada com os menores autovalores, enquanto que os primeiros CP principalmente o primeiro tem uma interpretação quase mediata.

1.1.3 ACP EM DOIS ESTUDOS DE CASO POR JEFFERS

JEFFERS(1967) realizou um estudo para verificar a qualidade das vigas de sustentação, feitas de madeira, e que serão empregadas para sustentar o teto dos túneis de minas. Como se vê, no processo de produção destas vigas a escolha da madeira adequada é de fundamental importância e, por isso, a aplicação de ACP para estabelecer quais os fatores determinantes na seleção e qual o grau de participação de cada fator no processo de escolha das toras é fundamental. Aparentemente, um critério baseado na capacidade de sustentar um peso de 2.440 lb é utilizado como fator de corte, as toras com menor capacidade serão rejeitadas, caso contrário serão aceitas como suficientemente fortes.

Foi observada alta correlação, principalmente, entre as variáveis físicas e concluiu-se que seria difícil interpretar os coeficientes de Regressão parcial individualmente. Segundo Ehrenberg(1962), a possibilidade de observar a formação dos "clusters" a partir da Matriz de Correlação é um modo tão eficaz quanto aquela realizada a partir da ACP, porém este último método garante melhor classificação dos resultados.

Após a aplicação de CP verifica-se que as 6 primeiras componentes já contém 87 % da variação total sendo considerados suficientes para evidenciar o modelo de comportamento dos dados. Jeffers(1967) utilizou como critério para o descarte dos CP, o descarte daqueles que apresentam o autovalor, $\lambda < 1$. A exposição de alguns primeiros CP e a observação dos seus autovetores permite acompanhar a análise realizada pelo autor na interpretação dos coeficientes.

$$Y_1=0.96X_1+X_2+0.31X_3+0.43X_4+0.14X_5+0.7X_6+0.99X_7+0.72X_8+0.88X_9+0.93X_{10}-$$
$$-0.03X_{11}-0.28X_{12}-0.27X_{13}$$

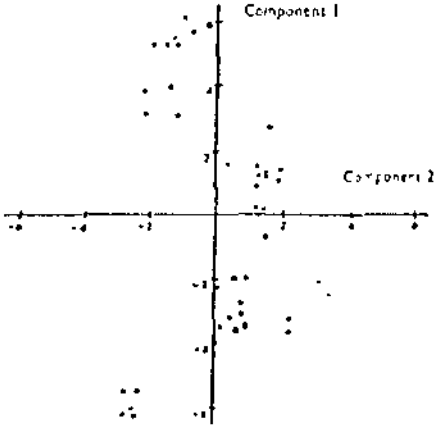
$$Y_2=0.40X_1+0.34X_2+X_3+0.84X_4-0.31X_5-0.26X_6-0.35X_7-0.35X_8+0.32X_9-0.46X_{10}$$
$$+0.38X_{11}+0.63X_{12}+0.57X_{13}$$

Em resumo, os coeficientes observados servem de contraste para avaliar o tipo de relação e de participação de cada uma das variáveis originais. Assim o primeiro CP tem altos coeficientes positivos de participação das 10 primeiras variáveis originais as quais estão relacionadas com as características físicas das vigas. O segundo CP dá uma idéia do grau de influência da sasonalidade uma vez que as variáveis X_3 e X_4 referem-se a tempo.

O emprego de ACP num segundo estudo de Jeffers(1967) resultou na comprovação que na realidade o posto da Matriz de Covariância pode ser reduzido de p = 19 para q = 4, pois os quatro primeiros CP contribuem para 92 % da variação total. Destas, ainda seria possível descartar mais duas, pois

$$\sum_{i=1}^{2} \lambda_i = 0.85 \quad / \sum_{i=1}^{p} \lambda_i$$

Gráfico n º 1.1 Plano Y_1 e Y_2 evidenciando a formação de "clusters"



O gráfico de Y_1, o primeiro CP, e Y_2, o segundo CP, mostra a formação de quatro grandes grupos de indivíduos. Com este destaque é possível ver que os "clusters" podem ser formados pelos contrastes que são apresentados nos dois primeiros CP. Um indivíduo pode então ser classificado pelas suas características em termos de participação das variáveis originais segundo os dois primeiros componentes.

CONCLUSÕES

Segundo Jeffers(1967) pode-se resumir os objetivos de ACP no que segue:

"1. O exame da correlação entre variáveis de um conjunto selecionado;
2. a redução da dimensão básica no conjunto medido para um número menor de dimensões significativas;
3. a eliminação de variáveis que contribuem com relativamente muito pouca informação extra;
4. o exame do agrupamento dos indivíduos num espaço n-dimensional;
5. determinação do peso objetivo das variáveis medidas na construção de índices significativos;
6. a alocação dos indivíduos para grupos previamente demarcados;
7. o reconhecimento de indivíduos identificados erroneamente;
8. cálculo de Regressão ortogonalizado. "

ACP no primeiro estudo teve o objetivo de contribuir na predição da força de compressão suportadas pelas vigas de madeira, enquanto que no segundo estudo a aplicação de ACP foi o elemento dominante na determinação dos "clusters" e na redução da dimensionalidade, sem o que efetivamente se tornaria difícil classificar os indivíduos através das 19 variáveis originalmente mensuradas. Jeffers põe restrições ao uso generalizado de ACP uma vez que sua utilização exige modelos lineares pois modelos não lineares exigem outro tipo de tratamento: bases logarítmicas, formas quadráticas, etc.

1.1.4 ACP em avaliação de perfil de textura alimentar

A avaliação de perfís dos alimentos tem sido realizada considerando seu aspecto qualitativo. Na verdade busca-se saber da presença ou da ausência de determinados atributos. Dois tipos de métodos contrapõem-se nesta avaliação e eles diferem não somente pela forma como a resposta é dada mas pela própria natureza do teste. Nos testes sensoriais, de natureza qualitativa, a base é a

consciensiosidade do degustador que o leva a detalhar um quadro tão preciso e completo quanto possível do alimento sob diversos aspectos previamente determinados. Este nível de consciência implica que o próprio degustador deve criar critérios não enunciados de associação do atributo ao alimento. Já no processo "quantitativo", após terem sido estabelecidas as variáveis em observação, o degustador é chamado a dar uma nota, enfim a atribuir um escore de presença ou ausência do atributo avaliado em termos de magnitude.

O método "quantitativo" pode ser, neste caso, uma extensão do qualitativo. Segundo Frutjers(1976) "Assumindo que cada rótulo do perfil representa uma sensação subjacente contínua e posteriormente, que isto continua independentemente, é possível representar que o resultado final da construção de um perfil é um sistema n-dimensional ortogonal dos eixos num espaço euclidiano. Por meio de uma métrica quantitativa é possível determinar para cada produto do tipo investigado as coordenadas neste espaço (isto é, as projeções nos eixos)."

Neste estudo, é altamente vantajoso, tanto em termos práticos como econômicos, a redução da dimensionalidade do conjunto de variáveis observadas. ACP pode ser útil como técnica multivariada com as qualidades requeridas. Segundo Frutjers(1976) "Os dados reunidos visando este perfil são analisados por Análise de Componentes Principais para encontrar qual a proporção de variáveis sensoriais que são independentes e quantas dimensões texturais subjacentes elas representam." Existe muita interrelação entre os atributos, isto é, entre as variáveis. A ACP é calculada a partir da matriz de correlação, para que estas intercorrelações possam ser avaliadas no julgamento final. Para a redução da dimensionalidade foi adotado o critério que propõe o descarte dos componentes cujos autovalores são menores do que 1, isto é, (λ_i (1).

Na matriz de correlação, observa-se que as variáveis que apresentam maior intercorrelação são rigidez x coesividade,

$\sigma_{12} = 0.85$; rigidez x elasticidade, $\sigma_{13} = 0.64$ e coesividade x elasticidade , $\sigma_{23} = 0.67$ relações diretas , positivas enquanto que a mastigabilidade relacionada com a rigidez apresentou $\sigma_{17} = - 0.76$, com a coesividade $\sigma_{27} = - 0.82$ e com a elasticidade $\sigma_{37} = - 0.79$, portanto sempre numa relação inversa.

Uma avaliação dos λ'_s pelo critério menor que 1 descartou ($p - q$) raízes características de modo que devem ser conservadas $q = 3$ autovalores e, conseqüentemente, os três primeiros CP. Além do critério já relacionando, obteve-se:

$$\frac{\sum_{i=1}^{3} \lambda_i}{\sum_{i=1}^{7} \lambda_i} \times 100 = 85.3\ \%$$

Os CP, assim, foram reduzidos a três e a interpretação dos coeficientes confere a Y_1 a identificação de " fator mecânico ", são altos os coeficientes de dureza, coesividade e de elasticidade; quanto à mastigabilidade também é alto o coeficiente porém apresenta um valor negativo logo um contraste. No segundo componente, Y_2 , são a secura e a aspereza que caracterizam o chamado fator fluidicidade. O terceiro componente, Y_3, está relacionado com a participação do ítem gordura com um c_{ij}=0.99. Este ítem ainda não tinha estado presente de modo significativo nos dois componentes anteriores, portanto este componente está totalmente dominado pela avaliação do grau de gordura da carne de peito de frango cozida.

A conclusão é de que nem todos os ítens medidos servem para discriminar os indivíduos da população no estudo do perfil de textura de carne de peito de galinha cozida. A ACP demonstrou que estudos posteriores podem basear-se numa estrutura tri-dimensional: coesividade, secura e gordura, considerando os altos coeficientes apresentados nos CP e observando o valor de intercorrelação das mesmas. A redução da dimensionalidade pode permitir maior precisão

pois o degustador terá menos fatores a observar. Pelo que foi demonstrado certos fatores decorrem de resultados anteriores confundindo a sua classificação.

1.1.5 ACP em procedimentos de controle de resíduos

A importância de ACP reside muito na sua versatilidade. Como forma de desenvolver novas aplicações e interpretar seus resultados é necessário ter o domínio do comportamento de seus parâmetros. A presença de ACP torna-se rotineira em análise de dados pelas características de praticabilidade que apresenta ao nível do trabalho computacional, suas propriedades ótimas como método para a redução da dimensionalidade, também pela sua aplicação em modelos preditivos em regressão multivariada e como instrumento do Controle Estatístico de Qualidade Multivariada. Necessariamente a redução dos CP realizada pela pressuposição de que :

$$\frac{\sum_{i=q+1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \longrightarrow 0 \quad \text{quando} \quad \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \longrightarrow 1$$

leva-nos a avaliar os resíduos num subconjunto de CP, o que resulta numa perda de parte da informação sobre a variação dos dados.

Segundo Jackson e Mudholkar(1979) " Se Componentes Principais são usados como uma técnica de redução de dados , um instrumento de diagnóstico ou um dispositivo de controle então os resíduos associados com ele são úteis para checar o ajuste e testar os "outliers" ."

Para a definição dos autovetores é necessário o desenvolvimento de :

$$C'\Sigma C = \Lambda \qquad (9)$$

onde Λ é a matriz diagonal com os autovalores distintos $\lambda_1 > \ldots > \lambda_p > 0$ e C é uma matriz ortogonal de coeficientes definida de tal modo que $C'C = I$. Os autovetores têm comprimento máximo um(1), devido a condição de ortonormalidade e passa a ser instrumentos utilizados para descartar variáveis originais. Sua maior participação nos menores CP é um indicativo de que a variável aleatória deva ser descartada.

A questão de que os autovetores podem ser diferentemente escalados provoca alguma confusão entre os autores. A seguir algumas das definições para o cálculo dos autovetores.

Seja:

$$W = C\Lambda^{-1/2} \qquad (10)$$

onde W : vetor característico utilizado como dispositivo de controle, C: autovetores associados aos autovalores de Σ, matriz de covariância Λ: matriz diagonal dos autovalores. Usando (10), Y é reescrito :

$$Y = W'X = \Lambda^{-1/2}C'X \qquad (11)$$

tal que

$$Y \sim N_p (0 ; W'\Sigma W)$$

mas

$$W'\Sigma W = \Lambda^{-1/2} C'\Sigma C \Lambda^{-1/2} \quad , \quad C'\Sigma C = \Lambda$$

$$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2}$$

$$= \Lambda^{-1} \Lambda$$

$$= I$$

então,

$$Y \sim N_p (0 ; I)$$

e neste caso:

$$T^2 = X' \Sigma^{-1} X \qquad (12)$$

reduz para:

$$T^2 = Y' Y$$

Outra normalização é

$$\boxed{V = C \Lambda^{1/2}} \qquad (13)$$

tal que

$$V' \Sigma V = \Lambda^{1/2} C' \Sigma C \Lambda^{1/2} \ , \ C' \Sigma C = \Lambda$$

$$= \Lambda^{1/2} \Lambda \Lambda^{1/2}$$

$$= \Lambda^{2}$$

então

$$V'V = \Lambda^{1/2} C'C \Lambda^{1/2}, \quad C'C = I$$

$$= \Lambda^{1/2} I \Lambda^{1/2}$$

$$= \Lambda$$

logo

$$V'\Sigma V = \Lambda^2$$

$$= \Lambda' \Lambda$$

$$= V'V\ V'V$$

$$\therefore \quad \Sigma = V\ V'$$

Com esta última normalização, os CP ficam escalados nas mesmas unidades das variáveis originais. Com a redução da dimensionalidade, a matriz de covariância dos resíduos será dada pela diferença entre a matriz de covariância original e a correspondente aos componentes remanescentes, ou seja, sendo Σ uma matriz pxp , se q<p é o posto da matriz de covariância, então esta fica reduzida a Σ_q deduzindo-se que

$$C_q'\Sigma\ C_q = \Lambda_q \qquad (14)$$

onde Λ_q é uma matriz diagonal qxq cujos valores da diagonal são os q maiores autovalores de Σ, selecionados por alguma das técnicas usuais. Conseqüentemente supõe-se que se $(p - q) \longrightarrow 0$, então $\Sigma_p - \Sigma_q \longrightarrow 0$, e esta diferença constitue-se no resíduo.

$$\Sigma_p - \Sigma_q = \Sigma - C_q \Lambda_q C_q'$$

$$= \Sigma - V_q\ V_q' \qquad (15)$$

Segundo Jackson e Mudholkar(1979) "(...) a adequacidade do modelo pode ser obtida pela predição de um valor de X através da relação :

$$\hat{X} = C_q\ \Lambda_q^{1/2}\ Y \qquad (16)$$

donde calcula-se a soma dos quadrados dos resíduos:

$$Q = (X - \hat{X})'(X - \hat{X}) \qquad (17)$$

que é uma medida geral de ajuste de uma observação com o modelo (16).

$$\text{Seja } \theta_i = \sum_{j=k+1}^{p} \lambda_j{}^i, \; i = 1,2 \text{ e } 3 \text{ e } h_0 = 1 - (2\theta_1\theta_3)/3\theta_2{}^2. \qquad (18)$$

E (...) que a quantidade:

$$c = \frac{\theta_1 \left[(Q/\theta_1)^{h_0} - 1 - \theta_2 h_0 (h_0 - 1)/\theta_1{}^2 \right]}{\sqrt{2\theta_2 h_0{}^2}} \qquad (19)$$

é aproximadamente normalmente distribuída com média zero e variância um. O limite de controle para Q vem a ser

$$Q_\alpha = \theta_1 \left[\frac{c_\alpha \sqrt{2\theta_2 h_0{}^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1{}^2} \right]^{h_0{}^{-1}} \qquad (20)$$

onde c_α é o percentil superior $(1 - \alpha)$ da distribuição normal.

As funções das últimas $(p - k)$ raízes características de Σ aparecendo nesta aproximação pode ser obtida com grande acuracidade pelas relações:

$$\theta_i = Tr\, \Sigma^{\,i} - \sum_{j=1}^{k} \lambda_j{}^i, \quad i = 1, 2, 3 \qquad (21)$$

especialmente quando as $(p - k)$ remanescentes raízes são pequenas e numerosas."

O enfoque, neste caso, é com a adequacidade do modelo empregado em termos de suas qualidade previsivas e em termos de sua capacidade de detectar a presença de "outliers". Pretende-se organizar uma técnica baseada na análise dos resíduos concluindo-se que se os resíduos tiverem presença mais significativa que a capacidade previsiva dos CP ou então do que a estatística T^2, o modelo adotado não é adequado para os dados coletados naquele momento ou então a presença de "outliers" está interferindo na atribuição de importância decisiva à variância dos dados. Existe presença forte de covariação entre as variáveis do modelo.

Os menores autovalores serão testados sobre o nível de significância de sua rejeição. A aplicação das estatísticas de teste aqui propostas baseiam-se na análise da soma dos quadrados dos resíduos em contrapartida com a performance das estatísticas calculadas a partir da ACP e outras a seguir relatadas.

A análise da adequacidade do modelo pode ser realizada pela comparação entre as estatísticas T^2 de Hotelling, CP, e Q, uma estatística dos resíduos . Como T^2 e CP analisam os dados que permanecem após o descarte de componentes, pode-se considerar que um modelo que apresente, em termos previsivos, valores altamente significativos confirmaria a adequacidade do modelo. Pois nesse caso ter-se-ia Q, soma de quadrados dos desvios entre o valor observado e o valor predito, minimamente significante. Caso as evidências sejam contrárias conclue-se que o modelo ACP não é adequado para aquele conjunto de dados.

Por outro lado, deve-se observar que se as matrizes não tem posto completo, então a presença de um simples "outlier" pode afetar de modo geral os resíduos. Para esse caso Jackson e Morris(1957) propõem:

$$\text{Res. SS} = \frac{(p - k) Q}{\theta_1} \qquad (22)$$

De tal modo que:

$$\text{Res SS} \sim \chi^2 (p - k)$$

Se o modelo adotado não contiver o mesmo número de CP que as p variáveis originais então o Res SS tende a apresentar os limites em Q confundindo os efeitos dos "outliers".

Em resumo os instrumentos de controle relatados são:

T^2 : para casos especiais de controle multivariado geral, pode ser empregada associada à ACP ou não.

Q : para processar com resíduos relacionados com ACP.

No caso de uso do estatístico Q num vetor de observações individual é necessário previamente realizar o teste do Q. Se Q não é significativo então o modelo ACP é adequado e T^2 pode então ser testado. Se T^2 não é significativo isto significa que o modelo está sob controle. Se T^2 é significativo então o CP individual pode ser testado no sentido de avaliar a causa de sua perturbação. No entanto, se Q é significativo deve-se olhar para os resíduos e Q pode ser visto como detector de "outliers".

Outras estatísticas propostas:

Q_0 e Q_L são estatísticas alternativas calculadas como estudo dos resíduos e vão ser utilizadas de forma conjunta para a obtenção de informações adicionais sobre a adequacidade do modelo empregado e sobre a descoberta da presença ou não de "outliers".

$$Q_0 = - 2 \sum_{i=1}^{n} \ln F_i, \text{ tal que } Q_0 \sim \chi^2_{2n} \qquad (23)$$

e

$$Q_L = \frac{1}{\Pi} \sqrt{\frac{3(5n+4)}{n(5n+2)}} \sum_{i=1}^{n} \ln\left(\frac{1-P_i}{P_i}\right), \qquad (24)$$

$$Q_L \sim t_{(5n+4)}, \quad \text{qualquer } n$$

outra estatística é

$$\chi^2_D = \sum_{i=1}^{n} T_i^2 - n\bar{Y}'\bar{Y}, \quad \chi^2_D \sim \chi^2_{(n-1)k} \qquad (25)$$

onde χ^2_D é uma medida da variabilidade dos valores amostrais em torno de sua própria média.

Estas três são medidas que procuram mensurar a estabilidade do processo. Se os dois primeiros são significativos, isto pode representar que existe uma falta de ajuste do modelo CP aos dados que estão sendo avaliados. Essa falta de ajuste pode ser uma condição geral do processo ou então, pode ser devida à presença de um ou mais "outliers". Se a falha se deve à variabilidade do processo em torno de sua própria média, esta havendo excessiva variabilidade nos dados. Um exame mais cuidadoso dos T^2 individuais pode detectar onde ocorrem os "outliers".

Para o caso em que nenhum destes estatísticos é significativo os autores propõem dois outros estatísticos, os quais estão intimamente interdefinidos, isto é, a definição de um deles é a condição de definição do outro: Q_M e χ^2_M.

$$Q_M = n(\bar{X} - \dot{\bar{X}})'(\bar{X} - \dot{\bar{X}}), \qquad (26)$$

$$\text{onde} \quad \dot{\bar{X}} = C_k \Lambda_k^{1/2} \bar{Y} \qquad (27)$$

que pode , se for significativo, indicar a inadequacidade geral do modelo CP. Se nenhum é significativo então

$$\chi^2_M = n\bar{Y}'\bar{Y} \qquad (28)$$

pode indicar alterações no processo que devem ser estudadas. Se χ^2_M não é significativo, através do CP médio é possível encontrar o gerador do problema.

1.1.6 ACP NA SEPARAÇÃO DE MISTURA DE NORMAIS

Desde Fleiss e Zubin(1969), Dempter(1969) e Gnanadesikan e Kettenring(1972) os critérios clássicos de descarte dos menores CP vem sendo criticados. Chang(1983) propôs a utilização de um critério assintótico para selecionar um subconjunto de CP.

Trabalhando com a distância de Mahalanobis entre as duas subpopulações, onde y é uma variável aleatória p-dimensional com mistura de duas distribuições normais tal que o vetor de médias das duas normais é $\mu' = (\mu_1, \mu_2)$, misturadas nas seguintes proporções p e (1 - p) com matriz de covariância comum Σ.

Seja:

V : matriz de covariância de Y,

tal que:

$$V = p (1 - p) d d' + \Sigma \qquad (29)$$

e

Δ : distância de Mahalanobis

tal que:

$$\Delta^2 : d' \Sigma^{-1} d \qquad (30)$$

onde:

$$d = \mu_1 - \mu_2 \qquad (31)$$

Se um subconjunto de CP de tamanho m, é utilizado, tal

que m<k onde c_i são os autovetores e λ_i são os autovalores, tenta-se estabelecer a relação entre a distância de Mahalanobis desta população e os autovalores através da proposição 1 à seguir:

Conforme estabelece Chang(1983), " (...) a relação Δ_m e os autovalores é estabelecida. Nós a estabelecemos em termos de Δ_m^2

Proposição 1

$$\Delta_m^2 = \frac{\sum_{i=1}^{m} \frac{(c_i'd)^2}{\lambda_i}}{\left[1 - p(1-p) \sum_{i=1}^{m} \frac{(c_i'd)^2}{\lambda_i} \right]} \qquad (32)$$

Em particular, para m=1,

$$\Delta_1^2 = \frac{\frac{(c_i'd)^2}{\lambda_i}}{\left[1 - p(1-p) \frac{(c_i'd)^2}{\lambda_i} \right]} \qquad (33)$$

Com isso estabelece-se que:

i - O CP com a maior quantidade de informação nem sempre é aquele que tem o maior λ_i. Como se pode perceber na proposição 1 a distância calculada é uma função monótona da razão entre o valor da componente e do autovalor e não simplesmente do autovalor.

ii- Através dessa proposição pode-se concluir que o subconjunto que apresenta a maior distância é o melhor pois tem maior poder de discriminação no caso de formação de "clusters". Mesmo com o emprego

de estimadores tanto para os autovalores como para os autovetores em função de se estar tratando com valores amostrais, a proposição número 1 é válida.

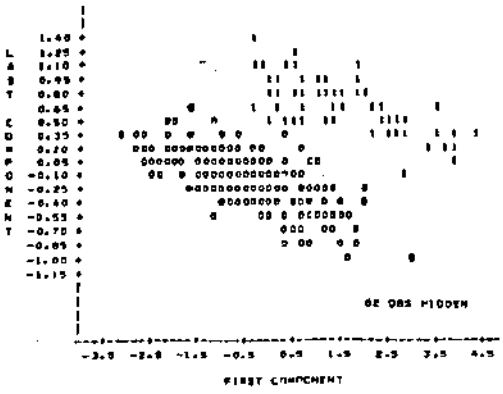
O descarte de componentes justifica-se apenas, quando entre os remanescentes pode-se garantir a maior quantidade de informações sobre os dados originais. A proposição 2 traz uma série de encaminhamentos para a avaliar o comportamento dos dados rotacionados e reduzidos na sua dimensionalidade em relação às distâncias máximas. Segundo Chang(1983), a segunda proposição pode ser colocada nos seguintes termos:

Proposição 2

"A informação está sendo distribuída em m ou menos que m Componentes Principais se Σ, a matriz de covariância populacional interna comum das duas sub-populações, tem m autovalores distintos."

Esta proposição pode ser estendida para o caso de mistura de mais de duas normais concluindo-se que a informação é inteiramente contida em $m(k-1)$ ou menos CP se Σ tem m raízes distintas. Os autores comprovam ainda que as informações podem estar contidas em outros CP que não os primeiros através de um exemplo, utilizando a matriz de correlação. Através de estudos simulados e de representação gráfica dos resultados é possivel existir um maior poder discriminatório entre os dados provenientes das duas normais misturadas quando se observa o gráfico de Y_m, o último CP contra Y_1, o primeiro CP. Este poder discriminatório que está sendo utilizado para separar as duas sub-populações normais misturadas perde grandemente seu poder quando se observar Y_1, o primeiro CP contra Y_2, o segundo maior CP. Seja:

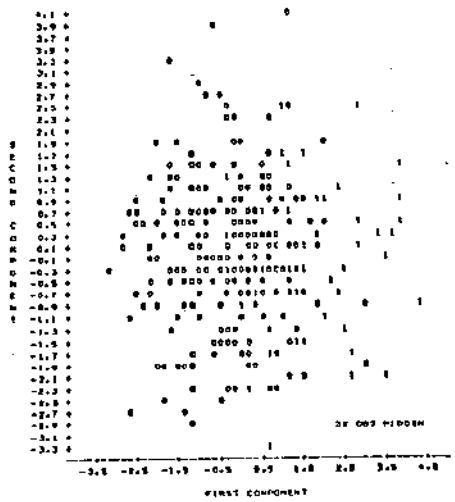
Gráfico nº 1.2 Dispersão de pontos de Y_1 versus Y_m



Este gráfico deixa muito claro a existência de dois grupos de pontos onde muito poucos pontos estão situados numa faixa intermediária na qual se poderia admitir dúvida quanto à sua classificação. Entretanto os "clusters" formados são muito evidentes. Este poder de discriminação deve-se ao fato de que a distância entre os pontos é mais evidenciada quando se avalia os pares de valores quanto ao seu desempenho no primeiro e no último componente.

Compara-se este resultado com o observado no gráfico de Y_1 com o Y_2. A diferenciação pretendida se confunde uma vez que as distâncias são muito semelhantes dificultando a visualização Neste caso, é possível questionar-se sobre a absolutização da escolha dos CP definida desde Pearson e Hotelling. É necessário, então, investigar não só a adequacidade do modelo mas também as distâncias entre as observações avaliadas como distâncias de Mahalanobis. A maior distância, provavelmente, estará entre Y_1 e Y_m.

Gráfico nº 1.3 Dispersão entre Y_1 e Y_2



Esta comparação permite demonstrar que a prática usual na escolha dos CP nem sempre é a mais correta, a mais indicada no caso de se estar pretendendo separar sub-conjuntos de dados pela formação de "clusters". Como cada "cluster" se forma pela variabilidade entre os pontos é necessário buscar o maior poder de discriminação. Este poder sempre foi atribuído à alta concentração da variância. A formação de "clusters" pressupõe

homogeneidade no interior do grupo, mas heterogeneidade entre os grupos. Isto poderia estar explicando a seleção da menor faixa de variação como sendo a que possibilita uma melhor associação de pontos

Segundo Chang(1983) " (...) o sucesso da Análise de Componentes Principais depende de vários fatores. Deve haver uma vantagem distinta sobre os dados não transformados na qual os melhores m componentes individuais constituem o melhor subconjunto de componentes, pelo menos assintóticamente. O teste e o gráfico podem ser utilizados para ajudar a selecionar aqueles componentes que contém as maiores informações. O sucesso disto é criticamente dependente do tamanho da amostra e da distribuição de informações."

O uso de CP pode trazer grandes vantagens em estudos de populações compostas por dois ou mais estratos, quando se desconhecem critérios de classificação dos indivíduos.

