Análise de componentes Principais

Introdução

Dos livros A.D.M (D.Peña) e A.D.M. (J.Lattin, et al)

[1]EACH-USP
Universidade de São Paulo

Outline

Outline

Introdução

Divisão das principais técnicas

Objetivos	Enf.Descritivo	Enf.Inferencial
Res. de dados	Descrição dos dados	Const. de modelos
Obter indic.	ACP	AF
	Escalado multidim.	
	A. Corresp.	
Classificação	A.conglom.	A. Discriminante
Const. grupos	A.conglom.	Class. com misturas
Rel. var.	Regress. Mult.	Correl. canónica

Método	Interd. *vs* depend.	Expl. *vs* Confirm.	Métr, *vs* não métr.	objetivos
ACP	Interd.	Expl.	Métr.	redução da dimensão
A.F.Exp	Interd.	Expl.	Métr.	padrões/traços latentes
A.F.Conf	Interd.	Conf.	Métr.	Verif. Modelos
Esc.Mult.	Interd.	Expl. princ.	Métr. ou não	Repr.espacial das similaridades
A.Cong.	Interd.	Expl	Métr. ou não	Agrupamentos das similaridades
C.Canon.	Dep	Expl. princ.	Métr.	Covariação entre dois conjuntos

Método	Interd. *vs* depend.	Expl. *vs* Confirm.	Métr, *vs* não métr.	objetivos
Eq. Estr. c.v.lat	Dep.	Conf.	Métr.	Dep. com erro de medida
ANOVA	Dep	Conf.	Métr. e não	Corr.canón. com v.X disc.
A.Disc.	Dep.	Expl. ou conf.	Métr. e não	Corr. canón. com v. Y disc.
Mod. Logit	Dep	Expl. ou conf	Métr. e não	Modelo de prob. não linear para resultados discr.

Outline

Introdução
ACP

- Objetivo Principal: Redução da dimensionalidade.
  Se for possível descrever com precisão os valores de p variáveis por um subconjunto $r < p$ delas, teremos reduzido a dimensão do problema com o custo de pequena perda de informação.
- as novas variáveis são combinação linear das originais.
- Quanto maior a dependência, menor número de variáveis (novas) serão necessários.
- ACP permite transformar as variáveis originais (normalmente correlacionadas) em variáveis não correlacionadas, facilitando a interpretação dos dados.

Introdução
ACP

- Objetivo Principal: Redução da dimensionalidade.
  Se for possível descrever com precisão os valores de p variáveis por um subconjunto $r < p$ delas, teremos reduzido a dimensão do problema com o custo de pequena perda de informação.
- as novas variáveis são combinação linear das originais.
- Quanto maior a dependência, menor número de variáveis (novas) serão necessários.
- ACP permite transformar as variáveis originais (normalmente correlacionadas) em variáveis não correlacionadas, facilitando a interpretação dos dados.

Introdução
ACP

- Objetivo Principal: Redução da dimensionalidade.
  Se for possível descrever com precisão os valores de p variáveis por um subconjunto $r < p$ delas, teremos reduzido a dimensão do problema com o custo de pequena perda de informação.
- as novas variáveis são combinação linear das originais.
- Quanto maior a dependência, menor número de variáveis (novas) serão necessários.
- ACP permite transformar as variáveis originais (normalmente correlacionadas) em variáveis não correlacionadas, facilitando a interpretação dos dados.

Introdução
ACP

- Objetivo Principal: Redução da dimensionalidade.
  Se for possível descrever com precisão os valores de p variáveis por um subconjunto $r < p$ delas, teremos reduzido a dimensão do problema com o custo de pequena perda de informação.
- as novas variáveis são combinação linear das originais.
- Quanto maior a dependência, menor número de variáveis (novas) serão necessários.
- ACP permite transformar as variáveis originais (normalmente correlacionadas) em variáveis não correlacionadas, facilitando a interpretação dos dados.

Existe uma distinção entre ACP e AF confirmatória, embora ambos os métodos possam ser usados para alcançar os mesmos fins, os modelos subjacentes são diferentes.
Na A.F. estamos preocupados em identificar as fontes subjacentes de variação comum a duas ou mais variáveis (fatores comuns). Um presuposto nesse modelo é que a variação em cada variável observada é atribuível aos fatores comuns subjacentes e a um fator específico (erro de medida). Em ACP, estamos preocupados em reexpresar os dados, não há modelo de medida subjacente. Cada componente é uma combinação exata linear das variáves originais.