1.2 SIMULAÇÃO

Para realizar os estudos que serão apresentados a seguir foram geradas amostras normais multivariadas com n=20; n=50 e n=200 que serão utilizadas como unidades experimentais de tal forma que a geração inicial será alterada de acordo com os objetivos que serão enfocados. Estes dados foram gerados numa distribuição normal p-variada com p=5, vetor de médias μ e matriz de covariância, Σ,

$$\mu_{1xp} = (\ 11.7166 \quad 25.7411 \quad 9.58361 \quad 46.2059 \quad 33.2548\ ) \qquad (34)$$

$$\Sigma_{pxp} = \begin{bmatrix} 4.09579 & -0.400363 & -0.482933 & -0.567155 & 0.98261 \\ & 42.90602 & -1.45298 & -13.3242 & -1.13109 \\ & & 12.3814 & -2.8519 & -3.63819 \\ & & & 166.653 & 5.87478 \\ & & & & 49.2751 \end{bmatrix} \qquad (35)$$

A matriz de covariância, (35) , apresenta algumas condições de interesse neste estudo. O espaço de variação das

variáveis é diferente sem se constituir em excessões, os coeficientes de variação variam de 17% a 37%, garantindo homogeneidade nos dados. Por outro lado é possível observar que muitas das variáveis são praticamente independentes com coeficientes de correlação que varia entre -0.07 e 0.07, mas r_{24}= -0.16 e r_{35}=-0.15

Após a geração de dados, originais e com as perturbações do tipo descrito, serão calculados os CP e avaliadas as mudanças verificadas nos coeficientes dos mesmos. Esta comparação permitirá que se conclua a respeito da sensibilidade dos CP.

Após a aplicação de ACP nos dados originais obteve-se os seguintes resultados:

Tabela nº 1.2 Resultados da aplicação de ACP nos dados com estrutura (34) e (35), para n=20, n=50 e n=200

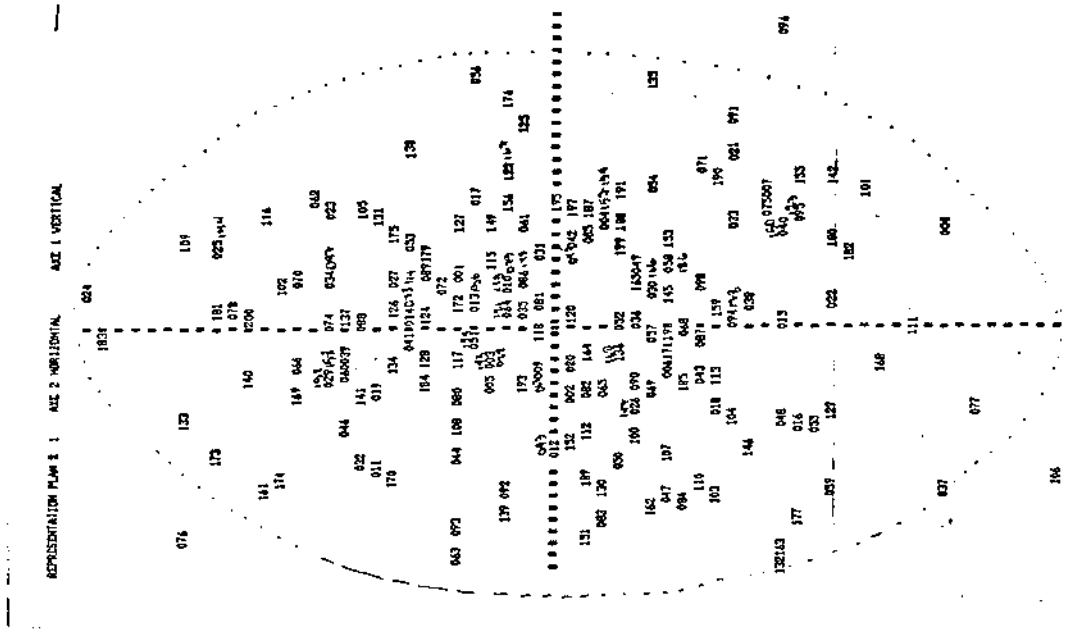
n=20

Componentes	%	c_{1i}	c_{2i}	c_{3i}	c_{4i}	c_{5i}
n=20						
X_1	59.64	-0.04326	0.02127	-0.09122	-0.28930	0.95165
X_2	19.11	0.14013	-0.12817	0.97538	0.02284	0.10968
X_3	18.67	0.04010	-0.24972	-0.09152	0.92188	0.27888
X_4	2.19	0.97793	-0.11007	-0.15345	-0.08914	0.00511
X_5	0.39	0.14324	0.95322	0.09149	0.24074	0.06716
n=50						
X_1	64.67	-0.02634	-0.00475	0.02332	-0.04286	0.99840
X_2	18.95	-0.08799	-0.32498	0.93975	0.05490	-0.02207
X_3	12.10	0.04021	-0.07532	-0.07917	0.99215	0.04558
X_4	2.94	0.99299	0.03276	0.10665	-0.03037	0.02439
X_5	1.34	-0.06167	0.94214	0.31412	0.09931	-0.00486
n=200						
X_1	61.16	-0.00281	0.02359	-0.01028	-0.05348	0.99823
X_2	17.92	-0.10568	-0.03895	0.99186	0.05757	0.01392
X_3	15.12	-0.01833	-0.09217	-0.06392	0.99207	0.05462
X_4	4.32	0.99294	-0.05628	0.10237	0.019373	0.00622
X_5	1.46	0.05049	0.99310	0.03902	0.09670	-0.01775

Os resultados da aplicação dos CP demonstram que o número de componentes pode ser reduzido para 3 pois o % de variância explicada pelos três primeiros CP é, respectivamente, 97.4%, 95.7% e

94.2%. Uma observação dos dois menores CP evidencia que as duas variáveis que mais contribuem na variância destes são respectivamente X_3 e X_1. É essencial observar que os coeficientes concentrando quase toda a informação em cada CP são os mesmos, ainda que com pequenas variações.

Gráfico nº 1.4 Plano Y_1 e Y_2 com a aplicação de ACP nos dados originais.



CAPÍTULO 2

ESTABILIDADE DOS CP NO USO DE ARREDONDAMENTOS

Os estudos que existem, em geral avaliam o efeito nos autovalores para casos em que são realizadas transformações nos autovetores do tipo arredondamento dos coeficientes c_i e como muitas vezes ocorre quando os $c_i \longrightarrow 0$, serão igualados a zero, isto é,

literalmente zerados. Os estudos realizados por Bibby(1980) e Green(1977) estabeleceram o tipo de sensibilidade que os autovalores apresentam ao serem efetivadas estas alterações nos autovetores.

Bibby(1980) trata o "erro de arredondamento" como um "mal necessário". Portanto, passa a ser um ponto importante, o conhecimento do comportamento dos resultados após o arredondamento dos coeficientes. Ele justifica o uso de arredondamentos pelo fato de que os resultados da aplicação das técnicas deverão ser interpretados e os coeficientes, c_i, arredondados

> i) facilitam a interpretação;
> ii) facilitam os cálculos dos autovalores;
> iii)" são apenas marginalmente inferiores aos estimadores "ótimos" mesmo usando seu próprio critério de otimalidade".

Para realizar seu artigo, Bibby(1980) enfocou exemplos sobre Análise de Regressão, ACP, sendo que , neste último, utilizou um exemplo de Fisher apresentado por Anderson(1958). A seguir apresentar-se-á os resultados do primeiro CP, neste estudo de Fisher, quanto ao comportamento do % de variância total abrangida pelo mesmo, além de demonstrar o tipo de arredondamento realizado, para efeito de comparação:

Tabela nº 2.1 Variações no % de variância explicada após as perturbações por arredondamento nos autovetores

estágio dos c_i	% da variância	a_1	a_2	a_3	a_4
original	78	0,6867	0,3053	0,6237	0,2150
aproximação	77	0.7	0,3	0.6	0,2
razão inteiro	76	2.0	1.0	2.0	1.0

Na tabela 2.1 é possível verificar uma ligeira perda da variação total contida no primeiro CP. As perturbações apresentadas na variância dos componentes, faz decrescer o % de variação explicada em apenas 1% e 2% respectivamente. O uso de aproximações nos

coeficientes, c_i, produziram pequenas perdas na variância.

O teorema, a seguir apresentado, traz os limites para a redução do percentual de variância explicada ao serem utilizados arredondamentos em ACP. Este teorema é apresentado por Ribby(1980) denotando por Δ_k a perturbação da variância associada a um vetor arbitrário h, seu desvio em relação a λ_k, a variância associada ao k-ésimo autovetor:

TEOREMA 1

Seja S uma matriz real simétrica pxp com autovalores $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ e correspondentes autovetores ortonormalizados γ_k, k=1, ... ,p. Seja h um vetor e seja

$$\Delta_k = \frac{h' S\, h}{h' h} - \lambda_k$$

Então,

$$r^2(\lambda_p - \lambda_k) \leq \Delta_k \leq r^2(\lambda_1 - \lambda_k) \qquad (36)$$

onde $r^2 = \| h - \gamma_k \|$

Prova

Seja $\xi = (h - \gamma_k)$, tal que $h = (\gamma_k + \xi)$ e $r = \| \xi \|$

então,

$$h'Sh = (\gamma_k + \xi)' S (\gamma_k + \xi)$$

$$= \gamma_k' S \gamma_k + 2 \gamma_k' S \xi + \xi' S \xi$$

$$= \gamma_k' S \gamma_k \gamma_k' \gamma_k + 2 \gamma_k' S \gamma_k' \gamma_k \xi + \xi' S \xi \gamma_k' \gamma_k$$

$$= \lambda_k + \lambda_k \gamma_k \xi + \xi' S \xi \qquad (37)$$

resultado este dado pelas próprias propriedades dos autovalores e dos autovetores.

$$S\gamma_k = \lambda_k \gamma_k \qquad e \qquad \gamma_k' \gamma_k = 1$$

Agora ξ tem uma única representação canônica. Pode-se escrever $\xi = \Sigma_i q_i \gamma_i$. Deste modo é possível substituir ξ em (36) escrevendo-o em termos de coeficientes q_k como $\xi = q_k \gamma_k$

Então

$$\gamma_k' \xi = q_k \gamma_k' \gamma_k$$

$$\gamma k' \xi = q_k$$

ainda

$$\xi' S \xi = \Sigma_i \, q_i \gamma_i' S \, \gamma \, q_i$$

$$= \Sigma_i \, \gamma_i' S \, \gamma \, q_i^2$$

$$= \Sigma_i \, \lambda_i q_i^2$$

substituindo então em (36), obtém-se

$$h' S \, h = \lambda_k + 2 \, \lambda_k \, \gamma_k \, \xi + \xi' S \, \xi$$

$$= \lambda_k + 2 \, \lambda_k \, q_k + \Sigma_i \, \lambda_i \, q_i^2$$

quanto a

$$h' h = (\gamma_k + \xi)'(\gamma_k + \xi)$$

$$= \gamma_k' \gamma_k + 2 \, \gamma_k' \xi + \xi' \xi$$

$$= 1 + q_k + r^2$$

logo

$$\Delta_k = \frac{h' S \, h}{h' h} - \lambda_k$$

$$= \frac{\lambda_k + 2\lambda_k \, q_k + \Sigma_i \lambda_i q_i^2 - \lambda k \, (\, 1 + 2 q_k + r^2)}{1 + 2 \, qk + r^2}$$

$$= \frac{\lambda_k + 2\lambda_k q_k + \Sigma_i \lambda_i q_i^2 - \lambda_k - 2\lambda_k q_k - \lambda_k r^2}{1 \quad 2 q_k + r^2}$$

$$= \frac{\Sigma_i \lambda_i q_i^2 - \lambda_k r^2}{1 + 2q_k + r^2} \quad , \; r = ||\xi||$$

$$= \frac{\Sigma_i \lambda_i q_i^2 - \lambda_k \xi'\xi}{1 + 2 q_k + r^2} \; , \qquad \xi'\xi = \Sigma_i q_i \gamma_i' \gamma_i q_i$$
$$= \Sigma_i q_i^2 \; , \; \gamma i' \gamma i = 1$$

$$= \frac{\Sigma_i \lambda_i q_i^2 - \lambda_k \Sigma_i q_i^2}{1 + 2 q_k + r^2}$$

$$= \frac{\Sigma_i (\lambda_i - \lambda_k) q_i^2}{1 + 2q_k + r^2} \qquad (38)$$

Então os limites para Δ_k podem ser encontrados. Para tanto, assumindo que k=1, se for assumido também um q_k, e r fixos obter-se-á um denominador $(1 + 2 q_k + r^2)$ também fixo. Nesse caso quem está variando são os autovalores λ_i, i = 1, ... , p. Para maximizar $\Sigma_i(\lambda_i - \lambda_k)q_i^2$ requer-se q_1 tão grande quanto possível, uma vez que q_1 está multiplicando o maior dos termos $(\lambda_i - \lambda_k)$. Isto é, $q_1^2 = r^2 - q_k^2$ e $q_j = 0$, $(j \neq 1,k)$. Conseqüentemente, o máximo que Δ_k pode assumir, para um dado valor de q_k e de r^2, é

$$\frac{(r^2 - q_k^2)(\lambda_1 - \lambda_k)}{1 + 2q_k + r^2} \qquad (39)$$

seja

$$f(q_k) = (r^2 - q_k^2)/(1 + 2q_k + r^2)$$

então

$$\Delta k \leq f(q_k)(\lambda_1 - \lambda_k)$$

desenvolvendo $f(q_k)$, obtém-se

$$f(q) = \frac{(r^2 - q^2)}{1 + 2q + r^2}$$

$$= \frac{r^2 - q^2}{1 + 2q + r^2 + q^2 - q^2}$$

$$= \frac{r^2 - q^2}{(1 + q)^2 + (r^2 - q^2)}$$

então

$$f(q) = r^2 - \frac{(r^2 + q^2)}{(1 + q)^2 + (r^2 - q^2)} \qquad (40)$$

Uma observação mais cuidadosa leva, então, a concluir que $f(q)$ é limitada pois para $q = -r^2$, obtém-se o limite máximo de (40) em $f(-r^2) = r^2$. Portanto

$$\Delta_k \leq r^2(\lambda_1 - \lambda_k) \quad , k \neq 1 \qquad (41)$$

$$\Delta_1 \leq 0 \quad , k = 1$$

Similarmente, prova-se o limite inferior de Δ_k.

2.1 A QUESTÃO DO ARREDONDAMENTO

Pode-se dizer que os arredondamentos são erros sobre os quais não se têm uma idéia precisa de qual foi o montante de afastamentos realizados. Não se pode prever com certeza quanto vai ser descontado de cada valor original. O que se pode avaliar é apenas o domínio desses valores; entre $(-c; c)$, dependendo do tipo de aproximação efetuada. Bibby(1980), baseado em estudos de Fisher(1944), Eisenhart(1947), procurou estabelecer as propriedades estocásticas dessas aproximações. Assim, seja

$$\delta = X_i - X_i^*, \qquad \text{onde } X_i^*: \text{ valor aproximado}$$

então os δ_i são as diferenças com o valor verdadeiro e, portanto, uma

variável aleatória contínua que pode tomar valores num intervalo [-c;c], de tal forma que a probabilidade da variável assumir valores num sub-intervalo é a mesma para qualquer outro sub-intervalo do mesmo comprimento.

Podemos assumir que

i) os δ_i são independentes entre si;
ii) os δ_i se distribuem uniformemente no intervalo entre (-c;c);
iii) os δ_i são independentes dos valores verdadeiros, não arredondados

Então, se δ_1, ... δ_n, onde n é o número de valores observados, são as diferenças entre os valores observados e os valores arredondados, tem-se que

$$a)\quad E(\delta_i) = \frac{(-c + c)}{2} = 0$$

$$b)\quad E(\delta_i^2) = V(\delta_i) = \frac{[c - (-c)]^2}{12} = \frac{c^2}{3}$$

$$c)\quad Var(\delta_i^2) = \frac{4\,c^4}{45}$$

Nesse caso, se for considerada a influência dos arredondamentos na função critério, isto é, na variância, deve-se avaliar o comportamento da soma dos quadrados dos arredondamentos.

Seja

$$r^2 = \sum_{i=1}^{p} \delta_i^2 , \qquad (42)$$

então

$$E(r^2) = 1/3\ pc^2 \quad e \quad Var(r^2) = 4/45\ pc^4$$

No caso de ACP o efeito do arredondamento dos valores originais pode ser assumido como seguindo a distribuição conforme foi descrita acima.

Se for atribuído a r um valor fixo este estudo passa a ter um caráter um tanto determinístico, no entanto valor do arredondamento em si é aleatório e com uma estrutura probabilística bem definida. Os valores observados não precisam apresentar uma estrutura de dados tão bem definida. Nesse caso Δ_k, definido anteriormente também é aleatório e pode-se supor que

a) $\delta_i \sim U(-0.05; 0.05)$;
b) δ_i, δ_j são independentes, identicamente distribuídos;
c) δ_i, x_i são independentes.

Suponha que p=5 e a expressão Δ_k pode ser reescrita:

$$r^2(\lambda_p - \lambda_k) \leq \Delta_k \leq r^2(\lambda_1 - \lambda_k)$$

$$E(r^2)(\lambda_p - \lambda_k) \leq E(\Delta_k) \leq E(r^2)(\lambda_1 - \lambda_k)$$

$$1/3\ pc^2(\lambda_p - \lambda_k) \leq E(\Delta_k) \leq 1/3\ pc^2(\lambda_1 - \lambda_k)$$

$$1/3(5)(0.05)^2(\lambda_p - \lambda_k) \leq E(\Delta_k) \leq 1/3(5)(0.05)^2(\lambda_1 - \lambda_k)$$

$$1/240\ (\lambda_p - \lambda_k) \leq E(\Delta_k) \leq 1/240\ (\lambda_1 - \lambda_k)$$

Desse modo quando se tem k=1, obtém-se

$$1/240\ (\lambda_p - \lambda_1) \leq E(\Delta_k) \leq 1/240\ (\lambda_1 - \lambda_1)$$

$$1/240\ (\lambda p - \lambda 1) \leq E(\Delta_k) \leq 0$$

O arredondamento dos dados originais para uma casa decimal, num vetor aleatório de quinta dimensão apresenta o % de variância explicada reduzido em menos do que $1/240(\lambda_p - \lambda_1)$. Assim o arredondamento, tendo por base a expressão acima definida pode ser esperado por crescer ou diminuir a variância explicada pelos primeiros CP. Pode-se, então, concluir que o preço pago pela sub-otimalidade está representado pelo montante de mudança ocorrida no % de variação explicada, isto é, na variância e, conseqüentemente nos coeficientes c_i.

Portanto avaliar as mudanças ocorridas nos autovetores como efeito dos arredondamentos consiste em conhecer o intervalo de variação do % de variância explicada, dada por Δ_k. No entanto, esta técnica não deixa explícito qual o autovetor mais perturbado, remetendo à necessidade de definir outras formas de detectar a falta de estabilidade individual.

2.2 RESULTADOS OBTIDOS COM OS ARREDONDAMENTOS DOS X_i

Utilizando um conjunto de dados gerados a partir de (34) e (35) obtém-se aplicando CP, os seguintes resultados, diferenciando as amostras geradas para n=20, n=50 e n=200. Foram realizados três tipos de alterações:

1º arredondamento nos dados com aproximação em décimos;

2º arredondamento nos dados aproximação para o inteiro mais próximo.

3º arredondamento na matriz de covariância com aproximação para o inteiro mais próximo

Como se observa na tabela 2.2, abaixo, o percentual de variação explicada, após a aplicação dos três tipos de arredondamentos citados acima, apresentou pequenas alterações porém muito pouco significativas. Esta proximidade cresce a medida em que n tende a infinito aproximando-se do valor original. O % de variância explicada apresentou mais alterações quando os dados foram aproximados para o inteiro mais próximo porém as alterações são semelhantes a pequenos ruídos, não significativos.

Tabela nº 2.2 Alterações no % de variância explicada em função dos arredondamentos dos x_i's(1º e 2º tipo de arredondamento) e na matriz de covariância (3º tipo de arredondamento).

Componentes		Y_1	Y_2	Y_3	Y_4	Y_5
n = 20		59.64	19.10	18.67	2.18	0.39
	1º	59.66	19.10	18.66	2.18	0.38
	2º	59.36	19.18	18.83	2.14	0.48
	3º	59.64	19.11	18.65	2.19	0.39
n = 50		64.70	18.95	12.09	3.02	1.21
	1º	64.72	18.94	12.07	3.03	1.22
	2º	64.79	18.84	12.29	2.90	1.16
	3º	64.69	18.95	11.98	3.09	1.26
n = 200		61.16	17.92	15.12	4.32	1.46
	1º	61.13	17.93	15.12	4.32	1.48
	2º	61.04	17.98	15.18	4.24	1.55
	3º	61.35	17.86	15.16	4.21	1.40

Com base no Teorema 1, calculou-se Δ_{p_i} para p=5 com a finalidade de detectar os limites de possível variação do % de variância explicada. Através deste intervalo, a maior falta de precisão está sendo detectada para os casos em que n = 20 e n = 50. Os intervalos demonstram a possível existência de desestabilizações com valor máximo de 0.87 e 0.83 respectivamente. Como se pode observar na Tabela 2.3 muitas questões ficam sem respostas neste caso. Não é possível verificar, por exemplo, quais são as variáveis mais perturbadas.

Tabela nº 2.3 Intervalos de Δ_k - limites de variação do percentual de variância explicada segundo o CP e segundo o tamanho de n.

k	n = 20	n = 50	n = 200
1	-0.87 ; 0	-0.83 ; 0	-0.68 ; 0
2	-0.28 ; 0.60	-0.23 ; 0.59	-0.19 ; 0.50
3	-0.27 ; 0.61	-0.14 ; 0.68	-0.16 ; 0.53
4	-0.03 ; 0.85	-0.02 ; 0.80	-0.03 ; 0.65
5	0 ; 0.87	0 ; 0.83	0 ; 0.68

Observe-se, através das Tabelas 2.4 à 2.8, os reflexos nos coeficientes, c_j, do uso de arredondamentos nos dados originais ou então na matriz de covariância. Porém não é só a praticabilidade do procedimento que torna interessante seu estudo mas também o fato de que muitos tratamentos estatísticos desenvolvidos com o apoio da informática já trazem embutidos em seu cálculo arredondamentos do tipo descrito.

Tabela nº 2.4 Variações do coeficiente C_1: 1º autovetor associado à λ_1, 1º autovalor, com aplicação de arredondamentos.

C_1		c_{11}	c_{12}	c_{13}	c_{14}	c_{15}
n = 20		-0.04326	0.14013	0.04010	0.97793	0.14324
	1º	-0.04379	0.14150	-0.03935	0.97746	0.14151
	2º	-0.05020	0.14425	0.03861	0.97690	0.14432
	3º	-0.04347	0.14071	0.03828	0.97817	0.14151
n = 50		-0.04968	-0.08851	0.04001	0.99189	-0.06514
	1º	-0.04974	-0.08881	0.04038	0.99186	-0.06495
	2º	-0.05166	-0.08701	0.04261	0.99191	-0.06362
	3º	-0.05123	-0.09082	0.03938	0.99157	-0.06605
n = 200		-0.00281	-0.10568	-0.01833	0.99294	0.05049
	1º	-0.00270	-0.10662	-0.01810	0.99287	0.04990
	2º	-0.00224	-0.10851	-0.01881	0.99245	0.05391
	3º	-0.00559	-0.10297	-0.01962	0.99315	0.05124

O 1º CP se apresenta estável independente do tamanho da amostra.

Tabela nº 2.5 Variações do coeficiente C_2: 2º autovetor associado à

Tabela nº 2.5 Variações do coeficiente C_2: 2º autovetor associado à λ_2, 2º autovalor, com o uso de aproximações.

C_2		c_{21}	c_{22}	c_{23}	c_{24}	c_{25}
n = 20		0.02127	-0.12822	-0.24971	-0.11006	0.95322
	1º	0.02780	-0.19356	-0.24348	-0.10123	0.94457
	2º	0.03407	-0.18762	-0.24519	-0.10049	0.94520
	3º	0.01623	-0.08160	-0.25712	-0.11575	0.95580
n = 50		-0.00826	-0.32795	-0.07463	0.03513	0.94104
	1º	-0.00800	-0.32508	-0.07534	0.03524	0.94198
	2º	0.00151	-0.35085	-0.06449	0.03195	0.93365
	3º	-0.00971	-0.33139	-0.07675	0.03479	0.93967
n = 200		0.02359	-0.03895	-0.09217	-0.05628	0.99310
	1º	0.02361	-0.03898	-0.09213	-0.05638	0.99311
	2º	0.02126	-0.04436	-0.09366	-0.06050	0.99254
	3º	0.02549	-0.02675	-0.10238	-0.05586	0.99248

A desestabilização pode ser verificada no caso de pequenas amostras, onde, por exemplo, c_{22} aumenta de importância quando n=20.

Tabela nº 2.6 Variações no coeficiente C_3: 3º autovetor associado à

Tabela nº 2.6 Variações no coeficiente C_3: 3º autovetor associado à λ_3, 3º autovalor com o uso de arredondamentos.

C_3		c_{31}	c_{32}	c_{33}	c_{34}	c_{35}
n = 20		-0.09122	0.97538	-0.09153	-0.15346	0.09154
	1º	-0.08929	0.96431	-0.10884	-0.16221	0.15479
	2º	-0.08597	0.96479	-0.11428	-0.16416	0.14751
	3º	-0.08872	0.98096	-0.07318	-0.14906	0.04752
n = 50		-0.00654	0.93898	-0.07934	0.10747	0.31687
	1º	-0.00715	0.93998	-0.07872	0.10758	0.31401
	2º	-0.01557	0.92953	-0.09724	0.10664	0.33895
	3º	0.000731	0.93641	-0.09814	0.11089	0.31812
n = 200		-0.01028	0.99186	-0.06392	0.10237	0.03902
	1º	-0.01097	0.99172	-0.06424	0.10330	0.03959
	2º	-0.00831	0.99083	-0.07281	0.10455	0.04404
	3º	-0.00068	0.99341	-0.04729	0.10061	0.02757

O 3º autovetor apresenta alterações, principalmente, quando se trata de aproximações nos dados originais.

Tabela nº 2.7 Variações do coeficiente C_4: 4º autovetor associado à

Tabela nº 2.7 Variações do coeficiente C_4: 4º autovetor associado à λ_4, 4º autovalor com o uso de aproximações.

C_4		c_{41}	c_{42}	c_{43}	c_{44}	c_{45}
n = 20		-0.28930	0.02283	0.92188	-0.08914	0.24074
	1º	-0.28794	0.02363	0.92209	-0.08931	0.24143
	2º	-0.25917	0.03354	0.92857	-0.09148	0.24715
	3º	-0.30137	0.01626	0.91781	-0.08679	0.24289
n = 50		-0.21346	0.05288	0.97010	-0.03887	0.09494
	1º	-0.21376	0.05227	0.97002	-0.03929	0.09528
	2º	-0.19061	0.06869	0.97385	-0.03966	0.09475
	3º	-0.30361	0.06992	0.94408	-0.04011	0.10011
n = 200		-0.05348	0.05757	0.99201	0.01937	0.09670
	1º	-0.05499	0.05775	0.99186	0.01925	0.09733
	2º	-0.05451	0.06583	0.99120	0.02048	0.09890
	3º	-0.11817	0.04211	0.98616	0.01767	0.10689

Tabela nº 2.8 Variações do coeficiente C_5: 5º autovetor associado à

Tabela nº 2.8 Variações do coeficiente C_5: 5º autovetor associado à λ_5, 5º autovalor no uso de aproximações.