Mecânica para obtenção das componentes (uso de matríz de correlação)

Nosso objetivo é encontrar um C.L das variáveis originais $\boldsymbol{X} = [\boldsymbol{x_1}, ..., \boldsymbol{x_p}]$ com variação máxima. Suponha que $\boldsymbol{X}$ seja padronizado (i.e, média 0 e variância 1). Se a C.L for representada por $\boldsymbol{u} = (u_1, ..., u_p)^{'}$, nosso objetivo é escolher $\boldsymbol{u}$ que maximize a variância de $\boldsymbol{z} = \boldsymbol{Xu}$ escrita como:

$$var(\boldsymbol{z}) = \frac{1}{n-1}\boldsymbol{u}^{'}\boldsymbol{X}^{'}\boldsymbol{Xu} \tag{1}$$

Como $\boldsymbol{X}$ é padrinizado, o termo $\boldsymbol{R} = \frac{1}{n-1}\boldsymbol{X}^{'}\boldsymbol{X}$ é a matriz de correlação, temos então:

$$var(\boldsymbol{z}) = \boldsymbol{u}^{'}\boldsymbol{Ru} \tag{2}$$

Se quisermos maximizar a equação 2 ainda podemos escolher $\boldsymbol{u}$ de modos que $var(z)$ vá para infinito. Por isso, impomos a condição:

$$\boldsymbol{u}'\boldsymbol{u} = \mathbf{1}$$

O problema passa a ser:

$$\begin{aligned} &\text{Escolha } \boldsymbol{u} \text{ para maximizar } \boldsymbol{u}'\boldsymbol{R}\boldsymbol{u} \\ &\text{sujeito à limitação } \boldsymbol{u}'\boldsymbol{u} = \mathbf{1} \end{aligned} \tag{3}$$

Podemos resolver o problema posto acima , formando o Lagrangiano, tomando a primeira derivada e atribuindo-lhe o valor zero:

$$L = \boldsymbol{u}'\boldsymbol{R}\boldsymbol{u} - \lambda(\boldsymbol{u}'\boldsymbol{u} - 1) \tag{4}$$

onde λ é chamado de multiplicador de lagrange. Observe que λ pode ser escolhido para penalizar a função objetivo se a igualdade $\boldsymbol{u}'\boldsymbol{u} = 1$ não for satisfeita.
Tomando a derivada de L em relação aos elementos de $\boldsymbol{u}$ temos:

$$\frac{\partial L}{\partial \boldsymbol{u}} = 2\boldsymbol{R}\boldsymbol{u} - 2\lambda\boldsymbol{u} \tag{5}$$

de onde

$$\boldsymbol{R}\boldsymbol{u} = \lambda\boldsymbol{u} \quad ou \quad (\boldsymbol{R} - \lambda\boldsymbol{I})\boldsymbol{u} = 0 \tag{6}$$

o vetor $\boldsymbol{u}$ é chamado de autovetor e o escalar λ é o autovalor. Dado que a matriz $\boldsymbol{R}$ é de posto completo, a solução consistirá em p autovalores e autovetores associados.

Variância explicada pelas componentes principais

Observe que os λ_i são iguais às variâncias para os componentes principais (veja o exemplo). A resposta está em que o objetivo é o de maximizar a variância de z, dada por $\boldsymbol{u}'\boldsymbol{R}\boldsymbol{u}$. Observe que para que as condições de primeira ordem se mantenham (6), $\boldsymbol{R}\boldsymbol{u} = \lambda\boldsymbol{u}$.
Substituindo $\lambda\boldsymbol{u}$ por $\boldsymbol{R}\boldsymbol{u}$ temos

$$var(z) = \boldsymbol{u}'\boldsymbol{R}\boldsymbol{u} = \boldsymbol{u}'\lambda\boldsymbol{u} = \lambda\boldsymbol{u}'\boldsymbol{u} = \lambda \tag{7}$$

Usamos $\boldsymbol{D}$ para representar a matriz de covariância diagonal das componentes principais (no exemplo: $\boldsymbol{D} = diag(\lambda_1, \lambda_2, \lambda_3)$). Como as componentes principais são não correlacionados, a variância da soma é a soma das variâncias, de onde $var(\sum_i z_i) = \sum_i \lambda_i$.

A variância para todas as componentes principais é $tr(\boldsymbol{D})$. Sabemos que $\boldsymbol{D} = U^{'}RU$ com $\boldsymbol{U}^{'}, \boldsymbol{R}, \boldsymbol{U}$ matrizes quadradas.
Uma propriedade do traço é que tem o mesmo valor para permutações do produto de matrices:

$$tr(\boldsymbol{ABC}) = tr(\boldsymbol{CAB}) = tr(\boldsymbol{BCA})$$

Portanto,

$$tr(\boldsymbol{U}^{'}\boldsymbol{RU}) = tr(\boldsymbol{RU}^{'}\boldsymbol{U})$$

Sabemos que $\boldsymbol{U}^{'}\boldsymbol{U} = \boldsymbol{I}$, então

$$tr(\boldsymbol{U}^{'}\boldsymbol{RU}) = tr(\boldsymbol{R})$$

Cargas do componente fatorial

A matriz de correlação do score do componente principal $\boldsymbol{Z}$ com os dados originais $\boldsymbol{X}$, ajuda a interpretar $\boldsymbol{Z}$ se soubermos o padrão de relacionamento com os dados $\boldsymbol{X}$ existentes:

$$corr(\boldsymbol{X}, \boldsymbol{Z}) = \frac{1}{n-1}\boldsymbol{X}`\boldsymbol{Z}_s \tag{8}$$

$\boldsymbol{Z}_s = \boldsymbol{Z}\boldsymbol{D}^{-1/2}$ é a matriz dos componentes principais padronizados. substituindo $\boldsymbol{XU}$ por $\boldsymbol{Z}$, temos:

$$corr(\boldsymbol{X}, \boldsymbol{Z}) = \frac{1}{n-1}\boldsymbol{X}`\boldsymbol{X}\boldsymbol{U}\boldsymbol{D}^{-1/2} \tag{9}$$

sabemos que $\boldsymbol{R} = \frac{1}{n-1}\boldsymbol{X}`\boldsymbol{X}$ e que $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}`$

$$corr(\boldsymbol{X}, \boldsymbol{Z}) = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}`)\boldsymbol{U}\boldsymbol{D}^{-1/2} = \boldsymbol{U}\boldsymbol{D}^{1/2} \tag{10}$$

Os resultados das correlações entre os componentes principais e as variáveis originais, às quais iremos nos referir como as cargas do componente principal, são representadas como:

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{D}^{1/2} \tag{11}$$

são determinadas por um escalonamento simples dos vetores subjacentes.
Das cargas do componente principal, determinamos a quantidade de variância em cada variável original explicada por qualquer número de componentes principais. A expressão para a variância explicada na variável X_i pelas primeiras c componentes é:

$$\sum_{j=1}^{c} f_{ij}^2 \tag{12}$$

onde c é o número de componentes principais retidos e f_{ij} a correlação entre X_i e Z_j da matriz $\boldsymbol{F}$
Se $c = p$ (todos os CP retidos) então $\sum_j f_{ij}^2 = 1$

Chegando à solução

Para obter a matriz padronizada dos componentes principais $\boldsymbol{Z}_s$, multiplicamos $\boldsymbol{Z} = \boldsymbol{XU}$ por $\boldsymbol{D}^{-1/2}$, ($\boldsymbol{D}$ é a matriz diagonal da variância de $\boldsymbol{Z}$)

$$\boldsymbol{Z}_s = \boldsymbol{XUD}^{-1/2} \tag{13}$$

com um pouco de manipulação algébrica, podemos expressar $\boldsymbol{X}$ como função de $\boldsymbol{Z}_s$. Temos:

$$\boldsymbol{X} = \boldsymbol{Z}_s\boldsymbol{D}^{1/2}\boldsymbol{U}' \tag{14}$$

O que mostra que qualquer matriz $\boldsymbol{X}$ pode ser expressa como o produto de três matrices:

- $\boldsymbol{Z}_s$ matriz de variáveis não correlacionadas (cada uma co variância unitária)
- $\boldsymbol{D}^{1/2}$ matriz diagonal que executa uma transformação extensora ("despadronizando"$\boldsymbol{Z}_s$, multiplicando pelos desvios padrões de $\boldsymbol{Z}$).
- $\boldsymbol{U}^{\text{‘}}$ matriz de transformação que realiza uma rotação ortogonal.

O modo de expressar $\boldsymbol{X}$ é chamado de Decomposição em valores singulares (DVS ou SVD)

Teste de Esfericidade de Bartlett

Este teste é realizado para saber quando é apropriado usar C:P. Até agora supomos que as variáveis eram altamente correlacionadas para justificar a redução de dimensões. Entretanto, se as variáveis forem altamente independentes, o uso de CP pode não ser apropriado. A questão colocada pelo teste de esfericidade é: A matriz de correlação deve ser decomposta em fatores em primeiro lugar?