C_5		c_{51}	c_{52}	c_{53}	c_{54}	c_{55}
n = 20		0.95165	0.10968	0.27888	0.00511	0.06716
	1º	0.95206	0.10975	0.27759	0.00056	0.06662
	2º	0.96008	0.10965	0.25115	0.01525	0.05393
	3º	0.94823	0.10480	0.29101	0.00529	0.07177
n = 50		0.97563	0.01058	0.21312	0.04302	0.02755
	1º	0.97555	0.01115	0.21341	0.04304	0.02760
	2º	0.98018	0.02408	0.19018	0.04621	0.01901
	3º	0.95136	0.01332	0.30270	0.04087	0.03773
n = 200		0.99823	0.01392	0.05462	0.00622	-0.01775
	1º	0.99813	0.01474	0.05611	0.00622	-0.01786
	2º	0.99824	0.01255	0.05547	0.00550	-0.01525
	3º	0.99264	0.00580	0.11989	0.00920	-0.01246

Como se pode observar os coeficientes do primeiro CP não demonstram sensibilidade tanto no uso de arredondamentos nas observações registradas como no caso de aproximações das variâncias e covariâncias, independente da utilização de pequenas ou de grandes amostras. Inclusive, é possível detectar, à medida em que n cresce, uma maior aproximação com o valor original. Quando o arredondamento determina perdas ou acréscimos nos dados observados que variam no intervalo (- 0.5; + 0.5), o 2º e o 4º CP apresentam pequenas perturbações. estas alterações podem também ser relacionadas com o tamanho da amostra pois se evidenciam quando se trata de pequenas amostras. É possível concluir que a estabilidade neste caso cresce quando $n \longrightarrow \infty$. As alterações observadas não são significativas pois não interferem nas interpretações dos coeficientes.

As perturbações mais significativas ocorrem quando a matriz de covariância é aproximada para o número inteiro mais próximos, significando perdas de no máximo ± 0.5, denominada de 3º tipo de arredondamento. Com excessão do 1º CP que permaneceu estável, todos os outros apresentaram algumas alterações que, em alguns casos, modificam a interpretação dos coeficientes, especialmente quando se trata de pequenas amostras. Observando-se a Tabela 2.5 verifica-se que o 2º CP tem o coeficiente c_{22} alterado; na Tabela 2.6 do 3º CP os coeficientes mais perturbados são c_{33} e c_{35}. Através da Tabela 2.7 é possível concluir que o 4º CP é sensível à todos os tipos de arredondamentos independente do tamanho da amostra, sendo alguns mais significativos como é o caso do coeficiente c_{41}. A análise da Tabela 2.8 leva a concluir que o último CP é bastante sensível à introdução de arredondamento nos dados originais ou na matriz de covariância qualquer que seja o tamanho da amostra. Neste caso observa-se em especial o comportamento dos coeficientes c_{51}, quando n=20 e n=50, e do coeficiente c_{53}, qualquer que seja n. Esta sensibilidade reforça a idéia de sua utilização para detectar a ocorrência de perturbações.

Desta forma, conclue-se que as aproximações realizadas nos valores observados originalmente não perturbam os coeficientes, c_i, quando $n \longrightarrow \infty$. Por outro lado, em caso de pequenas amostras os

coeficientes que sofrem alterações são os dos CP intermediários.

Os limites propostos por Bibby para Δ_k não apresentam grandes alterações no caso de n = 200, que também apresentou alterações. Constata-se também que não necessáriamente os % de variância explicada refletem o comportamento dos autovetores ainda que estes sejam definidos em função dos autovalores principalmente quando se trata, como neste caso, de pequenas alterações. O uso das aproximações altera estrutura de variância modificando as medidas de escala porém as proporções não refletem estas modificações. Os autovetores, no entanto, são associados aos autovalores portanto não permanecem estáveis. Sob estas condições, as limitações no uso da forma proposta por Bibby, é necessário encontrar outras maneiras de mensurar o desajuste dos coeficientes. Estas alternativas podem ser:

* Aplicar o mesmo método que será aplicado quando se verificar a ocorrência de perdas de otimalidade na variância;

* Comparar graficamente os intervalos de δ_k observados.

As conclusões de quase estabilidade está necessariamente associada a normalidade dos dados. Outros estudos são necessários para que se possa concluir quando esta normalidade não possa ser garantida.

CAPÍTULO 3

ESTABILIDADE DOS CP NA PRESENÇA DE "OUTLIERS"

No caso multivariado pode-se ter diferentes tipos de "outliers", valores extremos altamente improváveis. Uma observação pode ser considerada como um "outlier" por apresentar uma ou mais das p-variáveis com valor muito diferente do valor esperado. A influência

que pode apresentar um "outlier" sobre as medidas que estão sendo calculadas torna complicada a sua presença no conjunto de dados. Em dados multivariados, a presença de um "outlier" distorce as medidas de locação, as medidas de escala e, também, as medidas de orientação (correlação). Pode-se dizer, nesse caso, que a presença de "outlier" incha inapropriadamente a variância e a covariância ou as correlações. Todas estas medidas bem como as medidas de locação são de grande importância na ACP e por isso toda uma preocupação em detectar quanto os CP são sensíveis à presença dos "outliers". Aliado à Análise de Sensibilidade se pode realizar um processo de identificação da observação que se caracteriza como um "outlier". Um estudo de Hawkins(1974) trata dos sérios problemas resultantes da presença de erros em dados multivariados, verificando, inclusive, a grande dificuldade de reabilitar grandes conjuntos de dados. Quando a base do conjunto de dados segue uma distribuição normal multivariada, existem três estatísticas, derivadas de ACP, que demonstram comportamento superior quando comparadas com testes do tipo X^2. Na pesquisa da presença de "outlier" através de uma Análise de Sensibilidade aliado com ACP, pretende-se a utilização de métodos para uma análise exploratória informal e outros processos simultaneamente.

3.1 A IMPORTÂNCIA DA EXPLORAÇÃO DE DADOS UTILIZANDO O MENOR COMPONENTE

O emprego de ACP tem por objetivo avaliar a estrutura subjacente dos vetores de dados. Este procedimento, baseado na análise da variabilidade dos dados, classicamente utiliza os maiores CP. Esta escolha leva em consideração que associado a eles tem-se as maiores variâncias supondo-se que devem abranger as maiores diferenças observadas entre os dados. Uma vez comprovado que o posto da matriz de covariância é r, tal que $r < p$, estar-se-á aceitando que os $(p - r)$-últimos vetores são linearmente dependentes, podendo reescreve-los como combinações lineares dos r-primeiros vetores. Deste modo, colaboram, a rigor, com nenhuma informação extra para o

conhecimento da composição dos dados. Com esse argumento a grande maioria dos autores considera esta informação dispensável e portanto os Menores Componentes Principais(MCP) são descartados.

Hawkins e Fatti(1984) propuseram um trabalho que resgata a importância da informação contida nos MCP, aqueles que estão associados aos menores autovalores. Sua utilização está programada para detectar a interrelação entre variáveis no interior dos vetores de dados. " É bem conhecido que a inclusão de variáveis redundantes no vetor de dados pode resultar em mais componentes com falsos grandes autovalores, como uma consequência das altas correlações positivas entre as variáveis originais e suas redundantes contrapartidas." Define-se assim a capacidade dos MCP de detectarem justamente estas variáveis redundantes. Outro aspecto a ser considerado é a possibilidade de usar o MCP pela ocorrência de uma variável dependente e outras preditoras. Segundo Chang(1984) "É um caso de identificar as subregressões interpoladas entre a variável dependente e diferentes conjuntos de variáveis preditoras." Outra alternativa de uso dos MCP está em seu poder de detectar a presença de "outliers" em dados multivariados, poder este que é maior que o apresentado pelos maiores CP.

Jolliffe(1982) e Chang(1983) analisam a utilização dos MCP onde o último CP - neste caso o menor deles, pois os autovalores são selecionados de modo que $l_1 \geq \ldots \geq l_p \geq 0$, é utilizado num diagrama de dispersão , contra Y_1, o maior CP conseguindo atingir o máximo na separação dos pontos quando é o caso de duas normais misturadas. A padronização permite superar problemas causados pela invariância escalar dos CP assegurando com isto o poder de comparar os diversos coeficientes uma vez que é atribuída uma escala comum aos dados. Na verdade, a transformação aplicada é linear e garante vetor de média zero e variância 1. Os autovalores de S, a matriz de covariância amostral, são calculados de tal forma que $0 \leq l_1 \leq \ldots \leq l_p$ com os autovetores correspondentes $a_i' = (a_{i1} \ldots a_{ip})$, $i=1, \ldots, p$. Assim os últimos CP estão associados aos maiores autovalores, as maiores variâncias possíveis, com tr S = tr L, onde

$L = diag(l_1 \ldots l_p)$, e com isso detém o máximo em termos de explicação da variação dos dados.

Por outro lado, o primeiro CP, $W_1 = a_1X$, $a_1'a_1 = 1$, tem variância mínima e é não correlacionado com os outros CP, isto é, $Cov(W_1,W_i) = 0$, $i > 1$. A proposição de Mardia é de que os r menores CP são escolhidos como aqueles que estão associados à autovalores menores do que um, $l_i << 1$, $i = 1, \ldots, r$. Como se pode observar na maioria das vezes são quase todos constantes ou aproximadamente constantes pelo ínfimo valor que assumem. Segundo Hawkins e Fatti(1984), "(...) por causa do escalonamento de X, as médias de W_i, i=1, ... ,p são todas zero. As equações

$$W_i = 0, \qquad \text{isto é,}$$

$$a_iX = \sum_{j=1}^{p} a_{ij} X_j = 0 \ , \ i = 1, \ldots , r \qquad (43)$$

representam r hiperplanos ortogonais em torno dos quais os dados estão mais densamente agrupados, isto é, eles representam matematicamente quase constantes não correlacionadas e linearmente independentes (por causa da ortogonalidade) relações lineares entre os elementos de X" Esta propriedade dá suporte às idéias desenvolvidas no texto, que são procedimentos descritivos, não sendo consideradas suas propriedades inferenciais, o que garante sua aplicabilidade sem a suposição de normalidade e mesmo que algumas dela sejam não estocásticas.

1º caso IDENTIFICAÇÃO DE VARIÁVEIS REDUNDANTES

Na suposição de que os s primeiros autovalores são identicamente nulos então os CP correspondentes definem relações lineares exatas, independentes na amostra entre os elementos de X, isto permitiria reescrever s delas em termos das demais. A proposição fica em trabalhar com aquele subconjunto s de X que pode ser escrito

como combinação linear dos demais.

Para tanto definem-se dois vetores:

$X^{(1)}$: vetor do subconjunto s de X

$X^{(2)}$: vetor das restantes X_i

Particionando então a matriz dos autovetores correspondentes aos s autovalores nulos, A_s , obtém-se:

$A_s^{(1)}$: submatriz sxs de A_s correspondente à $X^{(1)}$

$A_s^{(2)}$: submatriz sx(p-s) de A_s correspondente a $X^{(2)}$

Então

$$X^{(1)} = - A_s^{(1)-1} A_s^{(2)} X^{(2)} \qquad (44)$$

$X^{(1)}$ reescrito, pode-se afirmar que seus elementos são redundantes nada acrescentando em termos de informação extra na compreensão do fenômeno. Assim $X^{(1)}$ pode ser estimado de maneira numericamente estável, como função de outros vetores e os CP associados aos menores autovalores, e que tendem à zero vão ser utilizados para, através dos coeficientes, identificar as variáveis que podem ser previstas em função das outras.

Uma nova padronização é proposta. São necessárias as seguintes suposições:

* variáveis redundantes identificadas entre as variáveis originais que foram removidas;

* variáveis restantes foram renumeradas de 1,...,p

A transformação proposta é a seguinte:

$$D = L^{-1/2} A \qquad (45)$$

$$Z = \begin{bmatrix} Z_1 \\ . \\ . \\ . \\ Z_p \end{bmatrix} = D X = L^{-1/2} W, \ \forall \ i, \ W_i = \sum_{j=1}^{p} a_{ij} X_j .$$

ou seja,

$$Z_i = \sum_{j=1}^{p} d_{ij} X_j$$

$$= \sum_{j=1}^{p} \frac{a_{ij} X_j}{\sqrt{l_i}}$$

Neste caso , $Z \sim N (0 , I)$, na amostra.

Para $Z_i=0$, seja X_m um vetor de dados tal que

$$X_m = - \sum_{j=m} \frac{d_{ij} X_j}{d_{im}} \qquad (46)$$

Z_i/d_{im}: bom preditor de X_m, não viciado com um certo erro de predição com variância residual -

$$\frac{1}{d^2_{im}} = \frac{l_i}{a^2_{im}} \quad \Longrightarrow \quad d^2_{im} = \frac{a^2_{im}}{l_i} \quad \text{e} \quad d_{im} = \frac{a_{im}}{\sqrt{l_i}}$$

Portanto d_{im} é grande quando está associado com grandes

coeficientes e menor autovalor (li) caracterizando o MCP.

Outra proposição baseada na ortogonalidade da rotação CP leva a uma nova rotação:

Seja

$L = ASA'$, $L = diag\ (\ l_1 ... l_p\ )$ e $D = L^{-1/2} A$

então

$$DSD' = L^{-1/2} A S L^{-1/2} A'$$

$$= L^{-1/2} A S A' L^{-1/2}$$

$$= L^{-1/2} L L^{-1/2}$$

$$= I$$

Seja, então

Φ : uma matriz ortogonal qualquer

Propõe-se uma nova rotação dos dados de tal forma que:

$D^* = \Phi D$

então

$$(\Phi D) S (\Phi D)' = \Phi D S D' \Phi'$$

$$= \Phi I \Phi'$$

$$= \Phi \Phi' = I$$

Comprova-se, então, que a predição baseada nas variáveis originais detecta casos extremos, a transformação D é mais

eficiente mas D^* apresenta melhores resultados pois é melhor estruturada.

Neste caso uma conclusão interessante segue pois se $d^*_{im} \longrightarrow \infty \implies X_m$ é altamente previsível pelo i-ésimo CP, $a_{ij} = 0$, $j = m$, indica que a subregressão inclue um subconjunto pequeno de outras variáveis.

Uma estrutura recomendada é aquela que leva os dados a apresentar grandes elementos e zeros em suas linhas. A sugestão apresentada neste caso é o uso do critério Varimax, simplificando as variáveis no seu conjunto e não apenas aquelas altamente previsíveis. A questão da melhor interpretação a partir do uso desta rotação foi introduzida por Gibson(1978).

2° caso SUBCONJUNTOS ALTERNATIVOS DE PREDITORES

Seja

$Y = X_1$, uma variável dependente, e ($X_2 \ldots X_p$) suas variáveis preditoras, nesse caso, " qualquer linha i de D com grande valor d_{i1} na 1ª coluna identifica uma equação de regressão para y:

$$Y = - \sum_{j=2}^{p} \left[\frac{d_{ij}}{d_{i1}} \right] X_j \qquad (47)$$

com variância residual $1/d^2_{i1}$.

Assim são definidos como "colinearidades preditivas" (...) os Menores Componentes Principais com acúmulo no elemento Y e " colinearidade não- preditiva" aqueles que não tem sobrecarga no 1º elemento." O primeiro caso indica subregressões alternativas para Y e

o segundo caso demonstra relações lineares apenas entre as variáveis restantes.

Uma forma alternativa proposta por Hawkins(1973), Webster et all(1974) é denominada de " Análise de regressão das variáveis latentes " que faz uma combinação da colinearidade preditiva com os mínimos quadrados.

$$Y = \sum_{j=2}^{p} b_j x_j,$$

onde

$$b_i = - \sum_{i \in I} \frac{d_{i1} d_{ij}}{\sum_{i \in I} d^2_{i1}} \qquad (48)$$

Sendo

I : conjunto de índices das linhas correspondentes às colinearidades preditivas, e

$1/\sum_{i \in I} d^2_{i1}$: variância residual

As técnicas propostas com base nos MCP são diferentes da ACP-Regressão que utiliza , como tradicionalmente, os maiores CP, Hawkins(1973) e Jolliffe(1982) confirmam que apesar de conseguirem captar a maior variabilidade nas variáveis preditoras não conseguem os melhores estimadores da variável dependente.

3º caso DETECTAR "OUTLIERS" EM DADOS MULTIVARIADOS

A proposição da utilização dos MCP para detectar "outliers" parte da suposição de que

X ~ N é uma normal multivariada, então

$$Z_i = \sum_{j=i}^{p} d_{ij} X_j, \quad i = 1, \ldots, p \qquad (49)$$

tal que $Z_i \sim N(0; I)$, independentes, se n é grande. Assim a presença do "outlier" será identificada pela existência de valor fora dos limites de variação de uma normal padrão. O valor limite aqui adotado é de $z_i \geq 3{,}34$. Justificando o uso dos MCP destaca-se que os coeficientes mais altamente sensíveis à presença de "outliers" nos CP estão associados aos menores autovalores $l_j << 1$, $i = 1, \ldots, r$, muito mais do que aqueles correspondentes aos maiores CP.

Duas técnicas para determinar "outliers" são apresentadas por Hawkins baseadas nos k Z_i's correspondentes aos MCP:

$$T_2 = \sum_{i=i}^{k} Z_i^2 \quad \text{e} \quad T_3 = \max_{i \leq i \leq k} |Z_i| \qquad (50)$$

Se foi avaliada uma observação que está fora do domínio sendo, então, considerada um "outlier", uma análise do tamanho e do sinal dos coeficientes permite saber qual dos X's é "outlier". O uso da matriz D e da rotação D^* facilitam a identificação e são indicados neste caso.

Uma conclusão é que " (...) um "outlier" junto dos maiores CP corresponde a uma observação geralmente muito grande(ou pequena) no "tamanho" geral que o resto da população, enquanto um

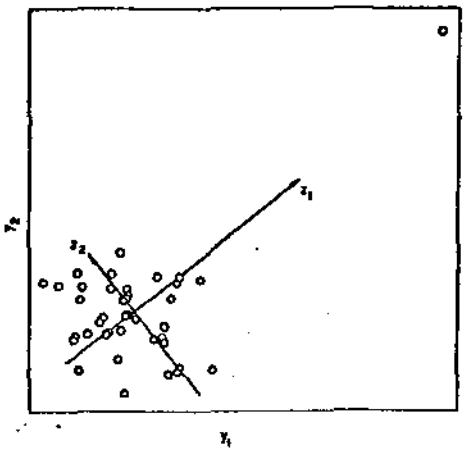
"outlier" junto dos MCP indica uma observação cuja "estrutura" multivariada difere daquela do resto da população."

3.2 CARACTERIZAÇÃO DE UM "OUTLIER"

Uma das formas de caracterizar um dado como "outlier" é a representação gráfica dos dados e outro é o estudo de um intervalo padrão de referência que apresentasse parâmetros para detectar um "outlier". De um modo geral este intervalo é construído com base na concentração central dos dados. Assim, percebe-se que este critério é perfeitamente válido quando se está tratando de variáveis que têm uma distribuição normal, ou, aproximadamente normal. Caso contrário, o ponto de corte pode podar valores que, apesar das diferenças entre ele e o restante dos dados, apresentam valores explicados por uma distribuição assimétrica com dispersão acentuada numa das extremidades da distribuição dos valores. O critério gráfico tem uma desvantagem quando se trata do caso multivariado. Se o número de variáveis é p> 3 opta-se por apresentar os dados parcialmente, combinando as variáveis que serão representadas graficamente e também alterando a entrada de variáveis. Através disto é possível caracterizar, graficamente, a existência de "outlier" multivariado.

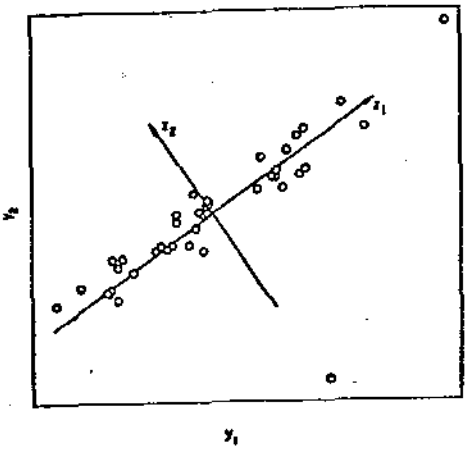
Alguns tipos diferentes de "outliers" podem ser caracterizados pelas conseqüências que geram nas observações. A importância de definir o "outlier" deve-se ao fato de que os diversos tipos vão gerar diferentes procedimentos de identificação e a Análise de Sensibilidade do próprio método empregada para detectar as alterações ante sua ocorrência. Pretende-se, especificamente, caracterizar dois tipos de "outliers":

Gráfico nº 3.1 "Outlier" do tipo que incha a variância e a covariância



→ Um "outlier" pode apresentar como conseqüência o "inchamento" da variância e da covariância e, deste modo, pode ser altamente prejudicial pelos efeitos nos resultados quando da aplicação de métodos como ACP, definidos em função da estrutura de variância.

Gráfico nº 3.2 "Outlier" do tipo que obscurece um comportamento singular



→ Um "outlier" pode obscurecer a percepção de um comportamento muito característico de um conjunto de dados, levando a redimensionar erroneamente a análise realizada. Observando o exemplo abaixo, verifica-se que o gráfico de dispersão utilizando os pontos já transformados pelo uso de CP mascara a concentração dos pontos em função da presença de "outliers". Uma análise cuidadosa revela que o eixo do primeiro CP contém praticamente toda a variabilidade dos dados, no entanto a presença de um "outlier" no eixo de Y_2, o segundo CP certamente

alterará a relação entre o percentual de variância explicada concentrada em Y_1 e o percentual de variância explicada concentrada em Y_2. Isto perturbará a análise a ser feita.

Toda a constatação da presença de "outlier" é fundamental no estudo e uso dos resultados devido às implicações que sua presença pode provocar. Uma das preocupações com a presença do "outlier" é a verificação de que tipo de erro está sendo cometido:

* O erro cometido pode ser num dos componentes do vetor resposta, e nesse caso, alterará as conclusões que partilharem aquele componente. Este tipo de erro pode alterar um dos componentes afetando apenas as interpretações que fossem feitas a partir de seus coeficientes.

* Erros sistemáticos podem ser cometidos,em todas as suas componentes e, nesse caso, podem afetar todos os componentes, alterando toda a ACP. Este fato, praticamente, invalida os resultados obtidos.

* O erro pode ser, no entanto, cometido de forma não sistemática porém em todas as componentes do vetor-resposta afetando desta forma os coeficientes c_i.

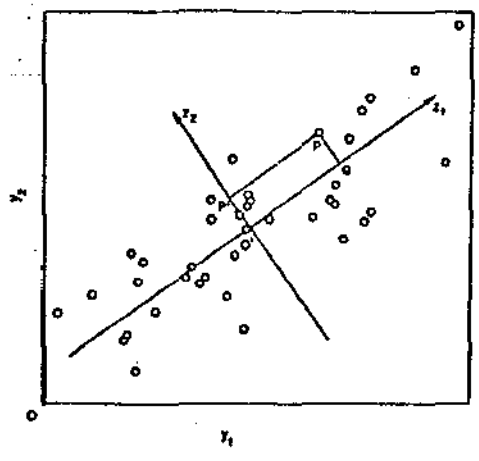
Os efeitos podem ser minimizados se for conhecido o tipo de perturbação que comumente ocorre. Estas considerações remetem a busca de procedimentos mais gerais de controle da presença de "outliers" bem como procedimentos que permitam realizar a Análise de Sensibilidade da técnica à presença de "outliers" .

3.3 PROCESSO DE IDENTIFICAÇÃO E ANÁLISE DE SENSIBILIDADE NO CASO DE "OUTLIERS"

As alterações que a presença de "outliers" podem provocar na variância e, consequentemente, nos coeficientes c_i, podem invalidar, na prática, todos os resultados obtidos com a aplicação da

ACP, daí a necessidade de realizar uma Análise de Sensibilidade que forneça elementos para avaliar a sensibilidade da técnica à presença de um "outlier", Pretende-se verificar que tipo de alterações se processam nestes casos, que mudanças ocorreram nos resultados como consequência desta presença. Quer-se saber como os coeficientes foram afetados, pois através de sua interpretação pretende-se estudar a estrutura de relações entre as p-variáveis. A própria ACP vai ser utilizada para avaliar, simultaneamente com sua aplicação, a Sensibilidade. Esta aplicação é mais apropriada, principalmente quando se tratam de dados de uma amostra que não apresenta uma estrutura de distribuição muito padrão. Este tratamento se justifica a medida em que a aplicação da técnica de CP independe do conhecimento da estrutura dos dados. As possíveis singularidades lineares apresentadas pelos valores obtidos podem ser observados e descritos através de um processo de ajustamento de uma reta aos valores já transformados como $Y_i = A'(X-\bar{X})$. É possível observar as projeções das observações nas coordenadas dos CP correspondentes aos menores autovalores, isto é, correspondentes às últimas linhas de Y. Supondo p=2, observa-se este efeito através da seguinte ilustração:

Gráfico nº 3.3 Coordenadas dos dois primeiros CP



O gráfico, apresenta X_1, X_2 as variáveis originais enquanto que Y_1, Y_2 denotam as coordenadas dos dois primeiros CP calculados a partir da matriz de covariância dos dados bivariados. Para analisar o comportamento destes dados observa-se que o ajustamento dos pontos dá-se em relação à reta Y_1, que é exatamente o ajuste obtido a partir da soma dos quadrados dos desvios

perpendiculares. Observa-se a projeção de um ponto típico dos dados, P', como mostra a figura acima. O resíduo ortogonal de um ponto P é o desvio deste ponto em relação a sua reta ajustada formando o vetor OP o qual se for observada a projeção de P, que é P', parece ser o mesmo que O'P' no eixo Y_2, o segundo CP. Segundo Gnanadesikan e Kettenring(1972): " Mais genericamente, com dados p-dimensionais, as projeções nos menores (isto é, com variância mínima) Componentes Principais pode ser relevante para estudar uma observação de um hiperplano de ajuste fechado nas q-menores coordenadas dos Componentes Principais. Podem ser relevantes para estudar os desvios de uma observação de um espaço sub-linear ajustado de dimensionalidade (p - q)."

Por outro lado, pretende-se através de uma série de mensurações estar capacitado a detectar, com uma certa precisão, quais das informações se constituem em "outliers". Este duplo aspecto referido, Análise de Sensibilidade e Processo de Identificação da informação que é "outlier", é essencial no processo de estudo do conjunto de dados.

Estas considerações remetem a busca de um sistema de proteção geral em relação à presença de "outliers" no conjunto de informações. Nesse caso, é necessário empregar técnicas com diferentes sensibilidades. É preciso ajustar o tratamento aos critérios de caracterização de um "outlier". Entre os procedimentos mais razoáveis citamos:

* os procedimentos caudais de detectação que avaliam situações extremas típicas. Neste caso, pode ser citado como exemplo a correlação entre as variáveis que vai apresentar alta distorção;

* o uso da padronização normal, $z = (x - \mu)/\sigma$, em casos que se trata de uma análise multivariada do tipo normal. Este procedimento permite detectar quando uma observação é muito isolada. Pode-se constatar que se trata de um ponto isolado quando, numa de suas componentes ou em várias e, ainda, em todas elas verificamos que a observação, quando

padronizada, segundo os critérios da distribuição normal padrão, apresenta z > 3,4 ou z < -3,4, pois a probabilidade de que ocorra um valor nesse intervalo é extremamente pequena com P $\longrightarrow$ 0. Isto é, se trata de um ponto quase impossível. Em geral, se observa que os pontos distorcidos na correlação podem prenunciar a presença de "outliers". O cuidado que se deve ter é com a suposição de normalidade em função das restrições impostas até o momento;

* Deve-se construir um grande conjunto de técnicas para avaliação da sensibilidade do procedimento à presença de "outliers" de modo a poder operar em diferentes situações. São casos em que os erros são num elemento do vetor resposta ou, então, casos em que erros sistemáticos são cometidos em todos os componentes do vetor resposta. Assim se um conjunto de técnicas gerais de detecção de "outliers" for empregada na mesma amostra pode-se considerar que a seleção foi eficiente e, indica, também, que se está de posse de um conjunto de dados perfeitamente avaliável estatisticamente. Na busca deste sistema de proteção em relação à presença de "outliers", é perfeitamente válido procurar uma técnica que se adapte à grandes quantidades de dados pois estatística opera com uma grande massa de dados e, ao mesmo tempo, interessa buscar simplicidade computacional. É interessante aliar ao próprio estudo da Sensibilidade e do processo de identificação de "outliers" o cálculo de outras técnicas que interessem na análise dos dados.