O teste é um teste Qui-quadrado aproximado com um teste estatístico que é função do determinante da matriz de correlação $\boldsymbol{R}$:

$$\chi^2 \left[\frac{(p^2 - p)}{2} \right] = - \left[(n-1) - \frac{(2p+5)}{6} \right] ln|\boldsymbol{R}|$$

onde :
$ln|\boldsymbol{R}| =$ log natural do determinante de $\boldsymbol{R}$
$(p^2 - p)/2 =$ número de g.l. associado ao teste χ^2
$p =$ número de variáveis
$n =$ número de observações

A lógica por trás do teste é: $|\boldsymbol{R}|$ é uma medida de variância generalizada, e é calculada como:

$$|\boldsymbol{R}| = \prod_{j=1}^{p} \lambda_j \tag{15}$$

se as variáveis forem mutuamente independentes, esperamos que $\boldsymbol{R}$ se aproxime de $\boldsymbol{I}$, neste caso, o diagrama de dispersão tem formato esférico em vez de oval. Todos os autovalores estão próximos de 1 (formato circular), portanto $|\boldsymbol{R}|$ é proximo de 1 e $ln|\boldsymbol{R}|$ é próximo de 0. À medida que o montante de correlação aumente, o diagrama de dispersão dos dados começa a parecer mais elipsoidal, neste caso, algus dos autovalores de $\boldsymbol{R}$ serão maiores que 1 e alguns próximos de 0. Como resultado, o produto dos autovalores se aproxima de 0 o que significa que $ln|\boldsymbol{R}|$ torna-se um número negativo maior.

Quando $|\boldsymbol{R}|$ é próximo de 1, o χ^2 de Bartlett é próximo de 0, indicando um bom ajuste e não nos dando razão para rejeitar $H_0 : \ esfericidade$.
Se não pudermos rejeitar H_0, concluimos que não é apropriado reduzir a dimensionalidade.
No caso contrário, quanto mais correlação presente, $|\boldsymbol{R}|$ estará mais próxima de 0, o que significa que $ln|\boldsymbol{R}|$ é um valor negativo grande, e o teste assume um valor positivo grande, indicando um mau ajuste dos dados à H_0, rejeitando a hipótese de esfericidade.

Como os dados deveriam ser escalonados?

Usamos dados padronizados (matriz de correlação em vez de matriz de covariância), e as razões principais são:

- A ACP busca maximizar a variância, ela pode ser sensível a diferencias de escala entre as variáveis
- A padronização garante que os dados sejam expressos em unidades comparáveis

Quando não padronizar:

- quando todas as questões sejam medidas na mesma escala (Likert por exemplo)
- Em casos em que houver razão para acreditar que a variância de uma variável é um indicador de sua importância em geral (ou alto valor de informação).
- no caso de uso de matriz de covariâncias, as cargas dos CP serão graduados como covariâncias em vez de correlações.

Quantos componentes devem ser retidos?

Caso seja apropriado reduzir a dimensionalidade, e quisermos saber quantos componentes devem ser retidos, podemos usar uma aplicação sequencial do teste de Bartlett.

- Se a esfericidade é rejeitada, extraímos o maior CP e depois testamos a matriz de correlação residual :

$$\boldsymbol{R} - \lambda \boldsymbol{\mu_1} \boldsymbol{\mu_1}'$$

  para determinar se seu determinante é diferente de zero.
- Infelizmente, dado o poder do teste χ^2, esta abordagem retém um grande número de componentes

Gráfico Scree

Proposta por Cattel (1966). Envolve desenhar a variância explicada por cada CP na ordem de maior a menor. Buscamos um "ângulo" na curva e retemos somente aqueles CP que estão acima desse ângulo.

Regra de Kaiser

Kaiser (1959) recomendou somente a retenção dos CP com autovalores que excedam a unidade (supondo que estamos tratando com variáveis padronizadas). (Deve ser encarada como uma orientação e não como uma regra inviolável).

Procedimento de Horn

Horn (1965) sugere o uso de um autovalor como limite, a partir da análise de CP de dados randômicos com o mesmo número de variáveis e o mesmo número de observações. Este procedimento é baseado na percepção de que a ACP ao procurar explicar a maior variabilidade possível, irá capitalizar sobre a variação da amostra randômica dentro dos dados.
Para calcular o limite de Horn, geramos uma matriz de daos da amostra de ordem $n \times p$ usando um gerador de números aleatórios. Aplicamos então os CP aos dados aleatórios e desenhamos os autovalores em um gráfico Scree junto com os autovalores dos dados reais. Somente aqueles CP com autovalores que excedem o limite de Horn são retidos.

Uma modificação adicionada ao prodecimento é:
Em lugar de usar um gerador de números aleatórios (com alguma distribuição subjacente), podemos melhorar a situação a partir dos dados reais. Em outras palavras, fazemos testes com a recolocação de cada variável no conjunto de dados. Observe que devemos avançar e melhorar a situação a partir de cada variável separadamente (i.e, da distribuição marginal de cada variável e não da distribuição conjunta) para gerar uma amostra na qual a distribuição conjunta subjacente seja completamente independente.

Variância explicada

Às vezes, será importante reter um número suficiente de componentes para reconstruir o conjunto de dados originais (i.e, para que sejamos capazes de explicar determinado valor da variância em cada uma das variáveis originais). Sabemos que

$$\boldsymbol{X} = \boldsymbol{Z}_s \boldsymbol{D}^{1/2} \boldsymbol{U}'$$

se retivermos c Componentes padronizados,a relação aproximada é:

$$\boldsymbol{X} \approx [\boldsymbol{z}_{1s}, \boldsymbol{z}_{2s}, \ldots, \boldsymbol{z}_{cs}] \boldsymbol{D}^{1/2} \boldsymbol{U}'$$

Podemos impor a limitação: reter um número suficiente de componentes para que possamos explicar pelo menos 50% (a escolha), da variância em cada variável original. O exame dos autovalores diz se podemos explicar 50% da variância em todas as variáveis, mas não se captamos pelo menos 50% da variância de cada variável.

Quando o critério é usado, deve ser baseado em uma avaliação da confiabilidade (e portanto, a porcentagem da variação do erro) das variáveis que estão sendo analisadas.

Como avaliar a validade da solução?

Estamos querendo questionar a posibilidade de generalização dos resultados de nossa análise além da amostra. Quando se lida com dados de corte transversal (tamanho n da amostra), podemos perguntar até que ponto os achados de nossa análise se extendem para uma nova amostra de tamanho n da mesma população subjacente?.
Uma sugestão é utilizar uma amostra de teste. Antes de usar ACP, separamos uma parte dos dados escolhida aleatoriamente (como regra dois terços para análise e um terço para teste).

O procedimento tem um problema quando o conjunto de dados é pequeno, pois aumenta a probabilidade de capitalização sobre o acaso (procedimento de Horn) .

Validação Jackknife

também conhecida como abordagem por método-U por amostra teste, é útil no caso de amostras pequenas. A idéia: Retemos uma observação, e conduzimos uma ACP sobre as $n-1$ restantes, depois, utilizamos o vetor $\boldsymbol{u_1}$ para calcular o valor de $\boldsymbol{Z_1}$ na única observação retida. Repetimos esse procedimento para todos os outros dados da amostra. Comparamos então a variância dos valores "jackknificados" com a variância do primeiro CP (determinado pelo conjunto completo). A vantagem é que não se usam quaisquer das informações da i-ésima observação para calcular-se o escore do seu CP.

Validação Boostrap

Na ausência de uma amostra teste, podemos obter um novo conjunto de dados reamostrando os originais, um processo conhecido como boostrapping. Se supormos que os dados na amostra são representativos de uma população subjacente, retirar várias amostras de tamanho n (com reposição), poderia imitar a variabilidade introduzida pela amostragem da população como um todo. O boostrapping permite-nos avaliar a distribuição de nossos resultados de análise dos CP mesmo quando suas prorpriedades estatísticas não sejam bem conhecidas.

O procedimento é: criamos uma nova amostra de dados (podem haver repetições), e depois formamos uma combinação linear dos dados "boostrappados" , usando o vetor $\boldsymbol{u_1}$ do conjunto inicial de dados. Podemos então comparar a variância dessa combinação linear com a variância do primeiro CP da amostra boostrapped (que sabemos ser a combinação linear com a variância máxima). Se a comparação for próxima (próximo de 1), concluimos que a variação é sistemática em ambas as amostras (i.e, comum à população subjacente). Se o quociente for pequeno, concluimos que nosso achado não pode ser generalizado para fora da amostra.

Exemplo

A tabela abaixo mostra a matriz de correlação para três variáveis X_1, X_2, X_3 padronizados (média zero e variância 1 para cada):

Tabela 4.5 Matriz de correlação para X_1, X_2 e X_3

	X_1	X_2	X_3
X_1	1,000	0,562	0,704
X_2	0,562	1,000	0,304
X_3	0,704	0,304	1,000
	$\text{var}(X_1) = 1{,}00$	$\text{var}(X_2) = 1{,}00$	$\text{var}(X_3) = 1{,}00$

É dificil representar esses dados tridimensionais de forma bidimensional. A Figura abaixo, mostra uma visão estilizada de uma diagrama de dispersão de três direções. O contorno tem um formato elíptico com um corte transversal oval. A projeção em cada plano das coordenadas resulta em um diagrama de dispersão bidimensional:

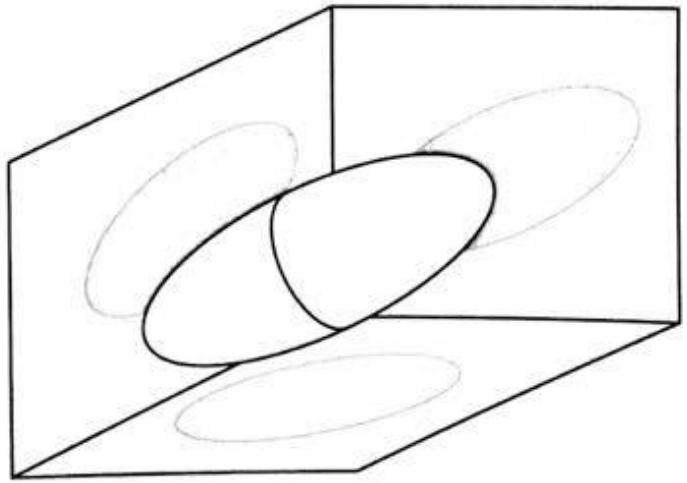


Figura 4.3 Visão estilizada tridimensional do formato de distribuição de X_1, X_2 e X_3. Sombras representam os formatos em duas dimensões.

Um diagrama de dispersão tridimensional dos dados reais é apresentado. Também é representado o diagrama de dispersão em pares, os que mostram uma correlação positiva entre todos os três pares de variáveis ($Corr(X_2, X_3) = 0,3$, $corr(X_1, X_3) = 0,7$)

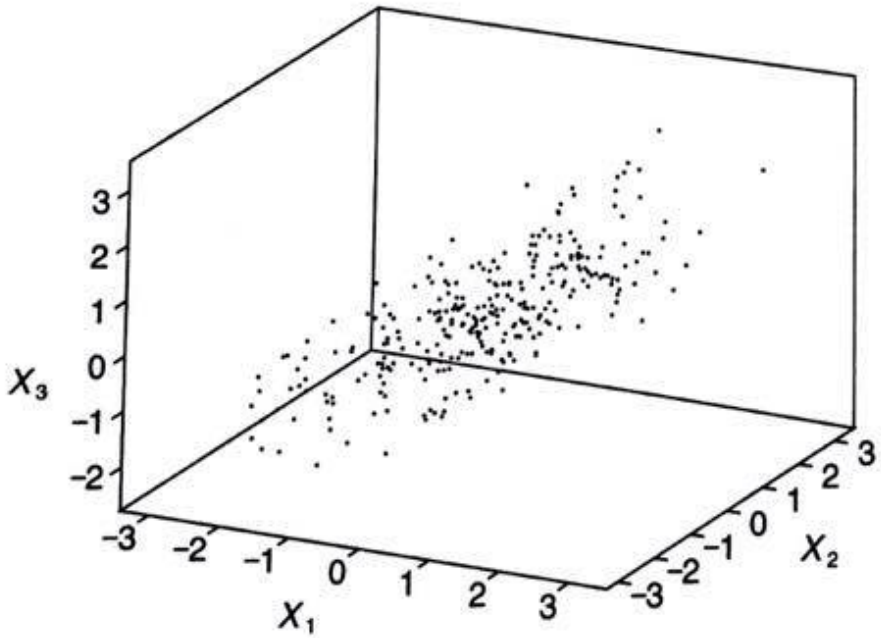


Figura 4.4 Diagrama de dispersão tridimensional dos valores reais de X_1, X_2 e X_3

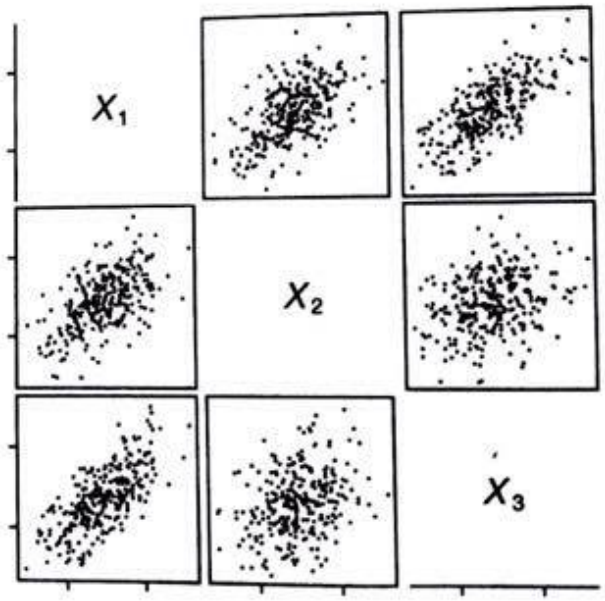


Figura 4.5 Diagrama de dispersão em pares de X_1 *versus* X_2, X_1 *versus* X_3 e X_2 *versus* X_3.