O desconhecimento quanto à presença ou não de "outliers" gera incerteza na utilização dos resultados obtidos e, em ACP, especificamente, a incerteza fica por conta da insegurança na interpretação dos coeficientes, c_{ij}. Por outro lado com base no montante de variação explicada pelos primeiros Componentes conforme a soma dos autovalores, λ_j, i=1,...,k, k<p pretende-se redimensionar o número de variáveis em estudo reduzindo os CP à um número menor que satisfaça em termos de variância. Necessariamente, este corte, pressupõe um domínio do grau de sensibilidade da variância frente a ocorrência de "outliers". O próprio método permite realizar este controle pois conforme pode ser observado, também na aplicação

realizada mais abaixo, e conforme a discussão encontrada em autores como Gnanadesikan e Kettenring(1972), os primeiros e os últimos CP são os mais interessantes do ponto de vista da Análise de Sensibilidade e de constatação sobre qual observação se constitue num "outlier". Este autores dizem que " Os primeiros são especialmente sensíveis a "outliers" que incham inapropriadamente a variância e covariância(se trabalhamos com S) ou as correlações (se trabalhamos com R). (...) Os últimos Componentes Principais são sensíveis às questões de "outliers" relacionados com o resíduo em relação à função ajustada." Uma observação problemática que pode ser detectada ao longo destes eixos apresentando resíduos pela adição de dimensões insignificantes ou, então, obscurecendo particularidades no próprio conjunto de dados.

O uso de CP não é a única fonte de referência para a análise pois se pode utilizar tanto os valores originais, X_i, como os CP, Y_i. Além disto, com uma amostra multivariada, pode-se utilizar algumas estatísticas univariadas do tipo gráficos de probabilidade sobre as linhas de X ou de Y. Estes procedimentos auxiliarão a acumular evidências quanto à improbabilidade das observações que são isoladas. Esta participação ao reunir as evidências colabora também no sentido de realizar de maneira mais eficiente a Análise de Sensibilidade. É importante que a presença de "outliers" seja corretamente identificada. Muitos são os métodos propostos para detectar a falta de ajuste de uma observação individual.

Rao(1964) propôs um estudo dos quadrados das somas dos comprimentos das projeções de uma observação nas últimas q coordenadas dos CP. Para tanto deve-se computar uma medida à qual Rao denominou de "d^2_j":

Seja

$$d^2_j = \sum_{i=p-q+1}^{p} [\ l_i'(Y_j - \bar{Y})\ ]^2 \qquad (51)$$

$$= (Y_i - \bar{Y})'(Y_i - \bar{Y}) - \sum_{i=1}^{p-q} [l_i'(Y_i - \bar{Y})]^2$$

A medida obtida será avaliada em função da magnitude dos afastamentos. Quando forem observados grandes valores de d^2_j, estes serão considerados como indicativos de que ocorre um ajuste (p - q)-dimensional muito pobre para a observação.

Esta estatística proposta por Rao pode ser complementada pelo estudo das projeções dos dados nas últimas coordenadas dos CP, isto é, naqueles componentes que apresentam os menores autovalores. Este estudo pode ser realizado utilizando-se:

⟶ Gráficos de dispersão, bi ou tridimensionais de subconjuntos bi ou tri-variados das últimas linhas de Y, em várias direções ao mesmo tempo de modo a agir como se fosse um fator.

⟶ Gráficos de probabilidade dos valores entre cada uma das linhas de Y. Unicamente porque a transformação de que trata ACP é linear não se pode aguardar que estes valores sejam normalmente distribuídos mesmo que esta seja a condição apresentada pelos dados originais. Nesse caso, o gráfico de probabilidades normal pode contribuir de forma um tanto eficiente para a análise. Quando se utiliza a estatística d_j^2 num gráfico de probabilidade gama, a análise que este gráfico proporciona pode ser ampliada com a utilização dos gráficos de probabilidade normal. A visão que este gráfico proporciona do comportamento dos pontos permite observar quais projeções definem uma observação como anormal.

⟶ Gráficos dos valores em cada uma das últimas linhas de y versus certas distâncias nos primeiros CP. Com este trabalho vai se verificar como a grande variabilidade dos dados está concentrada nos dois primeiros CP. Em se tratando de uma amostra multivariada

p-dimensional com p=5, é interessante observar as projeções das observações nos eixos das coordenadas dos CP associados aos menores autovetores a partir da distância de " ... um centróide de cada ponto projetado no plano bi-dimensional com os dois maiores autovalores. Isto pode mostrar uma certa inadequacidade do ajuste multidimensional, principalmente, se a magnitude dos resíduos nas coordenadas associadas com os menores autovalores está associada com o agrupamento dos pontos no espaço bi-dimensional dos dois autovetores correspondentes aos dois maiores autovalores." O tratamento aqui proposto não está considerando dados não-lineares. Este tipo de dados exige um tratamento próprio tanto a nível da própria definição de erro multivariado como da computação dos tratamentos e da expressão estatística das diferenças ocorridas entre a reta ajustada e as observações registradas.

Seja qual for o meio utilizado para detectar a observação duvidosa, distorcida por erro de mensuração ou de digitação é possível sugerir que

* o valor suspeito de ser um "outlier" seja excluído do conjunto de valores e que seja feita a recuperação da informação contida em ACP agora com (n-k) observações, onde k: nº de "outliers" existentes;

* seja feito uso de estimadores robustos para os parâmetros que foram afetados como os de escala (S) ou os de orientação (R).

3.4 APLICAÇÃO ATRAVÉS DE DADOS SIMULADOS

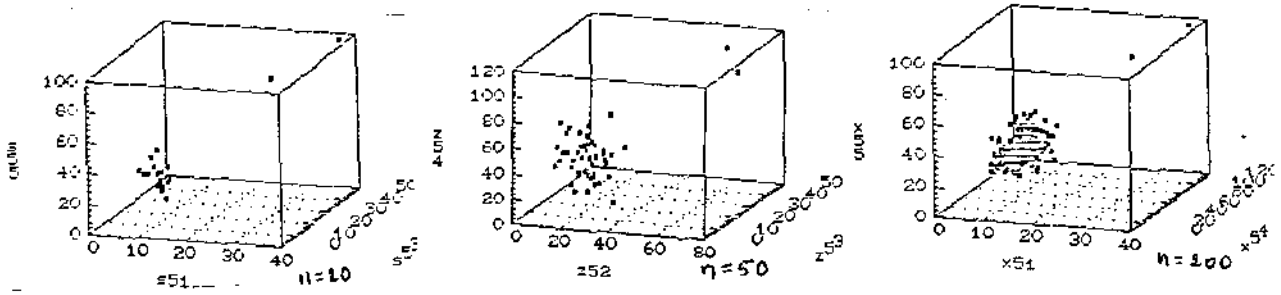
A partir da estrutura definida em (34) e (35), foram gerados três conjuntos de dados, com diferentes tamanhos de n, nos quais foram introduzidos "outliers". A introdução dos "outliers" foi conduzida de modo a caracterizar dois tipos de perturbações, entendendo-se como ponto isolado aqueles padronizados pela transformação Z , da normal padrão, é maior que 3,5 sendo considerado um ponto altamente improvável. Segue a definição das alterações.

A: "Outlier" que incha a variância e a covariância onde:

1º Alteração são pontos isolados em r das p variáveis do estudo, r < p.

2º Alteração são pontos isolados em p variáveis.

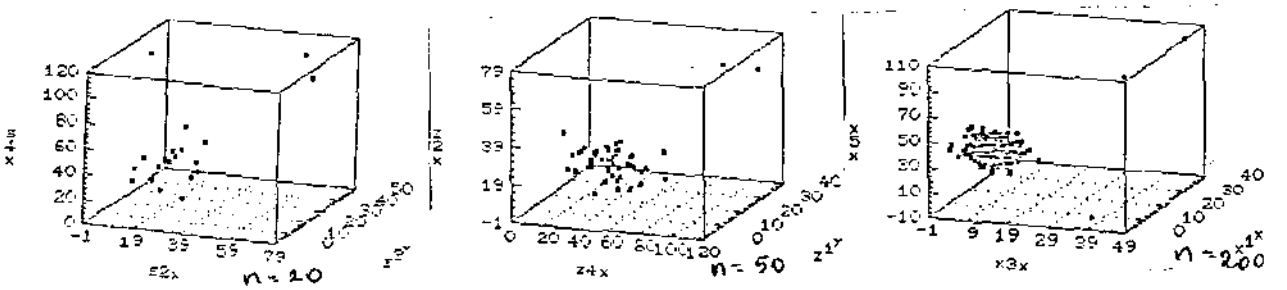
Gráfico nº 3.4 Comportamento dos dados sob o efeito da presença de "outliers" que incham as medidas de escala (1º e 2º alterações) segundo o tamanho da amostra.



B: "Outlier " do tipo que obscurece a percepção do comportamento típico do conjunto de dados.

3º Alteração Pontos isolados nas caudas das distribuições em oposição: alguns quando $x_{ij} \longrightarrow -\infty$ e outros quando $n \longrightarrow +\infty$.

Gráfico 3.5 Comportamento dos dados com a introdução de "outliers" que confundem as características dos mesmos(3º alteração), segundo a variação no tamanho da amostra.



Desta forma, varia-se, simultaneamente, o tamanho da amostra e o tipo de "outlier". Com a presença dos "outliers" ocorreram modificações nas medidas de locação e de escala, esta presença alterou a média das variáveis em maior ou menor grau onde as distorções dependem do tamanho de n, $n \longrightarrow \infty$ ou não.

Como se pode observar nos gráficos acima, a presença dos "outliers" tende a "inchar" a variância, ou então mascarar sua estrutura de variação, causando sérias alterações no cálculo de ACP. Este tipo de "outlier" desestabiliza as medidas de locação, de escala e de orientação. Todas estas medidas estão relacionadas de forma direta com o método de ACP. É portanto válido considerar que houve uma mudança geral nas condições de aplicação do método.

Tabela nº 3.1 Alterações no vetor de médias após a introdução de "outliers" de três tipos, com n=20, 50 e 200.

variáveis	X_1	X_2	X_3	X_4	X_5
n = 20					
	11.57	22.67	8.92	49.62	33.41
1ª alteração	12.17	29.69	10.99	47.87	38.69
2ª alteração	13.53	31.78	12.58	51.04	38.69
3ª alteração	12.94	30.56	14.05	53.13	36.70
n = 50					
	11.85	25.40	8.98	47.20	32.91
1ª alteração	12.05	26.12	9.61	48.49	35.12
2ª alteração	12.59	26.99	10.36	49.99	35.12
3ª alteração	12.36	26.51	11.10	51.07	34.62
n = 200					
	11.71	25.74	9.57	46.20	33.24
1ª alteração	11.80	25.91	9.71	46.44	33.90
2ª alteração	11.92	26.18	9.90	46.57	33.90
3ª alteração	11.87	26.06	10.04	46.85	33.74

Pelo tipo de alterações impostas aos dados era de se esperar modificações nas medidas de locação. No entanto, verificou-se uma variação muito pequena nas medidas de locação para o caso de grandes amostras evidenciando que o tamanho da amostra funcionaria como um redutor de perturbações. Neste caso, acredita-se que a presença dos "outliers" não deve perturbar muito os resultados de ACP. O mesmo não se verifica quando se trata de pequenas amostras que tendem a crescer em termos de média quanto maior for o número de

variáveis que apresentam alterações do tipo "outlier". A falta de estabilidade nos coeficientes produz alterações significativas nos mesmos, inviabilizando as interpretações decorrentes da aplicação de ACP. Outra observação que pode ser feita relaciona-se com as medidas que apresentaram alterações. Elas refletem a perturbação apenas naquelas variáveis que contém "outliers", pois as medidas de locação são calculadas de forma univariada ainda que sua apresentação e sua utilização seja como um vetor multivariado.

Tabela nº 3.2 Variância observada nos dados gerados após a introdução de "outliers", variando o tamanho da amostra.

Variáveis	X_1	X_2	X_3	X_4	X_5
n = 20					
	2.88	68.21	11.81	204.57	66.87
1ª alteração	14.77	114.58	43.58	363.28	385.28
2ª alteração	47.81	195.34	101.53	482.16	385.28
3ª alteração	55.69	248.84	138.58	600.96	464.04
n = 50					
	4.54	41.22	9.95	199.12	57.09
1ª alteração	8.46	65.74	22.80	270.36	186.21
2ª alteração	21.87	103.59	47.70	308.31	186.21
3ª alteração	24.56	119.01	63.88	357.83	210.08
n = 200					
	4.09	42.96	12.38	166.65	49.27
1ª alteração	5.24	48.82	15.61	184.28	81.46
2ª alteração	8.66	58.08	21.78	194.42	81.46
3ª alteração	9.26	61.74	26.31	208.61	87.49

Conforme se pode observar pelos dados acima, a variância apresenta grande desestabilidade quando n=20 e n=50, por exemplo a v.a. X_1 cresce 1833% entre a variância da amostra não alterada(2.88) em relação à variância da amostra com o 3º tipo de alteração(55.69) quando n = 20 e cresce 126% quando n = 200. Concluindo-se desta maneira que a perturbação é mais significativa nas pequenas amostras do que nas grandes amostras. Este fato, como se poderá verificar nas tabelas abaixo, refletir-se-á fortemente nos resultados dos CP.

Tabela nº 3.3 Variações do % de variância explicada segundo o tipo de alteração introduzida nos dados e o tamanho da amostra.

	Y_1	Y_2	Y_3	Y_4	Y_5
n = 20					
	59,6	19,1	18,7	2,2	0,4
1ª alteração	68,4	21,4	6,1	3,6	0,3
2ª alteração	83,6	8,0	5,3	2,5	0,3
3ª alteração	67,2	24,8	5,0	2,4	0,4
n = 50					
	64,7	18,9	12,1	3,0	1,2
1ª alteração	54,7	30,9	9,7	3,7	0,8
2ª alteração	68,4	20,0	8,2	2,5	0,7
3ª alteração	59,4	28,7	7,0	3,9	0,8
n = 200					
	61,1	17,9	15,1	4,3	1,5
1ª alteração	58,0	22,1	13,8	4,5	1,4
2ª alteração	58,1	22,9	12,4	4,9	1,4
3ª alteração	55,9	25,1	11,5	5,8	1,4

A presença de "outliers" altera, em grande parte, o percentual de variância explicada pelos componentes. Observa-se que, em geral, o primeiro componente apresenta um percentual decrescente de variância, na medida em que mais variáveis apresentam alterações em alguns registros. Esta perda de poder de explicação do primeiro vai "inchar" o % do segundo CP, na maioria dos casos, e algumas vezes também do terceiro e até do quarto componente. O % de variância do último componente mantém uma relativa estabilidade, variando muito pouco. Esta variação cresce com o número de variáveis perturbadas.

Tabela nº 3.4 Comportamento dos autovalores diante de perturbações do tipo "outliers" com amostras de tamanho n = 20, 50 e 200.

Autovalores	λ_1	λ_2	λ_3	λ_4	λ_5
n = 20					
	211,35	67,71	66,15	7,75	1,38
1ª alteração	630,80	197,41	56,94	33,25	3,10
2ª alteração	1014,31	97,72	64,92	31,02	4,15
3ª alteração	1014,44	374,68	76,25	36,51	6,22
n = 50					
	201,84	59,13	37,71	9,44	3,79
1ª alteração	302,55	171,33	53,66	20,86	4,52
2ª alteração	457,29	133,95	54,89	16,64	4,92
3ª alteração	461,08	223,04	54,78	30,24	6,21
n = 200					
	168,42	49,34	41,63	11,91	4,04
1ª alteração	194,59	74,18	46,56	15,38	4,70
2ª alteração	212,08	83,60	45,49	18,01	5,22
3ª alteração	220,27	98,99	45,61	22,82	5,72

Os autovalores refletem diretamente as alterações na variância, apresentando uma tendência a estabilidade, ainda que relativa, apenas quando se trata de grandes amostras. Os autovalores apresentam valores crescentes a medida em que aumenta o número de variáveis que apresentam observações isoladas em relação às restantes observações. Como os autovetores são definidos a partir dos autovalores qualquer perturbação que neles ocorrer vai significar re-alocamento de pesos nas variáveis que compõem o estudo.

Tabela nº 3.5 Perturbações no 1º autovetor com a presença de "outliers", segundo a alteração introduzida e o tamanho da amostra.

C_1		c_{11}	c_{21}	c_{31}	c_{41}	c_{51}
n = 20		-0.04	0.14	0.04	0.98	0.14
	1º	0.00	0.30	0.06	0.66	0.68
	2º	0.18	0.37	0.28	0.65	0.57
	3º	0.18	0.36	0.28	0.66	0.56
n = 50		-0.04	-0.08	0.04	0.99	-0.06
	1º	-0.01	0.17	0.00	0.87	0.45
	2º	0.16	0.31	0.26	0.74	0.51
	3º	0.15	0.27	0.28	0.79	0.45
n = 200		0.00	-0.10	-0.02	0.99	0.05
	1º	0.01	0.00	0.00	0.96	0.29
	2º	0.08	0.11	0.10	0.93	0.32
	3º	0.06	0.06	0.13	0.96	0.25

Observa-se na tabela 3.5 que C_1, autovetor associado à λ_1, o primeiro autovalor e o maior deles, continua concentrando valor máximo em c_{41}, como originalmente, porém de modo decrescente aumentando os outros coeficientes principalmente c_{51}. Quando n = 200 dois fatos podem ser observados: a estabilidade de c_{11} à c_{41} e a sensibilidade de c_{51} à presença de "outliers".

Tabela nº 3.6 Perturbações no 2º autovetor com a presença de "outliers" segundo o tipo de alteração e o tamanho da amostra.

C_2		c_{12}	c_{22}	c_{32}	c_{42}	c_{52}
n = 20		0.02	-0.13	-0.25	-0.11	0.95
	1º	0.22	-0.18	0.25	-0.64	0.67
	2º	0.22	-0.15	0.06	-0.63	0.72
	3º	0.17	0.40	-0.28	-0.63	0.57
n = 50		0.00	-0.33	-0.07	0.03	0.94
	1º	0.10	0.20	0.13	-0.47	0.84
	2º	0.15	0.40	0.15	-0.67	0.59
	3º	0.19	0.46	-0.06	-0.55	0.66
n = 200		0.02	-0.04	-0.09	-0.06	0.99
	1º	0.05	0.31	0.00	-0.27	0.91
	2º	0.14	0.54	0.18	-0.35	0.72
	3º	0.16	0.55	0.07	-0.26	0.77

O 2º autovetor fica altamente perturbado com a introdução dos "outliers". Todos os coeficientes se alteram causando erros de interpretação e esta conclusão é pertinente independente do tamanho da amostra. Quando n = 200 ocorre um "inchamento" de c_{22}. Este coeficiente, quando é o caso de amostras com tendência à assimetria invertem a concentração com o coeficiente c_{52}.

Tabela nº 3.7 Perturbações no 3º autovetor com a presença de "outliers", segundo o tipo de alteração e o tamanho da amostra.

C_3		c_{13}	c_{23}	c_{33}	c_{43}	c_{53}
n = 20		-0.09	0.97	-0.09	-0.15	0.09
	1º	-0.10	0.91	-0.12	-0.38	-0.03
	2º	-0.03	0.90	0.06	-0.40	-0.15
	3º	-0.17	0.83	-0.15	0.09	-0.51
n = 50		0.00	0.94	-0.08	0.11	0.32
	1º	-0.06	0.95	-0.20	-0.07	-0.23
	2º	0.04	0.82	0.08	0.01	-0.57
	3º	0.04	0.81	0.10	0.00	-0.57
n = 200		-0.01	0.99	-0.06	0.10	0.04
	1º	-0.05	0.94	-0.11	0.09	-0.29
	2º	0.00	0.81	0.06	0.09	-0.58
	3º	0.00	0.82	0.10	0.08	-0.56

O 3º autovetor apresenta uma relativa estabilidade, com perdas de valor do coeficiente c_{23}, perda esta que vai se acumular em

c_{53}, excetuando-se o caso de n = 20 quando todos os coeficientes ficam desestabilizados menos o c_{23}.

Tabela nº 3.8 Perturbações no 4º autovetor com a presença de "outliers" segundo o tipo de alteração e a variação no tamanho da amostra.

C_4		c_{14}	c_{24}	c_{34}	c_{44}	c_{54}
n = 20		-0.29	0.02	0.92	-0.09	0.24
	1º	0.23	0.18	0.91	0.10	-0.27
	2º	0.40	-0.14	0.84	-0.11	-0.32
	3º	0.32	0.14	0.84	-0.37	-0.19
n = 50		-0.21	0.05	0.97	-0.04	0.09
	1º	0.34	0.17	0.90	0.07	-0.19
	2º	0.33	-0.27	0.86	-0.09	-0.24
	3º	0.12	-0.16	0.94	-0.27	-0.06
n = 200		-0.05	0.06	0.99	0.02	0.10
	1º	0.11	0.11	0.99	0.00	-0.04
	2º	0.19	-0.18	0.95	-0.04	-0.16
	3º	0.10	-0.14	0.98	-0.11	-0.05

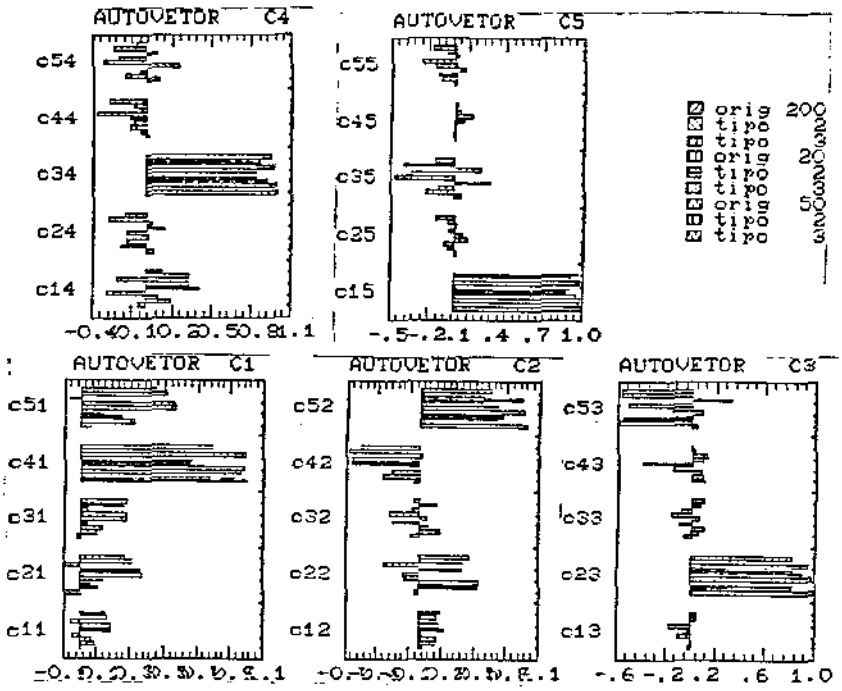
Em geral, as perturbações mais evidentes no 4º autovetor estão relacionadas com os "outliers" que mascaram o comportamento típico dos dados. É significativa, no entanto, a estabilidade de c_{34} que apresenta pequenas perdas apenas quando n = 20.

Tabela nº 3.9 Perturbações no 5º autovetor causadas pela introdução de "outliers", segundo o tipo e o tamanho da amostra.

C_5		c_{15}	c_{25}	c_{35}	c_{45}	c_{55}
n = 20		0.95	0.11	0.28	0.00	0.07
	1º	0.94	0.09	-0.30	0.07	-0.10
	2º	0.87	0.06	-0.45	0.06	-0.16
	3º	0.90	-0.04	-0.33	0.13	-0.25
n = 50		0.97	0.01	0.21	0.04	0.03
	1º	0.93	-0.01	-0.35	0.04	-0.03
	2º	0.91	-0.06	-0.39	0.00	-0.07
	3º	0.96	-0.15	-0.15	0.01	-0.17
n = 200		0.99	0.01	0.05	0.00	-0.01
	1º	0.99	0.02	-0.12	0.00	-0.07
	2º	0.97	-0.05	-0.22	-0.02	-0.11
	3º	0.98	-0.08	-0.12	0.00	-0.14

A variável que mais contribue para a variância do último CP é também a mais estável, qualquer que seja o tamanho da amostra e o tipo de "outlier".

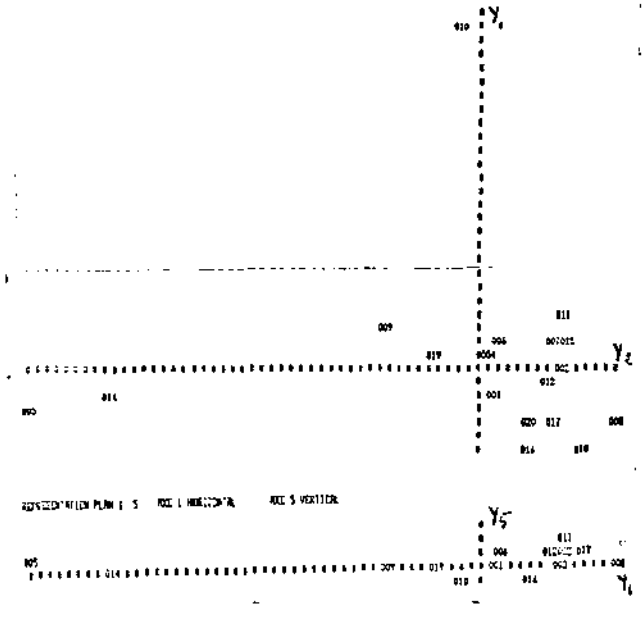
Gráfico nº 3.6 Comparação do comportamento dos coeficientes, c_i, com a introdução de "outliers".



De um modo geral é possível definir o primeiro CP e o último como sendo os componentes que menos se perturbam com a presença de "outliers". Analisando as tabelas de 3.5 à 3.9, comprova-se que Y_1, apresenta valores modificados quando se trata de casos de pequenas amostras, n=20 e n=50. O último CP é o que apresenta a maior estabilidade. Qualquer conclusão baseada no segundo CP pode estar sendo prejudicada pelas significativas alterações apresentadas pelos coeficientes das p-variáveis. Como já foi dito anteriormente a variância apresenta grandes alterações no segundo CP, em geral o "inchamento" da variância resulta em crescimento do 2º

autovalor, conforme se pode constatar nas tabelas 3.3 e 3.4. As modificações nos coeficientes com a presença de "outliers", produz realocamento dos pesos das variáveis. Na maioria dos casos o 2º, o 3º e o 4º autovetor apresentam inversão dos coeficientes, alocando maior peso em variável diferente da observada nos dados originais, sem a ocorrência de "outliers". Porém verifica-se que o componente C_4 é estávelno caso de grandes amostras, n = 200. Estes fato,altera a forma de análise da participação de cada variável no fenômeno em estudo. Estes resultados, vêm ao encontro da teoria de Hawkins(1985), que propõe o último CP para uma análise da presença de "outliers". É aconselhável, no entanto, que o primeiro e o último sejam utilizados na análise do comportamento dos dados. Um gráfico dos mesmos pode revelar a presença das observações que apresentam distorções. Sabe-se também que este tipo de dispersão é bastante poderoso quando se trata de separar duas normais misturadas conforme Chang(1983). É possível utilizar o mesmo para melhor avaliar as grandes aglomerações de observações que podem estar formando "clusters". Por outro lado, esta relação sugere que é possível construir uma medida para Análise de Sensibilidade dos CP na ocorrência de "outliers" baseada no % de variância explicada.

Gráfico nº 3.7 Plano Y_1 e Y_5 comparado com o plano Y_1 e Y_2, destaque dos "outliers".



Como se observa acima, o plano Y_1 e Y_5 demonstra claramente quais são as observações que estão inchando a variância descaracterizando o estudo de ACP. Enquanto isto o plano Y_1 e Y_2 se presta mais para evidenciar casos de "outliers" que obscurecem o comportamento típico do conjunto de dados em termos de variância. Portanto, a utilização

do último componente auxilia na pesquisa de "outliers" pela sua sensibilidade.

CAPÍTULO 4

ESTABILIDADE DOS CP FRENTE À PERTURBAÇÕES NA VARIÂNCIA

A técnica de ACP têm por base uma transformação linear, $Y_i = c'X$, $X = X - \bar{X}$, capaz de rotacionar ortogonalmente os pontos no espaço, de modo que novas variáveis são escritas a partir das originais, sem que se perca a variabilidade original, isto é,

conservando o espalhamento original.