A pergunta que tentamos resolver é: É possível usar uma única dimensão para captar e transmitir a maior parte das informações contidas nos X_i?.
o primeiro CP, é a combinação linear dos X_i que exibe variância máxima, e lembrando que uma combinação linear é simplesmente a projeção de todos os pontos no espaço, em um único eixo. A Figura mostra a orientação do primeiro CP. É o eixo mais longo da configuração em formato elipse (a linha que corre ao longo da elipse). Se projetarmos cada ponto do dado sobre este eixo, a variável resultante (Z_1), exibirá maior variância possível.

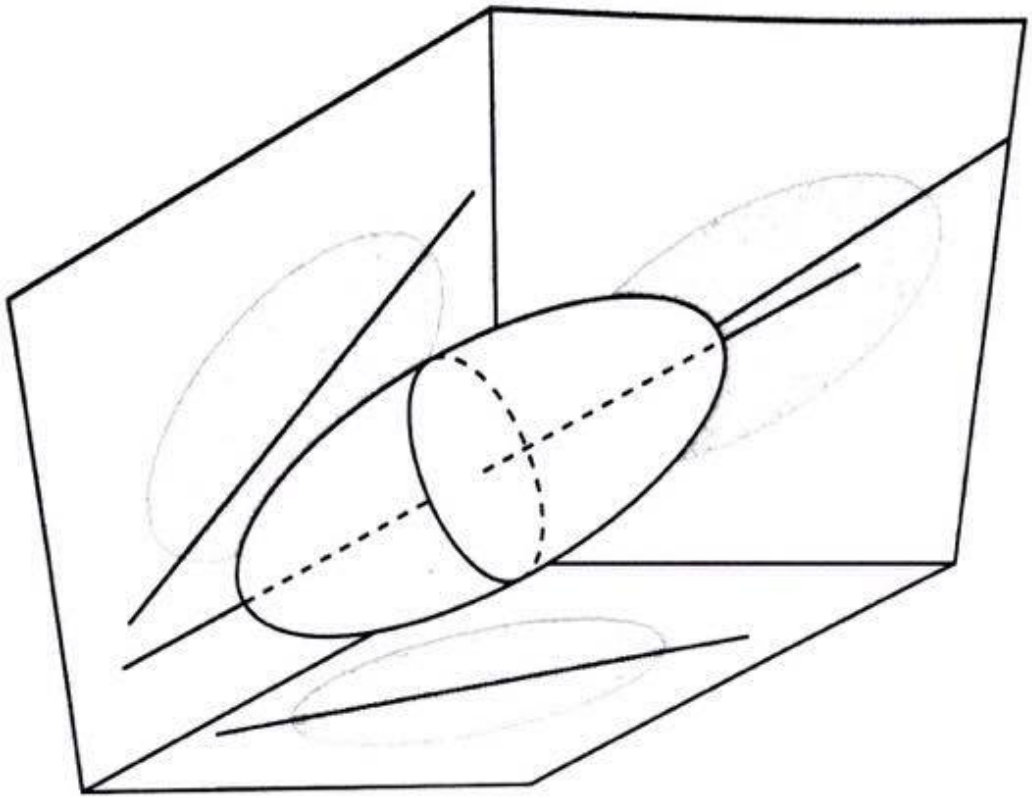


Figura 4.6 Visão tridimensional estilizada identificando o *primeiro componente principal*.

Suponhamos que $\boldsymbol{u_1} = (u_{11}, u_{12}, u_{13})$ represente um vetor de comprimento 1 orientado ao longo do eixo mais longo do elipsoide, de modo que $\boldsymbol{z_1} = \boldsymbol{X u_1}$, em que $\boldsymbol{z_1}$ é um vetor n-dimensional consistindo em valores de Z_1 e $\boldsymbol{X}$ é a matriz $n \times 3$.
A variância de Z_1 foi 2,5, mais que o dobro da variância de qualquer uma das variáveis originais. Podemos pensar na variância de Z_1 como a variância explicada pelo primeiro componente, quanto maior, mais informações dos dados originais estarão contidos nesse único componente.
Embora Z_1 explique grande parte da variação dos X_i's, não explica toda a variação. Pela aplicação sequencial da lógica, podemos agora identificar uma segunda combinação linear que explica a maior parte possível da variação residual

As dimensões restantes são dadas pelo plano perpendicular ao eixo central (linha orientada do vetor $\boldsymbol{u_1}$. Uma visão estilizada é apresentada na figura abaixo, assim como o diagrama de dispersão tridimensional dos dados reais com Z_1 removido.