De tal modo que:

$$tr\ \Sigma = tr\ \Lambda,$$

onde

$$tr\ \Sigma = \Sigma_i\ \sigma_{ii} \quad e \quad tr\ \Lambda = \Sigma_i\ \lambda_i\ ,\ i = 1, \ldots , p$$

A transformação, no entanto, reestrutura a variância garantindo que ao primeiro CP corresponda a maior variância permitindo que os primeiros CP captem o máximo da diversidade dos dados observados ao apresentarem os maiores afastamentos dos valores em relação ao vetor média. Desse modo, tem-se, segundo Krzanowski(1979), que "Análise de Componentes Principais é simplesmente, uma rotação dos eixos para novas posições. Estes novos eixos são tais que as projeções ortogonais dos pontos amostrais têm espalhamento decrescente."

Sob este ponto de vista, qualquer alteração na estrutura de variância deve ser avaliada quanto à seus efeitos nos coeficientes, c_i, uma vez que estes estão definidos como os autovetores associados aos autovalores da matriz covariância. Krzanowski vai definindo a questão, ponto a ponto, numa sucessão de artigos: (1971), (1979a), (1979b), (1982), (1983), (1984). Estes artigos começam com estudos sobre os métodos de Análise Multivariada mostrando suas semelhanças algébricas e culminam no artigo de 1984 sobre Análise de Sensibilidade dos CP. Neste último estudo, o autor definiu uma expressão analítica simples para a realização da Análise de Sensibilidade, baseada no ângulo θ formado entre o autovetor original e o autovetor calculado a partir de retiradas de otimalidade, perdas não maior do que ε, na função critério V, definida por Krzanowski(1984) como $V = c'\Sigma\ c$. A magnitude de θ representa o grau de afastamento entre o valor original e o valor perturbado. Este teste é bastante simples e pode ser realizado simultaneamente com a ACP. A Análise de Sensibilidade conduzida desta forma permite que se verifique a estabilidade dos CP e, neste caso,

referenda ou não, qualquer interpretação realizada a partir dos autovetores encontrados. O estudo de Krzanowski parte da avaliação da função critério utilizada para o cálculo de ACP.

Observa-se que ACP é utilizado para a redução da dimensionalidade, a redução das variáveis originais, a identificação de "outliers", a formação de "clusters", o cálculo de ACP -Regressão, etc. Porém a mais clássica finalidade, ao lado da redução da dimensionalidade, têm sido o uso dos coeficientes, c_i, para interpretar a composição dos dados em termos de estrutura de relações entre as variáveis na construção do vetor-resposta. Este aspecto remete à percepção do rigor necessário nos cálculos para sustentar esta interpretação. Fundamentalmente, é preciso garantir a sua estabilidade ante a desestabilização da variância, neste caso, o principal aspecto a ser considerado é o de pequenas perdas na mesma. Estas pequenas retiradas de otimalidade significam aumentos ou decréscimos na variância, qualquer que seja a causa. Existe um trabalho de Green(1977) tentando estabelecer uma forma de mensurar a estabilidade que, no entanto, não consegue definir uma técnica que se ajuste aos usuais métodos que são aplicados em ACP.

A proposição do método de Análise de Sensibilidade por Krzanowski(1984) foi fundamentada em estudos de De Sarbo et al.(1982) que, no decorrer de investigações em torno de Análise de Correlações Canônicas, chegou a uma linha geral de aproximações em casos de Análise de Sensibilidade.

4.1 UMA APROXIMAÇÃO PARA DEFINIR A REGIÃO DE INDIFERENÇA

Com a finalidade de embasar o estudo de uma aproximação para definir uma região de indiferença em torno de c, o que leva a função critério V, definida por Krzanowski(84) como $V=c'\Sigma c$, a um ponto de máximo, é necessário definir essa rotação ortogonal.

Seja

$\left[\; T \; \right]_{\alpha}^{\alpha}$: matriz simétrica positiva definida

β : um vetor 1 x p

α e β: bases ortonormais

$$\beta = \begin{bmatrix} z_1 \\ z_2 \\ . \\ . \\ z_p \end{bmatrix}$$

Pretende-se realizar uma mudança de base nos dados originais com a finalidade de conseguir expressá-los através de variáveis não-correlacionadas. Esta mudança leva à expressão de uma elipse. Segundo Boldrini(1986),

Seja

$$\left[v \right]_{\alpha}' \left[T \right]_{\alpha}^{\alpha} \left[v \right]_{\alpha} \qquad (52)$$

uma função a ser definida em termos de mudança de base.

como $\left[v \right]_{\alpha} = \left[I \right]_{\alpha}^{\beta} \left[v \right]_{\beta}$, então

$$\left[v \right]_{\alpha}' \left[T \right]_{\alpha}^{\alpha} \left[v \right]_{\alpha} = \left(\left[I \right]_{\alpha}^{\beta} \left[v \right]_{\beta} \right)' \left[T \right]_{\alpha}^{\alpha} \left[I \right]_{\alpha}^{\beta} \left[v \right]_{\beta}$$

$$= \left[v \right]_{\beta}' \left(\left[I \right]_{\alpha}^{\beta} \right)' \left[T \right]_{\alpha}^{\alpha} \left[I \right]_{\alpha}^{\beta} \left[v \right]_{\beta}, \quad (53)$$

mas como α e β são bases ortonormais então $\left[I \right]_{\alpha}^{\beta}$ é uma matriz ortogonal e portanto

$$\left(\left[I \right]_{\alpha}^{\beta} \right)' = \left(\left[I \right]_{\beta}^{\alpha} \right)^{-1} = \left[I \right]_{\beta}^{\alpha} . \qquad (54)$$

Assim, tem-se em (53), utilizando (54)

$$\left[v \right]_{\beta}' \left[I \right]_{\beta}^{\alpha} \left[T \right]_{\alpha}^{\alpha} \left[I \right]_{\alpha}^{\beta} \left[v \right]_{\beta} \qquad (55)$$

como

$$\left[I \right]_{\beta}^{\alpha} \left[T \right]_{\alpha}^{\alpha} \left[I \right]_{\alpha}^{\beta} = \left[T \right]_{\beta}^{\beta} \qquad (56)$$

e, utilizando (56) em (55), vem

$$\left[v \right]_{\beta}' \left[T \right]_{\beta}^{\beta} \left[v \right]_{\beta} = \sum_{i=1}^{p} \lambda_i z_i^2$$

$$= \lambda_1 z_1^2 + \lambda_2 z_2^2 + \ldots + \lambda_p z_p^2 \qquad (57)$$

O fato da ACP ter sido fundamentalmente delineada para acumular o máximo de variância nos primeiros CP remete à necessidade de se estabelecer uma Análise de Sensibilidade quando se pretende avaliar os efeitos da retirada de otimalidade da função critério. Esta maximalidade em termos de variância nos CP implica em encontrar os limites para o seu máximo. Com isso, querer-se-á o tipo de comportamento dos coeficientes, c_i, nos limites da função critério, quando esta passa por um máximo. Quer-se conhecer como se comportam os coeficientes, c_i, nos limites de perturbações da função critério, maximizada. Vai-se avaliar esta região utilizando uma aproximação de

De Sarbo et allii(1982) baseados numa série de Taylor.

Seja

$V(c)$ uma função critério quadrática em c, que deve ser maximizada.

$V^* = V(c^*) \longrightarrow$ ponto de máximo atingido em $c=c^*$.

Um erro suportável caracteriza uma região em torno do ponto de máximo tal que pequenas oscilações, pequenas perdas ou acréscimos em V, são definidas como no máximo ε de V^*. Com isto está-se construindo uma região de "indiferença" ($c| V^* - V \leq \varepsilon$) que deve ser limitada de modo que se possa obter ($c| V^* - V = \varepsilon$). Para obter este limite utiliza-se uma expansão em série de Taylor próximo a c^*, o ponto onde V atinge seu máximo.

Desenvolvimento em série de Taylor:

$$f(a + v) = f(a) + \partial f(a)v + 1/2\ \partial^2 f(a)v^2 + \ldots + 1/p!\ \partial^p f(a)v^p + r(v)$$

Como $V(c)$ é a função a ser maximizada e tem-se que $c = c^* + r$ onde $r + (c - c^*)$

então:

$$V(c) \simeq V(c^* + r) = V(c^*) + \partial V(c^*)r + 1/2\ \partial^2 V(c^*)r^2 \qquad (58)$$

o que, em linguagem matricial pode ser traduzido por:

$$V(c) \simeq V(c^*) + \left[\frac{\partial V(c^*)}{\partial x_1} \ldots \frac{\partial V(c^*)}{\partial x_p}\right]' r + \frac{1}{2} r' \begin{bmatrix} \frac{\partial^2 V(c^*)}{\partial x_1^2} & \ldots & \frac{\partial^2 V(c^*)}{\partial x_1 \partial x_p} \\ \vdots & & \vdots \\ \frac{\partial^2 V(c^*)}{\partial x_p x_1} & \ldots & \frac{\partial^2 V(c^*)}{\partial x_p^2} \end{bmatrix} r$$

$$V(c) \simeq V(c^*)+\left[\frac{\partial V(c^*)}{\partial x_1}\ldots\frac{\partial V(c^*)}{\partial x_p}\right](c-c^*)+\frac{1}{2}(c-c^*)'\begin{bmatrix}\frac{\partial^2 V(c^*)}{\partial x_1^2}\ldots\frac{\partial^2 V(c^*)}{\partial x_1 \partial x_p}\\ \vdots \\ \frac{\partial^2 V(c^*)}{\partial x_p x_1}\ldots\frac{\partial^2 V(c^*)}{\partial x_p^2}\end{bmatrix}(c-c^*)$$

$$V(c) \simeq V^* + g^{*\prime} r + 1/2\ r' H^* r \qquad (59)$$

onde

$r = (c - c^*)$

g^* = vetor gradiente de $V(c)$ avaliado em $c=c^*$

H^* = matriz Hessiana de $V(c)$ avaliada em $c=c^*$

Quando V atinge o ponto máximo, g^* é nula e H^* negativa (semi) definida, portanto:

$$V(c) \simeq V^* + 1/2\ r' H^* r \qquad (60)$$

Como se impõe que a região de "indiferença" seja limitada a perturbações no máximo de tamanho ε, obtém-se:

$$(c \mid V^* - V \leq \varepsilon)$$

então, faz-se

$$| r' \, H^{*} \, r | \leq 2 \, \varepsilon$$

Como H^{*}, a matriz Hessiana é negativa (semi) definida, define-se $A = - H^{*}$, de modo que A é positiva (semi) definida e nesse caso pode-se fazer:

$$r'A \, r = 2 \, \varepsilon \qquad (61)$$

Como

$$\lambda r = A^{-1} \, r \Longleftrightarrow Ar = \lambda^{-1} \, r$$

Logo

$$r'A \, r = \lambda_1 z_1^{\,2} + \ldots\ldots + \lambda_p z_p^{\,2} = 2 \, \varepsilon$$

$$= \frac{z_1^{\,2}}{\frac{1}{\lambda_1}} + \ldots + \frac{z_p^{\,2}}{\frac{1}{\lambda_p}} = 2 \, \varepsilon \qquad (62)$$

Esta é uma equação p-dimensional que define uma elipsóide. Como se vê (59) está definida semelhante à função critério $V = c' \Sigma \, c$, logo, segundo Krzanowski(1984), esta elipsóide define uma região dentro do espaço dos coeficientes com r mudanças nos coeficientes as quais resultam numa redução de no máximo ε na função critério V.

No caso de ACP a perturbação que interessa é o limite máximo que pode ocorrer nos coeficientes, c_i, sem que as alterações na função critério sejam maior que ε. Como $r = c - c^{*}$, para obter a máxima diferença vai ser maximizada $r'r$, sujeito à seguinte

restrição: $r^{t}A\ r = 2\varepsilon$. A solução deste sistema é encontrada através da aplicação de multiplicadores de Lagrange, na função L, definida como:

$$L = r^{t}r - \lambda(r^{t}A\ r - 2\varepsilon)$$

$$L = r^{t}r - \lambda r^{t}A\ r + 2\lambda\varepsilon$$

Para conhecer quando L passa por um ponto de máximo deve-se obter a derivada primeira em relação a r e igualá-la a zero e a derivada segunda deve ser negativa. Diferenciando L em relação à r, obtém-se:

$$\frac{\partial L(r)}{\partial r} = \frac{\partial (r^{t}r)}{\partial r} - \frac{\partial(\lambda r^{t}A\ r)}{\partial r} + \frac{\partial(2\lambda\varepsilon)}{\partial r}$$

$$= 2r - 2\lambda A\ r \qquad (63)$$

Para que $\frac{\partial L(r)}{\partial(r)} = 0$, precisa-se que

$$r - \lambda A r = 0$$

$$(\lambda^{-1} I - A)\ r = 0 \qquad (64)$$

$\Longrightarrow$ Como está se buscando a perturbação máxima aceitável para c, $r \neq 0$ então resulta em (64). Logo quer-se que

$$\lambda^{-1} I = A$$

A solução existe e é dada pelo valor de r que satisfaz (62). Neste caso, o valor de r que satisfaz a equação é dado pelo autovetor associado ao maior autovalor de A^{-1}. Quando se olha para sua inversa, A(se A é singular), r estará associado ao menor autovalor , não-nulo, de A, sujeito à restrição (61) imposta acima.

No espaço p-variado as observações realizadas são n pontos p-dimensionais.

Supondo

c: uma direção no espaço p-dimensional

c^*: uma direção perturbada no espaço p-dimensional

Sendo c, c^* definidos como os autovetores associados com os autovalores calculados a partir de Σ: matriz de covariância e ortogonalizados com a indeterminação escalar retirada pela imposição da restrição $c'c = c^{*\prime}c^* = 1$. Logo, maximizar $r'r$, dado que $r = (c - c^*)$, é o mesmo que maximizar $(c - c^*)'(c - c^*)$

$$(c - c^*)'(c - c^*) = c'c + c^{*\prime}c^* + 2\, c^{*\prime}c$$

$$= 1 + 1 + 2\, c^{*\prime}c$$

$$= 2\,(1 - c^{*\prime}c),$$

mas

$$\cos\theta = \frac{\langle c^*, c\rangle}{\|c^*\|\,\|c\|}, \quad \|c^*\| = c'c = 1 \qquad (65)$$

então

$$\boxed{\cos\theta = \langle c^*, c\rangle} \qquad (66)$$

portanto

$$(c - c^*)'(c - c^*) = 2(1 - \cos\theta)$$

onde θ : ângulo entre c, c^* , assim a maior diferença entre c, c^* tal que $r`A\ r = 2\varepsilon$, é encontrar o componente c, cujo ângulo θ com c^* é um máximo mas cuja variância é pelo menos ε menor que a de c^*. Para realizar a Análise de Sensibilidade basta aplicar, concomitantemente com o estudo de ACP, o cálculo do ângulo θ formado entre c, c^* quando ocorre uma perda ou um crescimento de no máximo ε na função critério, $V = \wedge$, isto após a aplicação de CP.

Segundo Krzanowski(1984) " Na análise a seguir, estaremos olhando para os menores autovalores não-nulos e correspondentes autovetores da negativa da matriz Hessiana A. (...) o maior autovalor não-nulo e seu correspondente autovetor providencia (...) a mais sensível das direções de partida. Isto dá a menor perturbação em c o qual leva a um decréscimo ε na função critério."

4.2 O PRIMEIRO COMPONENTE PRINCIPAL

A aplicação das idéias acima, devem ser vistas inicialmente num dos CP, antes da generalização dos resultados. O primeiro CP, $Y_1 = c_1`X$, é de maior interesse neste caso, pois pela própria metodologia de CP, retém a variância máxima. Após terem sido calculados os autovalores e os autovetores associados a estes e se escolhe $\lambda^* = \lambda_1 = c_1`S\ c_1$ para ser o máximo. Por Mardia(1979), cap 8, tem-se: " **Teorema 8.2.2 Nenhuma Combinação Linear Padronizada tem uma variância maior que λ_1, a variância do primeiro Componente Principal.**"

Seja

$V = c`S\ c$ , restrito à $c`c = 1$

$V = c`S\ c - \lambda(c`c - 1)$,

$$c = \begin{bmatrix} c^1 \\ . \\ . \\ c^p \end{bmatrix} \qquad S = \begin{bmatrix} s_{11} & . & s_{1p} \\ . & . & . \\ . & . & . \\ . & . & . \\ s_{p1} & . & s_{pp} \end{bmatrix}$$

e

$$c'Sc = (c^1 \quad c^2 \quad \ldots \quad c^p) \begin{bmatrix} s_{11} & \ldots & s_{1p} \\ . & & . \\ . & & . \\ s_{1p} & \ldots & s_{pp} \end{bmatrix} \begin{bmatrix} c^1 \\ . \\ . \\ c^p \end{bmatrix}$$

$$= \Sigma_i \; c^i s_{i1} c^1 + \Sigma_i \; c^i s_{i2} c^2 + \ldots + \Sigma_i \; c^i s_{ip} c^p$$

$$= \Sigma_i \; \Sigma_j \; c^i \; c^j \; s_{ij}$$

e

$$c'c = (c^1 \; c^2 \; \ldots \; c^p) \begin{bmatrix} c^1 \\ c^2 \\ .. \\ c^p \end{bmatrix}$$

$$= \Sigma_i \; [\; c^i \;]^2$$

então

$$V = c'S \; c - \lambda(c'c - 1)$$

$$V = \Sigma_i \; \Sigma_j \; c^i \; c^j \; s_{ij} - \lambda(\; \Sigma_i \; [\; c^i \;]^2 - 1) \qquad (67)$$

Seja, então,

$$c^* = c_1 \quad e \quad \lambda^* = \lambda_1$$

E, seja

$$H = \begin{bmatrix} \frac{\partial^2 V}{\partial c^1 \partial c^1} & \frac{\partial^2 V}{\partial c^1 \partial c^2} & \cdots\cdots & \frac{\partial 2 V}{\partial c^1 \partial c^p} \\ \cdots & \cdots & & \cdots \\ \frac{\partial^2 V}{\partial c^p \partial c^1} & \frac{\partial^2 V}{\partial c^p \partial c^2} & \cdots\cdots & \frac{\partial^2 V}{\partial c^p \partial c^p} \end{bmatrix}$$

$$H = \begin{bmatrix} 2(s_{11} - \lambda\delta_{11}) & 2(s_{12} - \delta_{12}) & \ldots & 2(s_{1p} - \lambda\delta_{1p}) \\ \ldots & \ldots & & \ldots \\ \ldots & \ldots & & \ldots \\ \ldots & \ldots & & \ldots \\ 2(s_{p1} - \lambda\delta_{p1}) & 2(s_{p2} - \lambda\delta_{p2}) & \ldots & 2(s_{pp} - \lambda\delta_{pp}) \end{bmatrix}$$

com

$$\frac{\partial^2 V}{\partial c^1 \partial c^j} = 2s_{ij} - 2\lambda\delta_{ij}, \quad \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Como foi requerido que $c^* = c_1$ e $\lambda^* = \lambda_1$ de modo que

$$H^* = 2S - 2\lambda_1 I \qquad (68)$$

E, dado que λ_i ($i = 1, \ldots, p$), $\lambda_1 \geq \ldots \geq \lambda_p$ são os autovalores de S e os c_i seus correspondentes autovetores, pode-se concluir que os autovalores de H^* vão ser da forma $\lambda_h = 2(\lambda_i - \lambda_1)$, com

autovetores correspondentes c_i, (i=1,...,p). Como foi definido que $A = -H^*$, os autovalores de A são do tipo $2(\lambda_1-\lambda_i)$, seus autovetores associados aos seus autovalores são também do mesmo tipo' c_i, (i=1, ..., p).

Seja, então, os autovalores de A:

$$0 < 2(\lambda_1 - \lambda_2) < 2(\lambda_1 - \lambda_3) < \ldots < 2(\lambda_1 - \lambda_p),$$

respectivamente associados a $c_2, c_3, \ldots, c_p$.

Como foi demonstrado anteriormente, deve-se usar o menor autovalor, não-nulo de A e seu autovetor associado c_i, como sendo o autovetor que proporciona a maior das perturbações com uma perda de no máximo ε na função critério V. Seja, então, o menor autovalor não-nulo de A dado por $\boxed{2(\lambda_1 - \lambda_2)}$ e seu autovetor c_2, associado com o segundo maior autovalor de S.

Logo, com as seguintes restrições:

$r'Ar = 2\varepsilon$, a perturbação máxima que pode ocorrer em c_i com o máximo ε de retirada de otimalidade de λ_1 implica em definir $r = kc_2$. Deste requerimento, pode-se retirar o valor de k, como sendo

$$r'A\,r = 2\varepsilon\,, \qquad r = kc_2$$

$$(kc_2)'A\,(kc_2) = 2\varepsilon$$

$$k^2\,c_2'A\,c_2 = 2\varepsilon\,, \qquad A = 2(\lambda_1 I - S)$$

$$k^2\,c_2'(2\lambda_1 I - 2S)c_2 = 2\varepsilon,$$

$$2k^2\,[c_2'\lambda_1 I c_2 - c_2'S\,c_2] = 2\varepsilon$$

$$2k^2\,[\lambda_1 c_2'c_2 - c_2'S\,c_2] = 2\varepsilon,\ c_2'c_2 = 1\ e\ c_2'S\,c_2 = \lambda_2$$

$$k^2 [\ \lambda_1 - \lambda_2\ ] = \varepsilon \quad , \qquad (69)$$

segue deste resultado que

$$k = \left[\frac{\varepsilon}{(\ \lambda_1 - \lambda_2\ )} \right]^{1/2} \qquad (70)$$

Portanto o coeficiente que correspondente à uma alteração máxima em λ_1, e, conseqüentemente, com uma diferença máxima ε de c_1 é dado por $c = c_1 + r$, $r = k\ c_2$:

$$\begin{aligned} c &= c_1 + r \\ &= c_1 + k\ c_2 \\ &= c_1 \pm c_2 \left[\frac{\varepsilon}{(\lambda_1 - \lambda_2)} \right]^{1/2} \end{aligned} \qquad (71)$$

Ortonormalizado, deve-se requerer que $c`c = 1$, logo

$$\begin{aligned} c`c &= (c_1 + r)`(c_1 + r) \\ &= c_1`c_1 + 2c_1`r + r`r \\ &= c_1`c_1 + c_2`c_2 \left[ \frac{\varepsilon}{(\lambda_1 - \lambda_2)} \right] \pm 2c_1`c_2 \left[\frac{\varepsilon}{(\lambda_1 - \lambda_2)} \right]^{1/2}, \end{aligned} \qquad (72)$$

Considerando as propriedades dos CP, que garantem $c_1`c_1 = c_2`c_2 = 1$ e $c_1`c_2 = 0$, obtém-se:

$$c`c = 1 + \left[\frac{\varepsilon}{(\ \lambda_1 - \lambda_2\ )} \right] \qquad (73)$$

Após todas estas considerações, é possível encontrar o componente perturbado que provoca uma desestabilização na variância de no máximo ε. Chamar-se-á de $c_{(1)}$, este coeficiente. Seja, então

$$c_{(1)} = c/\sqrt{c'c}$$

$$c_{(1)} = \frac{c_1 \pm c_2 \left[\frac{\varepsilon}{(\lambda_1 - \lambda_2)}\right]^{1/2}}{\left[1 + \frac{\varepsilon}{(\lambda_1 - \lambda_2)}\right]^{1/2}} \qquad (74)$$

Procura-se, então, como medida da estabilidade o ângulo θ_1 entre $c_{(1)}$ e c_1, é dado por

$$\cos \theta_1 = \frac{(c_{(1)}, c_1)}{||c_{(1)}|| \; ||c_1||} \qquad (75)$$

$$\cos \theta_1 = [1 + \varepsilon/(\lambda_1 - \lambda_2)]^{-1/2} \qquad (76)$$

4.3 GENERALIZAÇÃO DA ANÁLISE DE SENSIBILIDADE

Examinar o efeito de uma retirada de otimalidade do j-ésimo CP, implica em verificar as perturbações ocorridas em c_j quando ocorre uma perda máxima ε na função critério V correspondente λ_j. Seja, então

$$S_{(j)} = S - \lambda_1 c_1 c_1' - \ldots - \lambda_{j-1} (c_{j-1})(c'_{j-1}) \qquad (77)$$

Obviamente λ_j é o autovalor máximo de $c'S_{(j)}c$ e será atingida no ponto $c^* = c_j$. A álgebra repete o que ocorre com o primeiro CP, seguindo o desenvolvimento já visto anteriormente. Dado que os autovalores de A são

$$0 < 2(\lambda_j - \lambda_{j+1}) < \ldots < 2(\lambda_j - \lambda_p) < 2\lambda_j$$

Desta definição, tira-se que o menor autovalor não-nulo de A para o j-ésimo CP é dado por $2(\lambda_j - \lambda_{j+1})$ e seu autovetor associado é c_{j+1}. Deste resultado, obtém-se que

$$c_{(j)} = \frac{c_j \pm c_{j+1}\left[\dfrac{\varepsilon}{(\lambda_j - \lambda_{j+1})}\right]^{1/2}}{\left[1 + \dfrac{\varepsilon}{(\lambda_j - \lambda_{j+1})}\right]^{1/2}} \qquad (78)$$

Neste caso, o ângulo θ_j formado entre o verdadeiro coeficiente e o coeficiente maximamente perturbado com uma retirada de tamanho limitado em ε, pode ser calculado da seguinte forma:

$$\boxed{\cos\theta_j = (1 + \varepsilon/(\lambda_j - \lambda_{j+1}))^{-1/2}}, \quad j = 1,\ldots,(p-1) \qquad (79)$$

Esta Análise de Sensibilidade pode ser calculada para diferentes tamanhos de ε, sendo que $\varepsilon = k\lambda_j$ e k é um valor muito pequeno pois representa a perda de otimalidade da função critério. Logo, k= 0.10; 0.05; ou 0.01.

Se, ao invés da pequena perda ε na função critério, se está olhando para um pequeno acréscimo ε na função critério, a forma de dimensionar as perturbações nos CP não precisa de novas

demonstrações matemáticas. Basta que se olhe para a minimização de S. Diferenciando em relação à $\tilde{c} = c - c_j$, no sentido de minimizar as diferenças, basta que a derivada segunda seja positiva. O coeficiente que apresenta a diferença máxima quando a variância apresenta um acréscimo de no máximo ε que a de c_j é dado por

$$c_{(j)} = \frac{c_j \pm c_{j-1}[\varepsilon/(\lambda_{j-1} - \lambda_j)]^{1/2}}{[\,1 + \varepsilon / (\lambda_{j-1} - \lambda_j)\,]^{1/2}} \qquad (80)$$

Quanto ao ângulo θ_j entre $c_{(j)}$ e c_j, pode ser obtido através do seguinte cálculo:

$$\cos \theta_j = [\,1 + \varepsilon/(\lambda_{j-1} - \lambda_j)\,]^{-1/2} \qquad j=2,\ldots,p \qquad (81)$$

O estudo das perturbações dos CP conforme pode ser visto nas equações derivadas para $c_{(1)}$, $c_{(j)}$ dão conta da arbitrariedade que existe em ACP quanto à escolha do sinal. Isto remete a um questionamento sobrea ordem de importância entre a direção e a magnitude. A direção significa perda ou acréscimo de um ε na função critério e basta que se multiplique A por (-1) para que os sinais sejam invertidos. Portanto, o interesse da Análise de Sensibilidade reside em acompanhar a magnitude da mudança nos coeficientes que irão definir a estabilidade dos componentes. Por outro lado, verifica-se que medir o ângulo θ_j entre o coeficiente original e o perturbado é uma função inversa da diferença que existe entre os autovalores. Isto resulta numa constatação de que a magnitude da variância não aloca estabilidade ao coeficiente, independente de sua posição no conjunto dos autovalores pode ocorrer perturbação ou não. Neste caso, a perturbação é aleatória e deve ser

constatada particularmente em cada conjunto de observações.

4.4 UM EXEMPLO NO ESTUDO DE KRZANOWSKI

Krzanowski utiliza dados apresentados por Mardia(1979) recalculados com a finalidade de evitar interferência quanto à adoção de arredondamentos. Dois valores são assumidos para k, k_1=0.10 e k_2=0.05, para o cálculo de $\varepsilon = k\lambda_j$. Examine-se, então os resultados obtidos comparando os coeficientes maximamente perturbados e o valor de θ, ângulo formado entre o coeficiente original e o perturbado.

Na Tabela 4.1, os componentes 2 e 5 são os que mostram as maiores
perturbações representadas pelo ângulo formado entre a direção original e a direção perturbada com um valor máximo ε de alteração na variância. Observando o ângulo formado entre os coeficientes originais e os que tiveram sua variância com uma perda de otimalidade ao nível de 5%, observa-se que θ = 14,19,16,17 e 24. Logo as maiores perturbações ficam por conta do segundo e quinto componente. Quando as alterações provocadas na variância estiveram ao nível dos 10%, o ângulo entre a direção original e a perturbada apresentou os seguintes resultados θ = 19,26,23,23 e 33 Sendo que novamente as alterações máximas ocorreram em Y_2 e Y_5. Com uma perturbação de 5%, existe uma alteração significativa em c_2 que apresenta um valor crescente enquanto que todos os outros coeficientes decrescem. Já, numa perturbação máxima de 10%, observa-se que o segundo componente não permaneceu estável pois diminuíram os coeficientes de praticamente todas as variáveis concentrando o peso máximo em X_2, alterando a interpretação do mesmo. Já o terceiro componente que apresenta o máximo de mudança em θ, apresenta estabilidade nos coeficientes pois não altera significativamente as interpretações dos componentes quanto à participação das variáveis originais. O quarto componente que apresenta o menor ângulo entre as direções observadas faz uma inversão de importância entre X_5 e X_6, em oposição. Enquanto cresce o coeficiente de X_6, decresce o de X_5. Por outro lado, ao se considerar a redução da dimensionalidade pelo % de variação retida em

CP. Este foi, também, o que apresentou os menores valores de θ e também a maior estabilidade nos coeficientes. É possível, então, proceder a interpretação das relações entre as variáveis ao nível do primeiro Componente pela sua estabilidade.

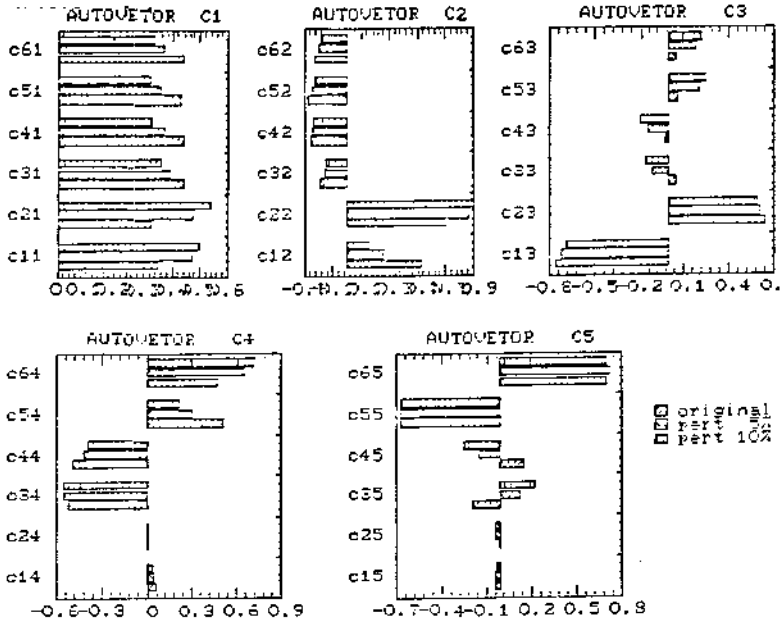
Tabela nº 4.1 Efeito nos autovetores e ângulo θ formado entre c_i, coeficiente original e $c_{(i)}$, coeficiente alterado.

COMPONENTE	VARIÂNCIA	COEFICIENTES						ÂNGULO
Y_i	λ_i	c_1	c_2	c_3	c_4	c_5	c_6	θ
Y_1	4.568	0.35	0.33	0.44	0.44	0.43	0.44	
pert. 5%	4.352	0.47	0.48	0.39	0.37	0.36	0.37	14
pert. 10%	4.159	0.50	0.54	0.36	0.33	0.32	0.34	19
Y_2	0.714	0.53	0.70	-0.19	-0.25	-0.28	-0.22	
pert. 5%	0.682	0.26	0.87	-0.16	-0.24	-0.24	-0.20	19
pert. 10%	0.656	0.15	0.90	-0.15	-0.23	-0.22	-0.18	26
Y_3	0.412	-0.76	0.64	0.05	-0.02	0.06	0.05	
pert. 5%	0.393	-0.72	0.61	-0.11	-0.15	0.21	0.18	16
pert. 10%	0.377	-0.69	0.59	-0.16	-0.20	0.25	0.22	23
Y_4	0.173	0.05	0.00	-0.52	-0.49	0.51	0.47	
pert. 5%	0.165	0.04	0.00	-0.55	-0.42	0.30	0.65	17
pert. 10%	0.156	0.03	0.00	-0.55	-0.39	0.21	0.71	23
Y_5	0.076	-0.04	-0.00	-0.19	0.15	-0.67	0.70	
pert. 5%	0.073	-0.03	-0.03	0.13	-0.15	-0.67	0.72	24
pert. 10%	0.068	-0.02	-0.04	0.22	-0.25	-0.64	0.70	33

Pode-se visualizar tanto a distância angular entre o coeficiente original e o perturbado, como observar graficamente os coeficientes, comparando-os. Desta forma, enriquece-se a análise pois aumenta o poder de comparação. É possível concluir que a ACP apresenta estabilidade quando o ângulo θ entre o autovetor original e os autovetores perturbados por alterações de tamanho máximo ε é mínimo. Neste caso, observando-se os gráficos comparativos dos autovetores verifica-se que as alterações são mínimas. Por outro lado, pode-se, também, concluir que os autovetores que forem máximamente perturbados e, ainda assim, não alteram significativamente o padrão de participação de cada variável no componente, podem permanecer auxiliando na interpretação do fenômeno.

Observe-se, nos gráficos abaixo relacionado com a visualização do ângulo θ. Os CP com θ máximo são Y_2 e Y_5

Gráfico nº 4.1 Comparação entre os autovetores após serem submetidos à alterações na variância com os valores originais



4.5 APLICAÇÃO EM DADOS GERADOS COM ESTRUTURA NORMAL.

Com a finalidade de comprovar a aplicabilidade dos resultados obtidos acima em (79) e (81), realizou-se estudos em amostras geradas a partir de um vetor de médias e de uma matriz de covariância, definidos em (34) e (35). Pretende-se aplicar ACP com as informações em diferentes situações, como é o caso de pequenas e grandes amostras. A questão principal é a da perturbação da variância para um ε máximo de (0,05; 0,10). Interessa avaliar quais os CP que se mantém estáveis frente a estas perturbações e, também, avaliar quais as mudanças máximas que ocorrem ao nível dos coeficientes nestes casos. Esta análise será confrontada com o poder de detectação do método de Análise de Sensibilidade dado por Krzanowski(1984).

Tabela nº 4.2 Variação dos coeficientes do 1º autovetor em função do crescimento da variância em 5 e 10%, segundo o tamanho da amostra e resultados da Análise de Sensibilidade

COMPONENTE Y_1	VARIÂNCIA λ_1	COEFICIENTES c_{11}	c_{21}	c_{31}	c_{41}	c_{51}	ÂNGULO θ
Y_1							
n = 20	211,354	-0,043	0,140	0,040	0,978	0,143	
5 %	280,334	-0,055	0,154	0,037	0,969	0,183	15
10 %	363,172	-0,062	0,165	0,035	0,961	0,210	21
n = 50	201,849	-0,050	-0,088	0,040	0,992	-0,065	
5 %	261,473	-0,033	-0,091	0,040	0,993	-0,051	15
10 %	333,529	-0,036	-0,094	0,039	0,993	-0,037	21
n = 200	168,424	-0,003	-0,106	-0,018	0,993	0,050	
5 %	223,303	-0,005	-0,112	-0,019	0,991	0,071	15
10 %	300,079	-0,002	-0,102	-0,023	0,990	0,095	21

Analisando os resultados de θ, é possível concluir que o 1º CP apresenta uma diferença entre c_1 e $c_{(1)}$, o autovetor com alterações na variância constante de 15 e 21, respectivamente, qualquer que seja o tamanho de n. Estes resultados indicam que o aumento do tamanho da amostra não se refletiu na estabilidade dos coeficientes.

Tabela nº 4.3 Variação dos coeficientes do 2º autovetor em função do crescimento da variância em 5 e 10%, segundo o tamanho da amostra e resultados da Análise de Sensibilidade

COMPONENTE Y_2	VARIÂNCIA λ_2	COEFICIENTES c_{12}	c_{22}	c_{32}	c_{42}	c_{52}	ÂNGULO θ
n = 20	67,712	0,021	-0,128	-0,250	-0,110	0,953	
5 %	90,226	-0,009	0,298	-0,245	-0,208	0,899	56
10 %	117,896	-0,004	0,324	-0,224	-0,242	0,887	64
n = 50	59,131	-0,008	-0,328	-0,075	0,035	0,941	
5 %	84,182	-0,005	-0,317	-0,064	0,022	0,946	20
10 %	115,304	-0,009	-0,311	-0,056	0,008	0,948	28
n = 200	49,347	0,023	-0,039	-0,092	-0,056	0,993	
5 %	71,136	0,030	-0,020	-0,092	-0,075	0,992	29
10 %	98,723	0,033	0,005	-0,089	-0,096	0,991	39

É surpreendente o valor assumido por θ na Análise de Sensibilidade. A desestabilização máxima ocorre quando se trata de amostra pequena, θ assume 56 e 64 para o caso de n=20, mas são

máximos também os valores de θ quando aumenta o tamanho da amostra. Este fato demonstra que o tamanho da amostra pode amenizar os efeitos das perdas ou acréscimos ocorridos na variância. A Análise de Sensibilidade indicou valores de 20 e 28 para n=50 e de 29 e 39 para n=200. Por outro lado, foi possível observar que este autovetor, juntamente com C_3, é o que mais apresenta problemas quando da geração de dados aleatórios. O problema , mais comum pode estar sendo indicado pelos coeficientes perturbados. Nestas amostras perturbadas ocorre uma mudança entre o os coeficientes de X_2 e X_5 em C_2 e em C_3.

Tabela nº 4.4 Variação dos coeficientes do 3º autovetor em função do crescimento da variância em 5 e 10%, segundo o tamanho da amostra e resultados da Análise de Sensibilidade

COMPONENTE Y_3	VARIÂNCIA λ_3	COEFICIENTES c_{13}	c_{23}	c_{33}	c_{43}	c_{53}	ÂNGULO θ
n = 20	66.151	-0.091	0.975	-0.091	-0.153	0.091	
5 %	84.035	-0.097	0.936	0.042	-0.095	-0.321	13
10 %	104.348	-0.101	0.925	0.062	-0.091	-0.348	19
n = 50	37.714	-0.006	0.939	-0.079	0.107	0.317	
5 %	50.593	0.043	0.942	-0.065	0.107	0.309	15
10 %	66.551	0.059	0.943	-0.055	0.105	0.306	20
n = 200	41.638	-0.010	0.992	-0.064	0.102	0.039	
5 %	55.846	-0.015	0.992	-0.055	0.110	0.024	15
10 %	74.206	-0.016	0.994	-0.046	0.101	0.001	21

Tabela nº 4.5 Variação dos coeficientes do 4º autovetor em função do crescimento da variância em 5 e 10%, segundo o tamanho da amostra e resultados da Análise de Sensibilidade

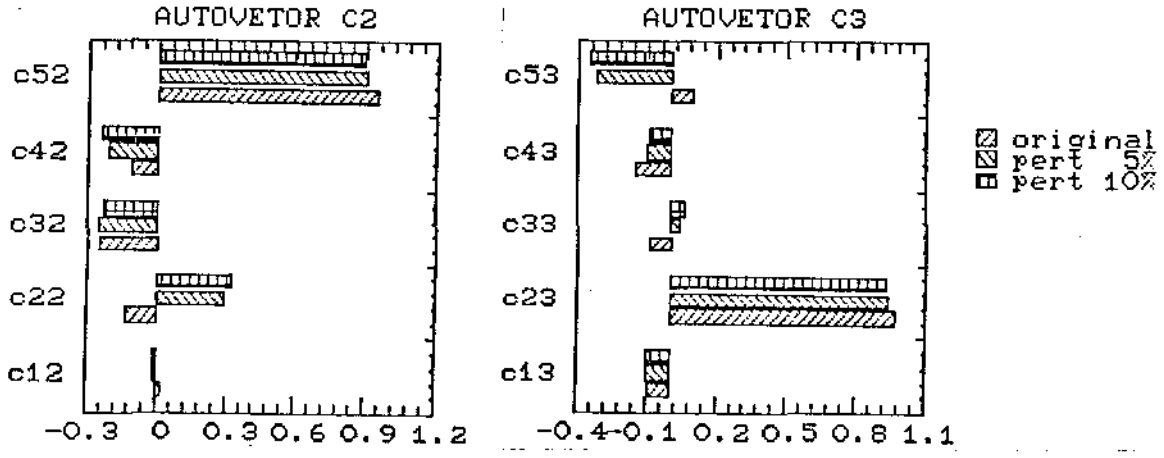
COMPONENTE Y_4	VARIÂNCIA λ_4	COEFICIENTES c_{14}	c_{24}	c_{34}	c_{44}	c_{54}	ÂNGULO θ
n = 20	7.753	-0.289	0.023	0.922	-0.089	0.241	
5 %	10.360	-0.369	-0.012	0.897	-0.095	0.222	14
10 %	13.733	-0.432	-0.038	0.872	-0.098	0.205	19
n = 50	9.445	-0.213	0.052	0.970	-0.039	0.095	
5 %	11.569	-0.021	0.046	0.995	-0.032	0.083	16
10 %	14.371	0.017	0.038	0.996	-0.032	0.072	22
n = 200	11.912	-0.053	0.057	0.992	0.019	0.097	
5 %	14.642	-0.056	0.050	0.992	0.018	0.096	15
10 %	17.991	-0.071	0.043	0.992	0.018	0.093	21

Tabela nº 4.6 Variação dos coeficientes do 5º autovetor em função do crescimento da variância em 5 e 10%, segundo o tamanho da amostra e resultados da Análise de Sensibilidade

COMPONENTE Y_s	VARIÂNCIA λ_s	COEFICIENTES c_{1s}	c_{2s}	c_{3s}	c_{4s}	c_{5s}	ÂNGULO θ
n = 20	1.382	0.952	0.110	0.279	0.005	0.067	
5 %	1.960	0.923	0.105	0.363	0.007	0.075	
10 %	2.758	0.894	0.099	0.430	0.008	0.079	
n = 50	3.794	0.997	-0.062	-0.013	0.030	-0.012	
5 %	6.279	0.998	-0.045	0.024	0.027	-0.008	
10 %	9.016	0.997	-0.062	-0.013	0.030	-0.012	
n = 200	4.043	0.998	0.013	0.055	0.006	-0.018	
5 %	6.201	0.998	0.018	0.058	0.010	-0.023	
10 %	9.100	0.997	0.018	0.073	0.009	-0.026	

Após a realização do cálculo de ACP, apresenta-se, no quadro acima, os principais resultados obtidos com os dados simulados. O cálculo de θ evidenciou alterações máximas nos coeficientes C_2 e C_3 quando n = 20. Porém o autovetor C_2 mostrou as diferenças máximas entre original e o valor perturbado, mesmo em amostras grandes como é o caso de n=50 e n=200. O 5º autovetor foi o mais estável neste caso.

Gráfico nº 4.2 Autovetores do 2º e do 3º CP após a perturbação da variância em 5 e 10%.



Tanto o primeiro CP como o quarto e o quinto apresentam

uma certa estabilidade. Y_1 e Y_5 podem ser observados como sendo os componentes que apresentam a maior estabilidade, seja qual for tamanho de ε, a perturbação da variância a que estão submetidos. Como se pode constatar, com os dados gerados seguindo o modelo normal, variou-se o tamanho da amostra e o tipo de perturbação. Independentemente do tipo de tratamento observado nos dados, os coeficientes de Y_1 apresentaram apenas ligeiras variações que não podem ser consideradas como significativas. Quanto ao CP que apresenta os maiores valores de θ, conclue-se que a desestabilização do C_2 faz com que haja um deslocamento do ponto de concentração em termos de coeficientes. Nas observações originais, constata-se que este componente tem como variável predominante X_5. Com $\varepsilon = 0.05$, qualquer que seja n, o coeficiente de X_2 cresce de importância alterando a interpretação do 2º CP. Em Y_2 e Y_3, quando $\varepsilon=0.10$, ocorre uma realocação de valores e os componentes passam a ser função de duas variáveis, X_2 e X_5. Ambas dividindo o poder de explicação dentro do componente. Observa-se que o comportamento dos componentes Y_2 e Y_3 fica desestabilizado quando a variância é perturbada em 5 ou 10 %. O mesmo pode ser observado no estudo Krzanowski com os componentes centrais, que concentram as variâncias de valor médio são os que apresentam alterações máximas nos coeficientes. Observa-se neste caso a validade da utilização do método de Análise de Sensibilidade proposto por Krzanowski.

Comparando-se estes resultados com os que foram apresentados por Krzanowski, reforça-se a tese de Hawkins(1974) que propõe a utilização do último CP para detectar erros nas observações. Estas conclusões devem-se ao fato de que o último componente não é sensível a desestabilização da variância. Não se observou mudanças neste componente independente, do tamanho da amostra. Pode-se também concluir que a análise de "clusters" será mais eficiente se forem analisados as representações gráficas dos dados de Y_1 e Y_p, primeiro e último componente. Chang(1983) propõe este procedimento quando se trata de separar duas normais multivariadas. A utilização do último CP apresenta grande eficiência quando se trata de erros que "incham" inapropriadamente a variância, como é o caso de retiradas de otimalidade na variância.

CAPÍTULO 5

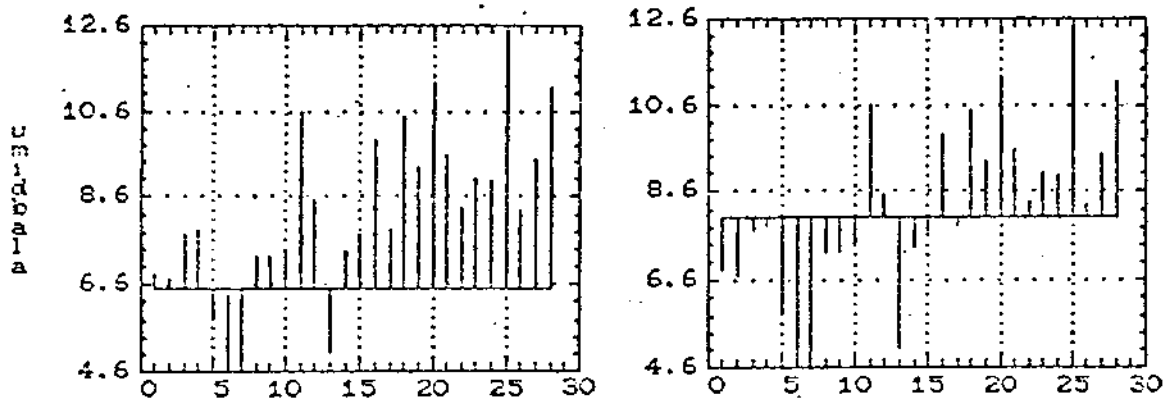
ANÁLISE DE SENSIBILIDADE UMA APLICAÇÃO

A modernidade traz consigo o aperfeiçoamento do computador pessoal permitindo a massificação no uso de técnicas estatísticas que, apesar de não envolverem matemáticas muito complexas, não se prestam a serem consumidas em larga escala. Este é

o caso do Controle Estatístico de Qualidade(CEQ) multivariado. Foi grandemente ativado durante a segunda guerra mundial que exigia altos níveis de qualidade nos produtos pois deles muitas vezes dependia o sucesso ou insucesso de uma missão. Foi, depois disso, um tanto desativado nos Estados Unidos, passando a ter maior desenvolvimento no Japão onde as teorias de Demming além da maior aceitação encontraram maior capacitação no suporte técnico para a sua divulgação e utilização. Recentemente reencontrou seu papel junto ao desenvolvimento industrial americano.

Neste trabalho pretende-se utilizar técnicas de CEQ como os gráficos de Controle de Qualidade e também da técnica de ACP acompanhada de uma Análise de Sensibilidade dos CP. O mesmo será aplicado em dados obtidos numa empresa de médio porte com processo de produção de balas comestíveis. Processo este que envolve diversas variáveis. Procura-se encontrar parâmetros para a redução da produção de artigos que não atendam à especificação da empresa, sendo que o fator principal de rejeição das balas é a umidade fora do padrão especificado para o produto. Quando esta é excessiva, o produto apresenta a primeira camada derretida sendo inutilizado para consumo imediato. Quando a umidade é insuficiente a dureza da massa faz com que a bala quebre durante o processo de produção. Como parte deste processo, pretende-se definir os fatores que mais interferem nos resultados e como mensurar o nível de participação de cada variável no vetor-resposta.

Gráfico nº 5.1 Gráfico da proporção de balas do tipo A306 rejeitadas com destaque das amostras que estão fora dos limites especificados.



Num primeiro momento, o tratamento estatístico obrigatoriamente deverá manter-se em torno das variáveis para as quais já existem instrumentos de mensuração na empresa com sistema de regulagem. Esta coleta de dados deu-se com a supervisão de estatísticos, buscando garantir a precisão e confiabilidade dos dados coletados. Inicialmente, dois aspectos serão considerados em relação às balas: fatores ambientais (externos), e fatores internos da massa. As variáveis que estão sendo controladas, atualmente, são umidade e temperatura do ambiente em que as balas estão sendo produzidas, denominados neste texto de fatores externos. Além destas duas, também os períodos em que as amostras foram coletadas. Como a questão principal é o estudo da umidade da bala, interessam neste estudo todos aqueles elementos que possam estar interferindo para explicar as variações que o fenômeno umidade apresenta. Um estudo descritivo das variáveis em estudo demonstrou que se pode considerar a normalidade dos dados, isto aliado ao fato de se tratar de amostras de populações infinitas, $N \longrightarrow \infty$.

As questões principais neste estudo referem-se ao processo produtivo: Está ou não sob controle? Como definir o padrão de regulagem ótima?

Com este estudo pretende-se realizar uma comparação entre os resultados do CEQ calculado a partir dos dados originais e os calculados após a aplicação de ACP e ACP-Regressão, no sentido de conseguir detectar como estes fatores devem ser ajustados simultaneamente. Na tentativa de manter o processo sob controle, pretende-se ao mesmo tempo reunir informações sobre a participação de cada variável no vetor resposta como um todo e mensurar o nível de participação destas quando o vetor resposta alcança índices aceitáveis ou ótimos. Isto é, pretende-se avaliar a participação das variáveis aleatórias originais de modo a obter um padrão de rotina multivariado para o ajustamento do processo de produção. Outro aspecto a ser desenvolvido no sentido de ajustar os fatores na produção são os "clusters". Através deles pretende-se evidenciar os grupos de vetores respostas segundo a participação dos fatores controlados. Ainda deve-se fazer uma Análise de Sensibilidade com a finalidade de observar a validade das interpretações dos coeficientes, c_i, encontrados após a transformação CP.

5.1 UMA BREVE REFERÊNCIA ÀS TÉCNICAS EMPREGADAS NOS DADOS AMOSTRAIS

COMPONENTES PRINCIPAIS

Os CP devem proporcionar um conjunto de variáveis não correlacionadas através da rotação ortogonal das variáveis aleatórias originais. Conforme propõe Anderson(1958), a transformação CP pode ser escrita da seguinte maneira, centrada na média:

Seja então:

$$Y_i = \Gamma' (X_i - \mu) , \quad \Gamma = (\gamma_1, \ldots, \gamma_p) \qquad (79)$$

autovetores associados aos autovalores da matriz de covariância Σ dos dados originais. Com esta rotação no espaço garante-se que Y_i, Y_j são

não correlacionados se i ≠ j, Pode-se recalcular o CEQ Multivariado, com mais eficiência, usando como base os CP. O primeiro CP contém o máximo de diversidade entre as observações pois foi calculado para maximizar a variância. Deste modo é possível utilizar os coeficientes tanto deste primeiro como dos outros componentes para captar o grau de participação de cada variável observada no vetor resposta. Esta rotação ortogonal favorece a análise de composição deste vetor resposta em termos de participação conjunta sem as distorções causadas pela correlação entre as variáveis de interesse. Com base na distribuição dos registros após a sua transformação e reescritos segundo a aplicação de CP num plano bivariado é possível observar a formação de "clusters". Com estas informações constata-se qual a melhor configuração em termos de ajuste tendo em vista determinados tipos de resposta.

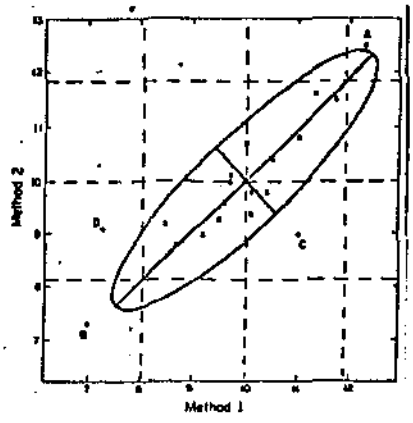
Propriedades da elipse:

1. O eixo maior é a linha de regressão ortogonal entre as duas variáveis originais. Esta linha é diferente da linha de regressão usual porque minimiza a soma de quadrados perpendiculares para a própria linha. A característica apresentada pode ser muito importante quando as variáveis são uma função do processo que está sendo mensurado.

2. A variação do eixo maior, no caso de um processo bivariado, representa a variação no processo e a do eixo menor representa a falta de ajuste entre as variáveis.

T^2- de Hottelling adaptado a Análise de Componentes Principais

Gráfico nº 5.2 Região de controle elipsóide proposta por Jackson após a aplicação de ACP.



Jackson(1980) propõe a construção de uma elipsóide que deve servir de região de controle no CEQ multidimensional. Sua aplicação foi desenvolvida para casos de pequenas amostras. Os pontos que caem no interior da região definida por este elipse indicam que o processo está sob controle. O "fora do controle" será indicado por pontos que estão fora da elipse. Este procedimento tem por base a estatística T^2 de Hotelling para a quantidade $X\ S^{-1}\ X'$. Esta estatística produz um gráfico com um limite de controle superior o qual indica quando o processo saiu de controle.

A estatística de Hotelling

$$T^2 = x'\Sigma^{-1}x, \quad x = X - \overline{X} \qquad (83)$$

está relacionado com a distribuição F do seguinte modo:

$$T^2_{p,n,\alpha} = (np)/(n-p+1)\ F_{p,n-p+1,\alpha} \qquad (84)$$

Este valor representa o limite superior do gráfico T^2. Este resultado pode ser aproximado para uma X^2_p.

Um problema que afeta a interpretação do modelo é, no entanto, a presença de correlação entre as variáveis observadas. No caso de correlação entre variáveis, um ponto fora do controle pode ser altamente pouco provável quando as variáveis estão dentro dos limites, mas quando ocorrer que uma observação tem valor mais alto que o esperado e quando outra tem valor menor que o esperado. Esta limitação pode ser superada empregando-se ACP que faz uma rotação ortogonal no espaço de respostas satisfazendo a restrição de que

$c_i' c_j = 0$, quando $i \neq j$. Esta restrição garante $Cov(Y_i, Y_j) = 0$, $i \neq j$. Portanto se os CP são avaliáveis, então o cálculo do T^2 pode ser feito da seguinte maneira:

$$T^2 = y'y \ , \qquad (85)$$

onde T^2 passa a ser justamente a soma dos quadrados dos CP. Esta estatística multivariada permite visualizar quando o vetor-resposta está "fora do controle" como resposta de um processo multivariado.

Análise de Sensibilidade

Realizar uma Análise de Sensibilidade implica em avaliar a estabilidade dos CP quando ocorre uma perturbação de no máximo tamanho ε na variância. Procurando as alterações máximas, pode-se definir o conjunto de variáveis cujo coeficiente pode ser devidamente utilizado para interpretar as relações das variáveis com o vetor-resposta. Por outro lado, é possível detectar aquelas que podem ser ignoradas por interferirem no processo de modo insignificante e que, por falta de estabilidade, comprometem as interpretações efetuadas a partir dos autovetores.

Os CP são analisados quanto à estabilidade, calculando-se o ângulo θ formado entre o coeficiente original e o coeficiente perturbado. O cálculo deste ângulo, conforme já foi detalhado, vai depender dos autovalores da função critério V, quando esta passa por um máximo $V(c^*)$.

Seja

$$\cos \theta = \left[1 + \frac{\varepsilon}{\lambda_j - \lambda_{j+1}} \right]^{-1/2} \qquad (86)$$

Onde, $\varepsilon = k \lambda_j$, $k = (0.05; 0.10)$

5.2 FATORES AMBIENTAIS E INTERNOS EM ESTUDO

A produção aqui avaliada segundo aspectos internos do produto e elementos externos ambientais que interferem no mesmo auxiliando a definir o resultado final, está representada por observações obtidas por amostragem. A amostra consta da coleta aleatória de cinco balas em períodos selecionados ao acaso. Nesta ocasião foram observados fatores internos e externos. Os fatores externos foram anotados pelos operários previamente educados para realizar o procedimento com o rigor necessário. Após o recolhimento a amostra foi encaminhada ao laboratório químico da empresa onde foram mensurados diversos aspectos do produto.

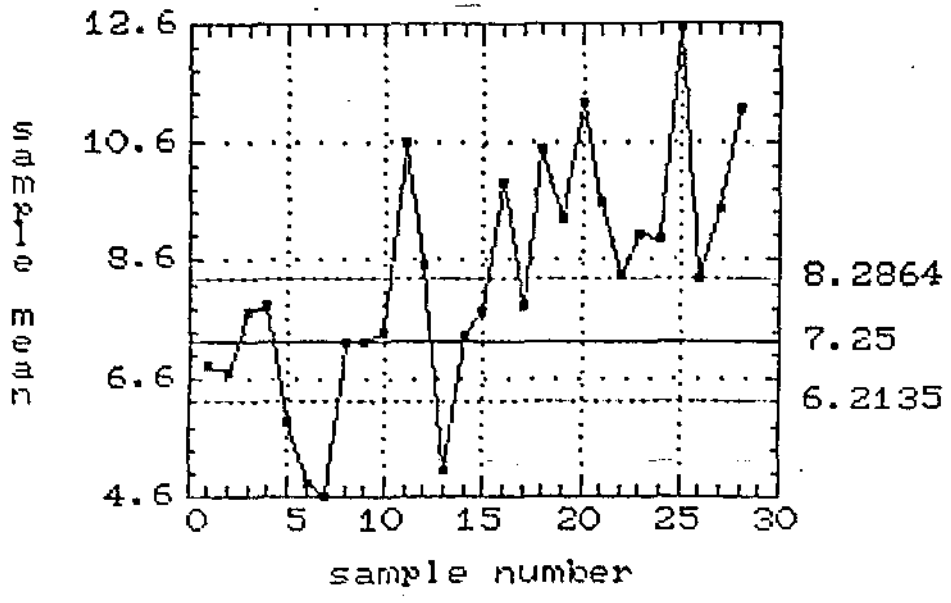
FATORES AMBIENTAIS
X_1: Hora
X_2: Temperatura do ambiente
X_3: Umidade do ambiente
Y : Umidade da bala

FATORES INTERNOS (LABORATORIAIS)
X_4: Grau brix (doçura)
X_5: PH (acidez)
X_6: Umidade da massa

A umidade da bala é a resposta que está sendo controlada. Observou-se uma variação de 4,6% a 12,6%, sendo que os limites padrões de controle para a umidade da bala vão de 6,2% a 8,3%. Das amostras realizadas mais unidades estiveram acima do padrão(50%) que abaixo do padrão(14%). Ainda assim qualquer dos extremos é prejudicial provocando rejeição do produto. Como se trata de balas comestíveis cujo teor de açúcar é alto, a umidade insuficiente provoca rigidez ou seja excessiva dureza. A rejeição vai ocorrer quase na etapa final do processo de produção. Quando o produto vai para o estampamento nessas condições quebra com

facilidade provocando resíduos e rejeição total com o envio então da bala para reaproveitamento em forma de recheio de outras balas. Mesmo contornável, porque existe o reaproveitamento das sobras, implica em alterações nos custos. Ainda em função do alto teor de açúcar nas balas quando a umidade é excessiva as balas irão "melar". Isto é, vai derreter a camada externa da bala a partir de um certo tempo de armazenamento(em geral 4 a 6 meses) provocando devoluções pelos clientes. Este tipo de rejeição é altamente prejudicial para a empresa pois além de fugir ao seu controle, provoca perdas de produto e de clientes. Estes dados podem ser observados no gráfico de controle apresentado abaixo para a umidade da bala.

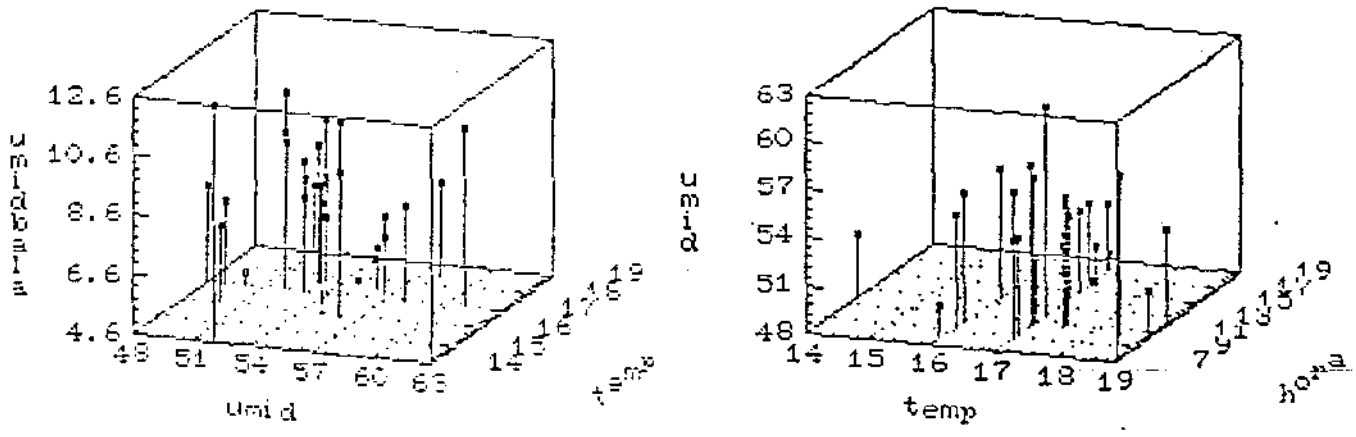
Gráfico nº 5.3 Controle da umidade da bala - Gráfico das médias



Ainda na questão dos fatores ambientais constata-se que duas das amostras: a de nº 7 e a de nº 25 apresentam valores extremos e inversos quando se observa a umidade da bala e a temperatura do ambiente. Na amostra nº 7 ocorreu temperatura máxima 19 graus e na amostra nº 25 ocorreu a temperatura máxima. Esta relação inversa confirma-se quando se aplica um estudo de correlação entre as duas,

r_{xy} = -0.586685. A equação de regressão definida é Y = 28.59 -1.198X. Porém este aspecto privilegia uma explicação univariada. Outro aspecto observado é que existe uma relação direta entre a umidade da bala e a umidade do ambiente. Umidade crescente correspondeu a umidade crescente da bala e umidade decrescente correspondeu a umidade decrescente da bala. Quanto ao período do dia em que o produto foi produzido observou-se que em geral quanto mais cedo maior a umidade da bala que decresce ao longo do período, porém esta relação é muito oscilante demonstrando que existem outros fatores que influenciam na variação desestabilizando a relação.

Gráfico nº 5.4 Aspectos tri-dimensionais das amostras observadas segundo os valores dos fatores ambientais



O modelo ajustado aos dados originais apresentam a seguinte equação de regressão

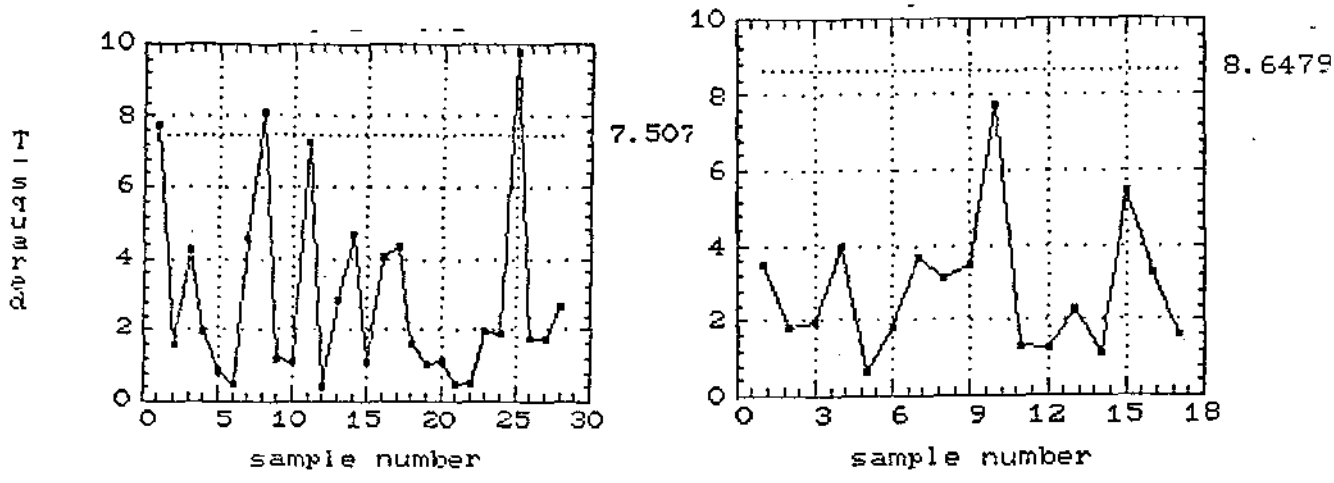
$$Y = 21.79 + 0.0785X_1 - 1.2237X_2 + 0.1167X_3$$

Estes valores ocorrem com um nível de probabilidade de 53% no caso do período do dia, 0.12% no caso da temperatura ambiente e 35% no caso da umidade ambiente. Estes resultados evidenciam a

relação observada entre a temperatura e a umidade da bala.

Quanto ao T^2 de Hotelling, gráfico de controle multivariado a um nível de α=0.10 considerou o sistema fora de controle em tres ocasiões: na 1º, 8º e 25º amostra. Em termos da variável umidade da bala que é a resposta de interesse do estudo, o gráfico T^2 coincide com o processo efetivamente fora do controle apenas na amostra nº 25, que é uma situação extrema.

Gráfico nº 5.5 T^2 de Hotelling para Fatores Ambientais e Fatores Laboratoriais, α=0.10.

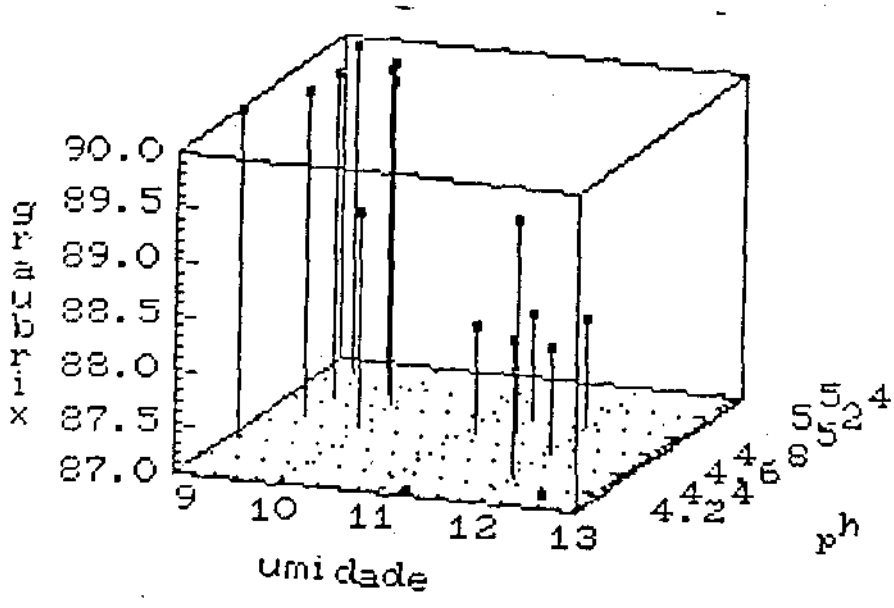


A estatística multivariada T^2 de Hotteling evidencia, que apesar de algumas amostras apresentarem valores um tanto fora dos padrões em termos da produção, nenhum ponto pode ser considerado fora de controle a um nível de 10%. A análise laboratorial da amostra de balas apresentou resultados que na sua íntegra em termos de análise não são muito diferentes dos obtidos com os fatores ambientais, porém menos extremos.

Observando o gráfico abaixo é possível distinguir em termos de fatores laboratoriais(internos) que dois grupos de

respostas podem ser definidos: o primeiro grupo apresenta baixa umidade e alto grau brix e ph e o segundo grupo apresenta valores com alta umidade , baixo ph e médio grau brix. Estes dois grupos reforçam a idéia de que a doçura da bala apresenta uma relação inversa com a umidade da massa.

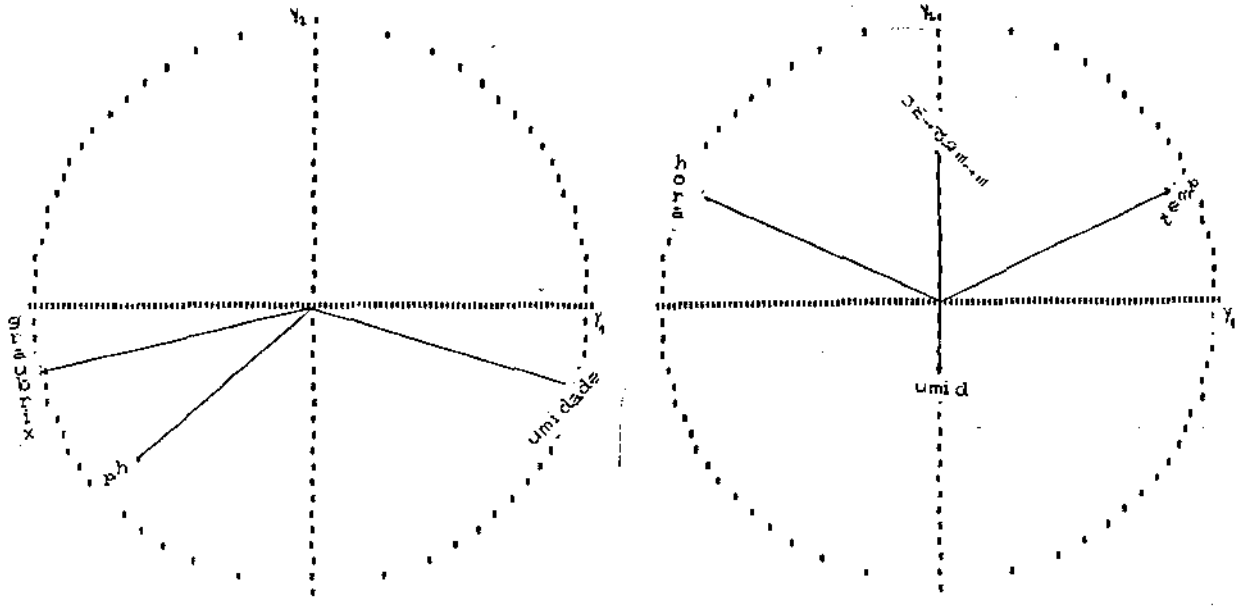
Gráfico nº 5.6 Distribuição dos pontos amostrais segundo os fatores internos, laboratoriais.



Constata-se que a umidade da massa está em oposição ao grau brix e ao ph da massa, porém estes dois últimos elementos estão definidos no mesmo subespaço. Isto quer dizer que em termos de informação contribuem do mesmo modo para a análise da umidade da massa. Um terceiro elemento foi considerado: a quantidade de massa produzida, porém observa-se que este montante apresenta uma relação direta fraca(0,21) com o ph apenas, insignificante e inversa com o grau brix(-0,11) e a umidade da massa(-0,10).Investigada apenas, esta informação foi abandonada no decorrer do estudo. Constata-se que a umidade está inversamente relacionada com o ph(-0,51), inversamente relacionada com intensidade forte com o grau brix(-0,90) deixando a idéia de que quanto maior a doçura da massa menor a umidade da massa.

O grau brix também é determinante no caso do ph(0,70) porém de modo direto. Deste modo é possível relacionar o controle da umidade da massa através do controle do grau brix e do ph.

Gráfico nº 5.7 Correlações entre as variáveis nos dois aspectos enfocados: Fatores Ambientais e Fatores Laboratoriais



Como se pode perceber fica difícil definir o grau de participação que deve ser adotado para cada variável no sentido de obter o processo dentro do controle quando se trata de avaliar as variáveis originais. Observe-se, agora, os resultados da aplicação de ACP aplicado aos dois conjuntos de informações:

Tabela nº 5.1 FATORES AMBIENTAIS

VARIÁVEL \ CP	%	λ_i	c_1	c_2	c_3	θ 5%	θ 10%
HORA	70	12,8345	-0,70378	0,71002	-0,02349	16	21
TEMPERATURA	25	4,5374	0,71036	0,70295	-0,03554	14	19
UMIDADE	5	0,9174	0,08715	0,04171	0,99909		

Tabela nº 5.2 FATORES AMBIENTAIS (Valores Padronizados)

VARIÁVEL \ CP	%	λ_i	c_1	c_2	c_3	θ 5%	θ 10%
HORA	49	1.4809	0.70209	0.13117	0.69989	21	29
TEMPERATURA	34	1.0031	-0.70805	0.02413	0.70575	18	24
UMIDADE	17	0.5160	0.07569	0.99107	-0.10982		

O primeiro CP pode ser interpretado como um contraste entre o fator temperatura e o fator período do dia. A relação de concorrência dos dois é de oposição. Este resultado define a necessidade de evitar um controle fixo da temperatura ao longo dia, como está instituído na empresa. Evidencia a importância do ajuste da temperatura sistematicamente. As condições ambientais em que as balas são produzidas formam um micro ambiente que deve favorecer a produção. Este micro ambiente interage com o macro ambiente constituído das condições atmosféricas externas. Desta forma um controle rígido e único da temperatura pode determinar perdas ao nível da produção. O segundo CP novamente apresenta uma contribuição destas duas variáveis como dominantes. O terceiro CP é o fator umidade do ambiente, uma vez que a contribuição desta variável no terceiro componente é de 99,9%. Neste caso, tendo sido verificado que o primeiro e o segundo componentes acumulam um percentual de variação explicada de 95%, poder-se-ia tratar o processo como bivariado reduzindo o número de componentes de 3 para 2. Como se observa, no entanto, caso assim fosse procedido perder-se-ia toda a informação sobre a umidade do ambiente. Uma padronização dos dados apresenta, uma inversão entre os coeficientes do segundo e terceiro componentes. A padronização assegura identidade em termos de unidade de valor das variáveis originais o que deve explicar esta alteração nos resultados. Pelo que pode ser observado a "invariância escalar" da ACP pode ser contornada pela padronização, e neste caso, é possível reduzir a dimensionalidade para r = 2. O 1º CP pode ser visto como o fator temperatura numa relação inversa com a hora e o 2º CP é o fator umidade. O fator umidade foi o que apresentou maior relação com a umidade da bala(r_{xy} = -0,587). Quanto à umidade é possível afirmar que as alterações que apresenta estão em parte sendo explicadas pela hora(r_{xy} = -0,478).

A Análise de Sensibilidade da função critério demonstra uma diversidade muito pequena em termos de observar a sensibilidade frente à perturbações máximas de $\varepsilon = 0,05$ e $0,10$. O componente E-maximamente perturbado, e causando uma perturbação máxima ε na função critério é o primeiro CP. Mas é preciso observar que esta alteração não difere significativamente da observada no segundo componente. Seja, então, os valores de $\theta = 16; 21$ para o o primeiro e $14; 19$ para o segundo, ligeiramente menor. O que se constata neste caso é de que os resultados padronizados, frente aos resultados centrados na média apresentam-se mais sensíveis, $\theta = 21 ; 29$ para o primeiro componente e $18; 24$ para o segundo.

Tabela nº 5.3 FATORES INTERNOS

VARIÁVEL \ CP	%	λ_1	c_1	c_2	c_3	θ 5%	θ 10%
PH	93	2.3411	-0.1170	-0.3819	0.9168	13	18
GRAU BRIX	6	0.1437	0.7054	-0.6818	-0.1940	14	20
UMIDADE MASSA	1	0.0327	-0.6991	-0.6240	-0.3492		

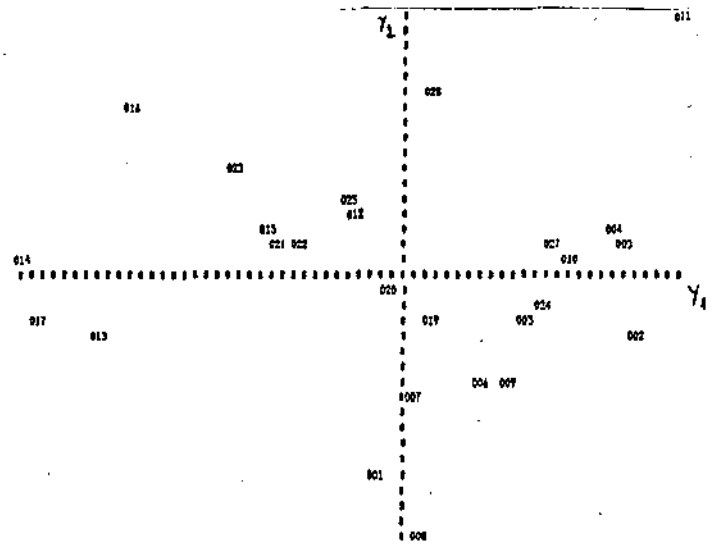
Tabela nº 5.4 FATORES INTERNOS (Valores Padronizados)

VARIÁVEL / CP	%	λ_1	c_1	c_2	c_3	θ 5%	θ 10%
PH	80	2.4118	0.5182	0.8234	0.2311	14	20
GRAU BRIX	17	0.5138	0.6244	-0.1796	-0.7601	13	19
UMIDADE MASSA	3	0.0744	-0.5844	0.5382	-0.6073		

Os fatores internos examinados a nível de laboratório demonstram que o primeiro CP é um contraste entre o grau brix: nível de doçura da massa e correspondente a massa compacta da mesma e a umidade da massa - que é a medida da parte líquida da massa. O segundo componente ainda dá maiores pesos à estas duas variáveis, o terceiro é um componente que discrimina os dados pela participação da variável "ph", com um peso pequeno para a variável umidade numa relação de contraste com o ph. O mesmo tipo de comportamento observado nos fatores ambientais pode ser constatado quando os resultados dos fatores internos observados são padronizados.

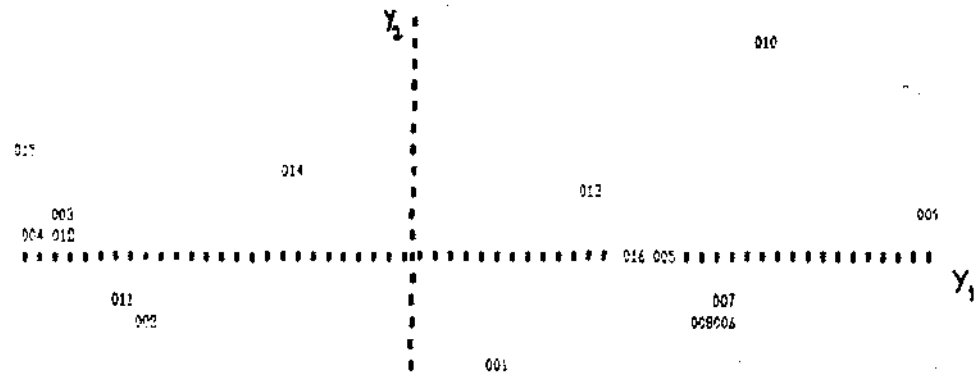
Um resultado interessante é a aplicação da Análise de Sensibilidade que apresenta um θ maior nos fatores ambientais , indicando maior sensibilidade de seus coeficientes.

Gráfico nº 5.8 Plano Y_1 e Y_2 - Fatores Ambientais



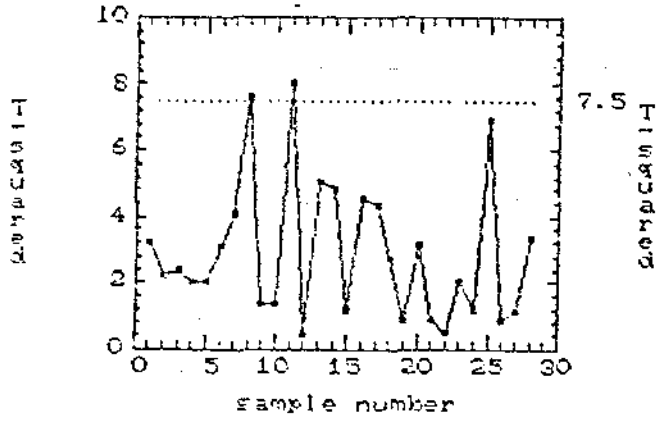
O % de variação explicada pelos dois primeiros CP(95% e 99% respectivamente), indica que o estudo pode ser reduzido para um caso bivariado. Com a redução da dimensionalidade, constata-se através do gráfico 5.7, a formação de grupos com os quais é possível estabelecer os parâmetros desejados. Observa-se que apenas uma das amostras apresenta valor altamente improvável ou distorcido por problemas técnicos. Relacionando este resultado com a Análise de Sensibilidade constata-se que esta detectou alterações nos coeficientes. Já no 2º conjunto de dados, sob o ponto de vista laboratorial não se observa nenhum valor que não possa ser agrupado junto com os demais. Esta situação pode ser confirmada pela estatística T^2 de Hotelling, onde nenhuma amostra multivariada se encontra acima do limite de contrôle.

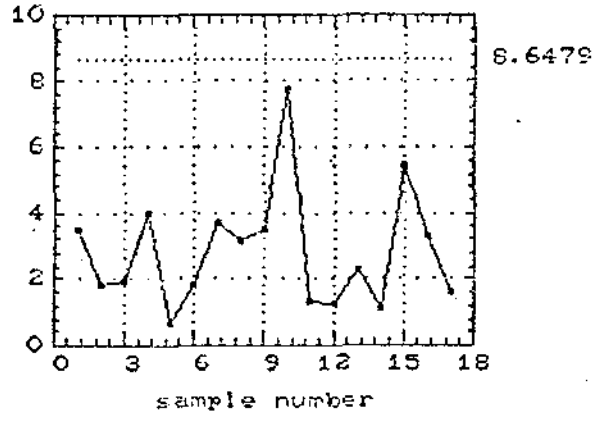
Gráfico nº 5.9 Plano Y_1 e Y_2 - Fatores Laboratoriais



Dois grupos podem ser observados no caso dos Fatores externos após ter sido reduzida a dimensionalidade de p=3 para r=2, e reescritos em função da aplicação de ACP. Confirmando o que pode ser observado pelas regressões.

Gráfico nº 5.10 T^2 de Hotteling definida como $T2 = yy'$, à um nível α=0.10 - Fatores Ambientais e Laboratoriais.





A aplicação de gráficos de controle à variável umidade da bala apresentou muitas amostras com uma umidade fora dos limites aceitáveis de umidade. O gráfico de T^2 de Hottelling aplicado como usualmente é utilizado, $c'\Sigma^{-1} c$, apresentou três amostras fora dos padrões. Após aplicação de ACP constatou-se que nos fatores ambientais apenas duas amostras continuam indicando o processo de produção fora de contrôle. Verifica-se, também que no caso dos fatores laboratoriais o gráfico da estatística T^2, com ou sem a transformação CP, apresenta os mesmos resultados com todos os pontos sob controle. Isto se deve aos valores observados pela não ocorrência de valores isolados. Estes resultados podem evitar que erros sejam cometidos durante a emissão de pareceres sobre os mesmos. Assim, a aplicação de ACP evita que o erro de coleta, de mensuração tenha efeito multiplicativo.

CONCLUSÃO

Exigir um certo rigor no tratamento de ACP torna-se um fato, pois baseado em suas ótimas propriedades, vem sendo ampliadas suas aplicações nas mais diversas áreas. Os resultados obtidos podem estar refletindo mudanças nas condições ideais de aplicação do

método, alterações estas que modificam a estrutura de variância dos dados.

Segundo Jackson(1959), ACP pode ser utilizado em CEQ, uma vez que a estatística, definida a partir dos coeficientes dos CP, apresenta uma distribuição T^2 com a vantagem que a nova base garante componentes não-correlacionados. A desvantagem de sua utilização está nas restrições impostas para a aplicação de T^2: os dados devem ter uma distribuição normal e a estrutura de variância conhecida.

Outra aplicação de ACP está relacionada com Regressão Multivariada. Segundo Massy(1965), isto permite tratar com dados que, em seus valores originais, apresentam multicolinearidade. As propriedades dos CP permitem que seja encontrada uma solução, o que não pode ser realizado com a regressão clássica pela impossibilidade de inverter a matriz. Pode ocorrer colinearidade entre as variáveis independentes e algumas das variáveis dependentes. Neste caso ACP oportuniza a deleção das variáveis redundantes. A utilização de ACP-R traz uma desvantagem justamente no redimensionamento do nº de variáveis. A deleção das variáveis redundantes implica em perda de informação e isto afeta o cálculo da regressão. Mesmo assim, o fato de superar os problemas de colinearidade o coloca em vantagem em relação à forma clássica. Outro problema pode estar na dificuldade de interpretar os coeficientes restringindo sua aplicabilidade pois estes representam também os coeficientes de regressão - β_i -. Porém com a aplicação dos CP-R, os perfís dos betas poderão ser avaliados através de uma Análise de Sensibilidade. A Análise de Sensibilidade dos CP-R pode ser realizada através de retiradas no valor esperado da forma alisada dos perfís dos betas.

A utilização de ACP na classificação de indivíduos a partir de um conjunto multivariado de informações foi tratado também por Mardia(1979). Contrastar o 1º e o 2º CP permite observar como os indivíduos se agrupam por influência do processo multidimensional. Dessa maneira, pode-se classificar os indivíduos em "clusters" pelo valor dos coeficientes. No entanto, nem sempre os CP que melhor

dicriminam os grupos são o 1º e o 2º CP. O maior contraste pode ser apresentado pela dispersão dos pontos observados a partir do 1º e do último CP. Chang(1983) avaliou a contribuição de ACP, e em particular do uso do último CP para separar a mistura de duas normais reforçando a idéia de que os dados apresentam maior contraste no diagrama de dispersão entre o 1º e o último CP. Uma outra análise está relacionada com os resíduos apresentados. Resíduos estes que podem ter origem no escalonamento das variáveis. Qualquer mudança de escala interfere nos resultados da aplicação do método. Diferentes definições dos autovetores permitem normalizá-los superando a dificuldade inicial. Assim são definidas formas de verificar a adequacidade do modelo, analisando-se o quadrado médio dos resíduos como uma medida geral do ajuste de uma observação com o modelo adotado. Com este procedimento é possível constatar a presença ou não de "outliers".

A difusão do emprego de ACP exige a realização de uma Análise de Sensibilidade que verifique a estabilidade dos coeficientes, c_i. Três principais perturbações devem ser inicialmente analisadas: presença de "outliers", uso de arredondamentos e perdas de otimalidade na variância.

Vários são os testes e técnicas relacionadas com a avaliação da estabilidade dos CP frente à ocorrência de "outliers". Estes são realizados, em geral, através de uma análise de resíduos, com dois motivos principais: constatar a presença de "outliers" e verificar a adequacidade do modelo ACP àquele conjunto de dados. Jackson e Mudholkar(1979) propõem a realização de uma análise dos resíduos através da estatística Q, tal que

$$Q = (X - \hat{X})'(x - \hat{X})$$

onde

$$\hat{X} = C_q \Lambda_q^{1/2} Y$$

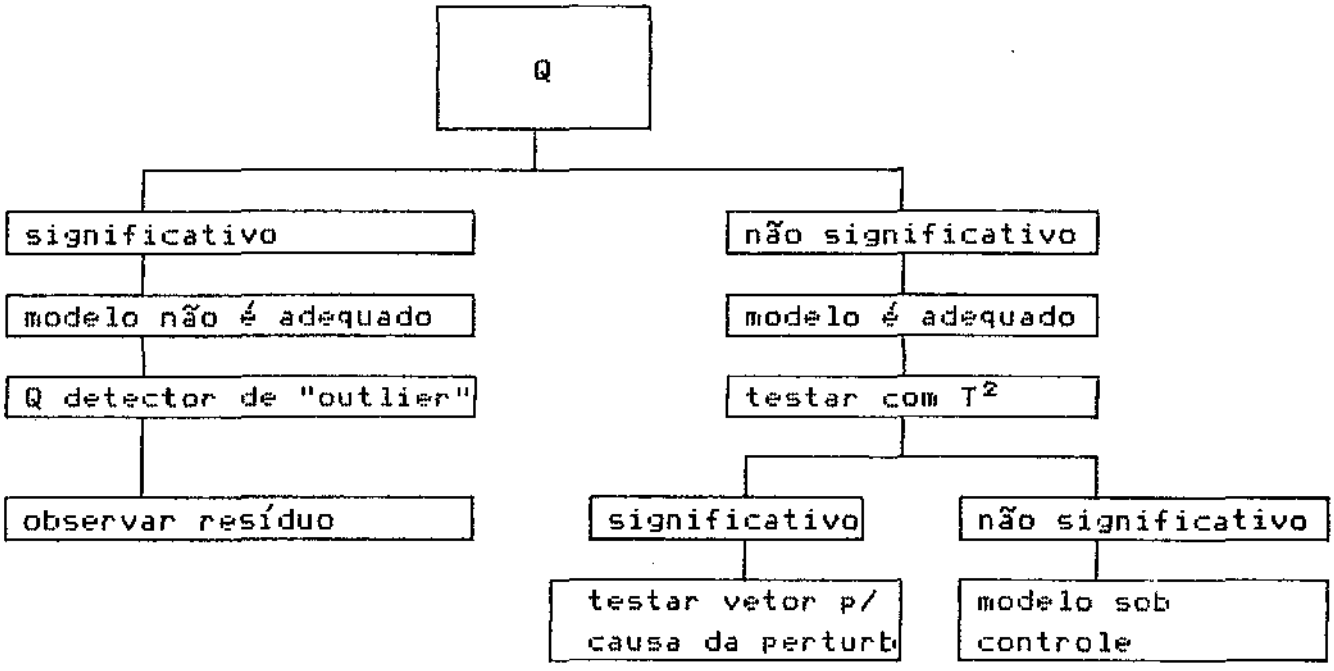
Uma regra de decisão para o caso é

$Q < \varepsilon$ modelo é adequado
$Q \geq \varepsilon$ modelo não é adequado

Quando existe suspeita de que haja a presença de "outliers", Jackson e Morris(1957) propuseram uma análise de resíduos através de

$$\text{Res. SS} = \frac{(p - k)\, Q}{\theta_1}$$

Esta análise de resíduos pode não ser eficiente para detectar "outliers" no caso de redução da dimensionalidade. Se o modelo adotado não contiver as p variáveis originais então o Res SS tende a apresentar os limites de Q, (20), alterados confundindo os efeitos da presença de "outliers". A expressão de θ_1 é dada em (21). O uso da estatística Q, (17), para um vetor individual pode ser resumida pelo quadro abaixo, acerca do resultado de Q:



Além destas, outras estatísticas alternativas podem ser definidas para o caso de verificação da sensibilidade dos CP na ocorrência de "outliers". Porém grande parte dos estudos realizados remetem à verificação de que nem sempre é o CP associado ao maior autovalor que contém a maior quantidade de informação. É possível demonstrar que o sub-conjunto de CP que apresentar a maior distância entre os pontos é o melhor pelo alto poder de discriminação que apresenta. Para Chang(1983) a seleção dos maiores autovalores na ACP não é essencial e atualmente não mais se justifica. Na questão da Sensibilidade dos CP verifica-se que o maior poder de discriminação está no estudo da dispersão entre os pontos no gráfico Y_1 e Y_p. Variâncias máximas e variâncias mínimas podem ser desejáveis sob o ponto de vista de que a separação dos "clusters" exige homogeneidade no grupo, mínima variância interna, e heterogeneidade entre os grupos, variância máxima entre os grupos.

Para detectar a falta de ajuste de uma observação individual, Rao(1964) propôs o estudo do quadrado da soma dos comprimentos das projeções de uma observação nas últimas coordenadas dos CP, avaliando a magnitude dos afastamentos.

$$d^2_j = (Yi - \overline{Y})' (Yi - \overline{Y}) - \sum_{i=1}^{p-q} [l_i'(Y_i - \overline{Y})]^2$$

A regra de decisão estabelecida é:

$$d^2_j \Longrightarrow \begin{cases} \text{grandes afastamentos} \longrightarrow \text{Ajuste pobre} \\ \text{pequenos afastamentos} \longrightarrow \text{Bom ajuste} \end{cases}$$

Outras alternativas podem ser o uso de gráficos: gráfico normal, gráfico de dispersão Y_1 x Y_p.

É possível observar que a falta de estabilidade dos CP na ocorrência de "outliers" cresce na medida em que o tamanho da amostra decresce. O tamanho da amostra desestabiliza a variância e pode ter pouca representatividade em relação ao tipo de distribuição da população. Para a aplicação de um modelo linear existe uma restrição baseada na normalidade dos dados. Considerando uma população com $N \longrightarrow \infty$, existe a suposição, questionável, da existência de normalidade. No caso de pequenas amostras a garantia de normalidade nem sempre se sustenta modificando as condições de aplicação de ACP. Neste caso dever-se-ia fazer primeiro uma investigação de normalidade, talvez aplicando um gráfico(difícil no caso multivariado) ou, talvez testar quanto ao quarto momento. Por outro lado seria possível definir um fator de correção para pequenas amostras que garantissem a validade dos resultados de ACP possibilitando o uso dos coeficientes na análise dos dados.

Quando os valores são arredondados é possível utilizar o teorema dado por Bibby(1980) que estabelece os limites da variação de Δ_k, a perturbação da variância associada a um vetor arbitrário h, que representa seu desvio em relação a λ_k, a variância de Y_k.

$$r^2 (\lambda_p - \lambda_k) \leq \Delta_k \leq r^2 (\lambda_1 - \lambda_k)$$

Na expressão acima são deduzidos os limites de variação de δ_k. No entanto este procedimento não permite que se verifique qual é o autovetor mais significativamente perturbado. Este conhecimento é importante porque a análise será realizada a partir dos coeficientes, c_i. Verificou-se que a estabilidade foi mantida no caso de grandes amostras quando $n \longrightarrow \infty$. Por outro lado houve uma coincidência entre os autovetores perturbados em casos de pequenas amostras na presença de "outliers" e na retirada de otimalidade na variância. Este fato pode estar indicando a necessidade de investigar se a Análise de Sensibilidade definida por Krzanowski(1984) poderia também detectar a sensibilidade no caso de arredondamentos. Esta relação precisa ser estudada com mais profundidade.

Esta estatística de Krzanowski(1984), baseada na suposição de perdas de otimalidade na variância permite verificar o grau de sensibilidade apresentado pelos CP. A sensibilidade vai ser avaliada pela magnitude do ângulo θ_j formado entre os coeficientes observados, c_j, e os coeficientes alterados, $c_{(j)}$.

$$\cos \theta_j = \left[1 + \varepsilon / (\lambda_j - \lambda_{j+1}) \right]^{-1/2} , \quad \varepsilon = k \lambda_j$$

De uma certa forma a Análise de Sensibilidade proposta atende às especificações quando é possível garantir uma estrutura de variância conhecida e a normalidade dos dados. No entanto inúmeros estudos envolvem relações não lineares e, também, relações não paramétricas e neste caso resta saber se a estatística definida por Krzanowski satisfaz igualmente as necessidades de análise.

Jeffers(1967) aponta para as restrições no uso generalizado de ACP, por se tratar de um modelo linear. No caso não-linear é necessário reformular as condições de aplicação ou, então, buscar outras formas de tratamento. Yohai(1985) considerou uma função, F_i, mais um erro como a expressão da transformação CP

$$X_i = g_{ii}(F_i) + \varepsilon_{ii}$$

onde ε_{ii} assume pequeno valor e g_{ii} pertence à uma classe de funções G_i. Estas funções podem ser não decrescentes ou não crescentes. A solução mais simples é a de escolher a classe de funções lineares não decrescentes:

$$G_i = [\ g: g(f) = af + b,\ a \geq 0\ ]$$

Restringir G_i à classe de funções lineares têm o inconveniente de que para garantir ε_{ii} tão pequeno quanto necessário é preciso que a relação entre quaisquer pares de variáveis originais seja linear. No caso em que é esperada uma função côncava, o modelo

linear não satisfaz. A melhor escolha revelou-se como sendo a classe dos segmentos de parábolas quadráticas não decrescentes. Estas funções g_i são do tipo

$$g(f) = af^2 + bf + c$$

com

$$\begin{array}{lllll} a > 0 & e & f \geq -b/2a & ou \\ a > 0 & e & f \leq -b/2a & ou \\ a = 0 & e & b \geq 0 & \end{array}$$

Uma aplicação realizada por Yohai(1985), envolvendo o estudo de 5 indicadores de desenvolvimento em países da América Latina e Caribe apresentou uma relação côncava entre as variáveis X_1 e X_3. O uso do modelo linear foi comparado com o uso do modelo quadrático não apresentando diferenças significativas nos resultados de Y_1, porém permitiu uma avaliação mais precisa sobre o montante de perda de informação, garantindo, também, resíduos menores. O 2º CP mostrou grandes diferenças nos resultados da aplicação de um ou outro modelo. De um modo em geral, observa-se que o modelo não linear consegue melhores coeficientes de explicação das variáveis X_{is}.

Outras questões remetem à um aprofundamento na definição de estatísticas que consigam realizar uma Análise de sensibilidade quando ocorrem mudanças na estrutura de variância como efeito de alterações na população. Uma situação em que o tempo varia entre uma aplicação e outra. Variação esta que faz parte do delineamento experimental pois se poderia estar interessado na construção de índices. Como comparar os índices construídos com intervalos de tempo ou com variação do espaço de referência, que seria o caso de comparação de regiões? Neste caso pretende-se, observar se os coeficientes cresceram ou decresceram como função da mobilidade do tempo e do espaço.

BIBLIOGRAFIA

1 - ANDERSON,TW. An introduction to multivariate statistical analysis. New York, Wiley, 1958.

2 - ANDERSON,TW. Asymptotic theory for principal components analysis. Annuary Mathematical and Statistical

34,122-148,1983.

3 - BAI,Z.D. SILVERSTEIN, Jack W. YIN, Y.Q. A note on the largest eigenvalue of a large dimensional sample covariance matrix. Journal of Multivariate Analysis, 26, 166-168,1988.

4 - BEN,MG. ALONSO,O. YOHAI,VJ. Componentes Principales usando un metodo suavizado univariado. Buenos Aires , Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

5 - BENASSENI J. Une contribution à l'etude de la stabilité en analyse factoriele. These 3º cycle. USTL Mont-Pellier, 1984.

6 - BENASSENI J. Influence des poids des Unités Statistiques sur les valeurs propres en Analyse en Composantes principales . Revue d'Statistique Appliquée, vol XXXIII nº4, 41-45, 1985.

7 - BYBBY, J.M. Some effects of rounding optimal estimates. Sankhyã, B, 42, 165-178, 1980.

8 - BICKEL, PJ and LEHMANN, EL. Descriptive statistics for nonparametric models. III. Dispersion Annuary of Statistical. 4, 1139-1158, 1976.

9 - BIRNBAUM, A. On the analysis of factorial Experiments without Replication. Thechnometrics 1, nº4. 343-349, Nov.1959.

10 - BOENTE, Graciela. Asymptotic Theory for Robust Principal Components. Journal of Multivariate Analysis 21, 67-78, 1987.

11 - BOLDRINI,José Luiz et allii. 2 ed. Álgebra Linear. SP, Harper & Row do Brasil, 1980.

12 - BRENOT,M. PARMENTIER, JP PAGES.Sur l'univers des variables et la Stabilite' en analyse factorielle. Data Analysis and Informatics, III, North Holland, 1984.

13 - BUHRMAN, JM and RUYMGAART, FH. An application of linearization in non-parametric multivariate analysis. Sankhyã Ser. A. in press, 1979.

14 - CHATFIELD, C. COLLINS, A. Introduction to Multivariate Analysis. London Chapman and Hall, 1980.

15 - CHANG, WC. On using Principal Components before separating a mixture of two multivariate normal distributions. Application Statistical 32, nº3, 267-275, 1983 .

16 - DANIELYAN, GM. ZHARINOV, T. T. OSIPOVA. Aplication of the

Principal Components method and the proportional Hazards Regression Model to Analysis of Survival Data. Biometrika Journal 28 - 1, 73-79, 1986.

17 - DAUXOIS, J. POUSSE, A. and ROMAIN, Y . Asymptotic Theory for the Principal Components Analysis of a Vector Random Function: Some Applications to Statistical Inference. Journal of Multivariate Analysis 12, 136-154, 1982 .

18 - DAVIS, A. Asymptotic theory for Principal Components Analysis: non-normal case. Austral Journal Statistical 19, 206-212, 1977.

19 - DEULIN , SJ. GNANADESIKAN, R, and KETTENRING, J, R. Robust estimation of dispersion matrices and principal components. Journal American of Statistical Association, 76, 354-362, 1981.

20 - _____, Robust estimation and outlier detection with correlation coefficients. Biometrika,62, 3, p 531. 1975.

21 - DRAPER, NR. Applied Regression Analysis. USA, Jonh Wiley & Sous. Inc. 2º ediçãO, 1981.

22 - DUBROU, AM. Procesing of Statistical Data by the Principal Components Methods. Statistika, Moscou , 1978.

23 - DUDZINSKI, M. L. NORRIS, J. M. CHMURA, JT & EDWARDS, CBH. Repeatability of PC in Samples: normal and non-normal data sets compared. Multivariate Behavioral Research 10 , 109-18, 1975.

24 - EASTEMENT, H. T. and KRZANOWSKI, WJ. Cross_ Validatory choice of the number of components from a Principal Components Analysis. Technometrics 24, 73-77, 1982.

25 - ESCOFIER, B. Stabilité et aproximation en analyse factorielle. Thêse de Doctorat d' Etat, Universitè P e MCurie, Paris VI, 1979.

26 - FANG, C and KRISHNAIAH, PR. Asymptotic distributions funcions of the eigenvalues of some random matrices for non-normal populations. Journal Multivariate Analysis 12, 39-63, 1982.

27 - FISHER, RA. The sampling distribution of some statistics obtained from nonlinear equation. London, Annuary Engenics

9, 238-249, 1939.

28 - FLURY, B. Some relations betwen the comparison of covariance matrices and Principal Components Analysis . Computational Statistics & Data Analysis. 1, 97-109, 1983. North Holland.

29 - FRUTERS, JER. Evaluation of a texture profihe for cooked chicken Breast Meat by Principal Components Analysis. Poultry Science, 55, 229-234, 1976.

30 - GNANADSIKAN, R. and KETTENRING, J. R. Robust estimates, residuals and outlier detection with multireponse data. Biometrics. 28, 81-124, 1972.

31 - GREEN, BF. The orthogonal aproximation of an oblique structure in factor analysis. Psychometrika, 17, 429-440, 1952.

32 - _____. Parameter sensitivity in multivariate methods. Journal Multivariate Behavioural Res., 12, 263-287, 1977.

33 - GUTTMAN, Louis. Some necessary conditions for common factor analysis. Psychometrika, 19, 149-161, 1954.

34 - HAWKINS, Douglas M. and FATTI, Paul L. Exploring data using the minor Principal Components. The Statistician, 33, 325-338, 1984.

35 - HAWKINS, DM. The detection of errors in multivariate data using Principal Components. JASA, 69, 340-344, 1974.

36 - HOTELLING, H. Analysis of a complex of statistical variables in to Principal Components. Journal Educ. Psychometrika, 24, 417-441, 498-520, 1933.

37 - _____ Relation between two sets of variables. Biometrika 28, 321-377, 1936.

38 - HUANG, Steel T. and CAMBANIS, Stamatics. Spherically Invariant Processes: Their Nonlinear Structure, Discrimination, and Estimation. Journal of Multivariate Analysis 9, 59-83, 1979.

39 - IMAN, RONALD L. HELTON, JON C. and CAMPBELL JAMES E. An Approach to Sensitivity Analysis of Computer Models: Part I - Introduction, Input Variable Selection and Preliminary variable Assesment . Journal of Quality Technology ,13, nº 3, july 1981.

40 - _____ _____ _____ An Aproach to Sensitivity Analysis of Computer Models: Part II-

Ranking of Input Variables, response Surface Validation, Distribution Effect and Technique Synopsis. Journal of Quality Technology. 13, nº 4, october 1981.

41 - JACKSON, JE. Quality control methods for several Related Variables. Technometrics, 1, nº4, 359-377, nov. 1959.

42 - _____. Principal Components and Factor Analysis: Part I , Principal Components. Journal Quality Technology. 12, nº4, 201-213, October, 1980.

43 - _____.Principal Components and Factor Analysis Part II, Additional Topics Related To Principal Componets. Journal of Quality Technology. 13, nº1, 46-58, Jan. 1981.

44 - _____. Principal Components and Factor Analysis Part III, Wath is Factor Analysis? Journal of Quality Technology. 13, nº2 , April, 1981.

45 - _____. Multivariate Quality Control. COMMUN. STATIST. THEOR. METH. 14(11), 2657-2688, 1985.

46 - JACKSON,JE. and MORRIS, RH. An application of multivariate quality control to photographic processing. J. Amer. Stat. Assoc. 52, 186-199.

47 - JACKSON, JE. MUDKOLKAR, GS. Control Procedures for residuals Associated whith Principal Components. Technometrics, 21, 341-49, 1979.

49 - JAMES, A. T. Normal Multivariate Analysis and the orthogonal group. Ann. Math. Statistical. ?

50 - _____. Test for a prescribed sub-space of Principal Components. In Journal Multivariate Analysis. IV (PR Krishnaiah, ed) 73-77, North Holland Amsterdam , 1977.

51 - JEFFERS, JN.R Two case studies in the aplication of Principal Components Analysis. Appl. Statistical. 16, 225-236, 1967.

52 - JOLICOEUR, P and MOSIMANN, J. Size and shape variation in the painted turtle: a Principal Components Analysis. Growth 24, 339-354, 1960.

53 - JONSSON, Dag. Some Limit Theorems for the Eigenvalues of a Sample Covariance Matrix. Journal of Multivariate Analysis, 12, 1-38, 1982.

54 - KATO, T. Perturbation Theory for Linear Operators. Springer,

Berlin, 1966.

55 - **KIEFFE, J.** Principal Components of random variables withvalues in a separable Hilbert space. Math. Operations Forch. Statistical 4, Heft 4, 391-406, 1973.

56 - KRZANOWSKI, WJ. The algebric basis of classical multivariate methods. The Statistician, 20, nº4, 51-61, 1971.

57 - _____ Between groups comparison of Principal Components. Journal of Amer. Statistical Assoc. 74, 703-707, 1979. With corrigenda in 76, 1022.

58 - _____ Between_Group comparison of Principal Components Some sampling results. Journal Statistical Comput. Sinul. 15, 141-154, 1982.

59 - _____ Cross_Validatory Choice in Principal Components Analysis: Some Sampling results. Journal Statistical Comput. Simul. 18, 299-314, 1983.

60 - _____ Sensitivity of Principal Components. Journal R. Statistical Soc. B. 46, nº3, 558-563, 1984.

61 - _____ Some exact percentage pointg of a statistic useful in analysis of variance and Principal Components Analysis. Technometrics 21, nº2, 261-263, mai. 1979.

62 - LIANG, WEN-QI and KRISHNAIAH, P.R. Multi - stage Nonparametric Estimation of Density Function Using Orthonormal Systems. Journal of Multivariate Analysis 17, 228-241, 1985.

63 - MANDEL, J Principal Components Analysis Variance and data Structure. Statistika, Neerlandica 26, 119-129, 1972.

64 - MARDIA, KV. The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model. Biometrika, 58, 105-121, 1971.

65 - MARDIA, KV. KENT, J.T. BIBBY, JM. Multivariate Analysis Academic Press London, 1979.

66 - MASSY, W.F. Principal Components Analysis Regression in exploratory data research. Journal American Statistical Association, 60, 234-256, 1965

67 - MORAN, R What should a professor of statistics do ? Austral Journal Statistical . 17 (3) 121-133, 1975.

68 - MUIRHEAD, R. J. and WATERNAUX, C. M. Asymptotic distributions

in canonical correlation analysis and other multivariate procedures for nonnormal populations. Biometrika, 67, 31-44, 1980.

69 - OLKIN, Ingram. Note On the Jacobians of certain matrix transformations useful in multivariate analysis. Biometrika, 40, 43-60, 1955.

70 - PEARSON, K On lines and planes of closest fit to sustems of points in space. Philosophical Magazine, série 6 , 2, 559 -572, 1901.

71 - PERES, C. A. Análise Multivariada. FIPE: Fundação Instituto de Pesquisas Econômicas. Cidade Universitária "Armando de Salles Oliveira", cap 9-34, São Paulo.

72 - RAO, C.R. On the distance between two populations. Sankhyã, 9, 246, 1948.

73 - _____ Analysis of dispersion for multiply classified data with unequal numbers in cells. Sankhya, 15, 253-280, 1955.

74 - _____ The use and interpretation of principal components analysis in aplied research. Sankhyã, A, 26, 329-357, 1964.

75 - _____ Tests for dimensionality and Interactions of mean Vectors under general and Reducible Covariance Structures. Journal Multivariate Analysis 16, 173-184, 1989.

76 - RUYMGAART, F. H. A Robust Principal Components Analysis Journal Multivariate Analysis 11, 485-497, 1981.

77 - SILVERSTEIN, J. W. Some Limit Theorems on the Eigenvectors of Large dimensional Sample Covariance Matrices. Journal of Multivariate Analysis. 15, 295-324, 1984.

78 - TOMASSONE. BIOMETRIE Institut National Agronomique Paris Grignon Mathematique et Informatique. Julho, 1987.

79 - TYLER, D. A Class of Asymptotic tests for Principal Components Vectors. Annuary Statistical. 11, nº4, 1243-1250, 1983.

80 - _____ The asymptotic distribution of principal components roots under local alternative to multiple roots. Annuary Statistical. 11, 1232-1242, 1983.

81 - _____ Inferences for eigenvectors. Annuary of Statistical 9, nº 4, 732-736, 1981.

82 - ULEBART, L. MORINTAU, A. Statistical Significance criteria in

multiple-choice data reduction and visualization . Psychometric Society Meeting, CHAPEL HILL, May 27-29, 1981.

83 - ZHAO, L.C., KRISHNAIAH P.R. and BAI, Z.D. On Detection of the Number of Signals in presence of White Noise. Journal of Multivariate Analysis 20, 1-25, 1986.

84 - WATERNAUX, C. M. Asymptotic distribution of the sample roots for a nonnormal population. Biometrika, 63, 3, 639-645, 1976.

85 - _____ Principal Components in the nonnormal case: the test for sphericity. Journal Multivariate Analysis to appear. 1977.

86 - _____ In the Nonnormal Case: The test of Equality of Q Roots. Journal of Multivariate Analysis, 14, 323-335, 1984.

87 - WOLD, S. Cross _ Validatory estimation of the numbers of components in factor and Principal Components models. Technometrics 20, 397-405, 1978.

88 - YOHAI, V. J. ACKERMANN, W y HAIGH, C. Nonlinear Principal Components. Quality and Quantity, 19, 53- 69, 1985.