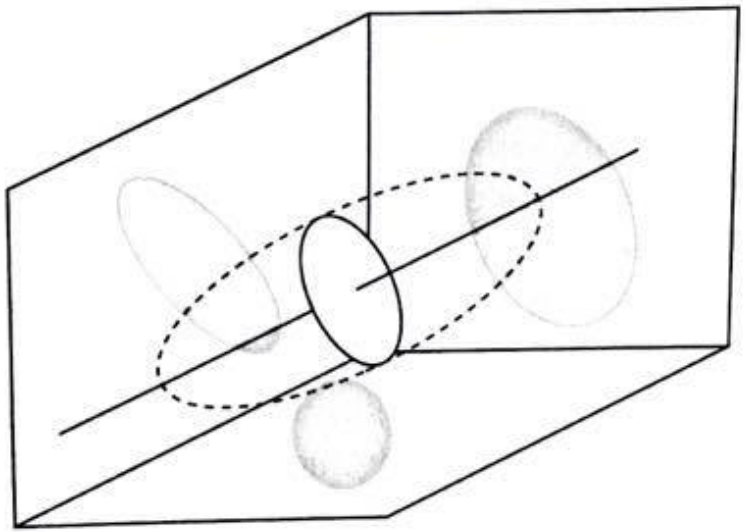


Figura 4.7 Visão tridimensional estilizada após a remoção das informações explicadas pelo primeiro componente principal.

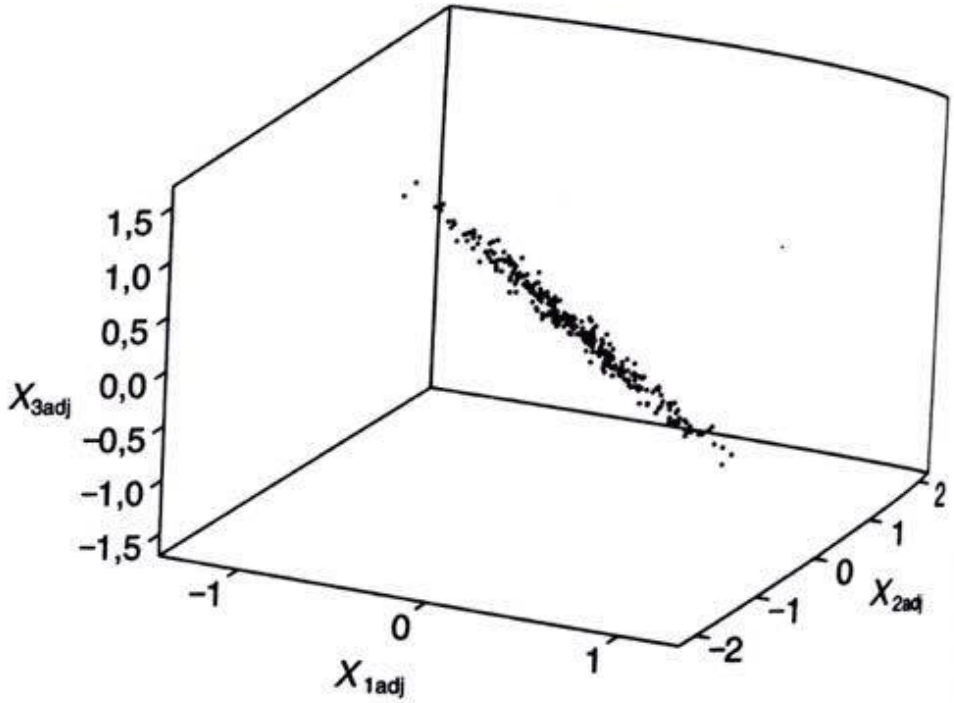


Figura 4.8 Diagrama de dispersão tridimensional de X_1, X_2 e X_3, após a remoção das informações em Z_1.

Para obter uma melhor ideia sobre o formato da configuração no corte transversal, projetando-o de volta no espaço da coordenada original e olhando para o diagrama de dispersão em pares X_1 vs X_2, X_1 versus X_3 e X_2 versus X_3 . Esses diagramas são apresentados na figura abaixo. São análogos à figura apresentada acima, com a remoção da variação nos dados explicada por Z_1 .

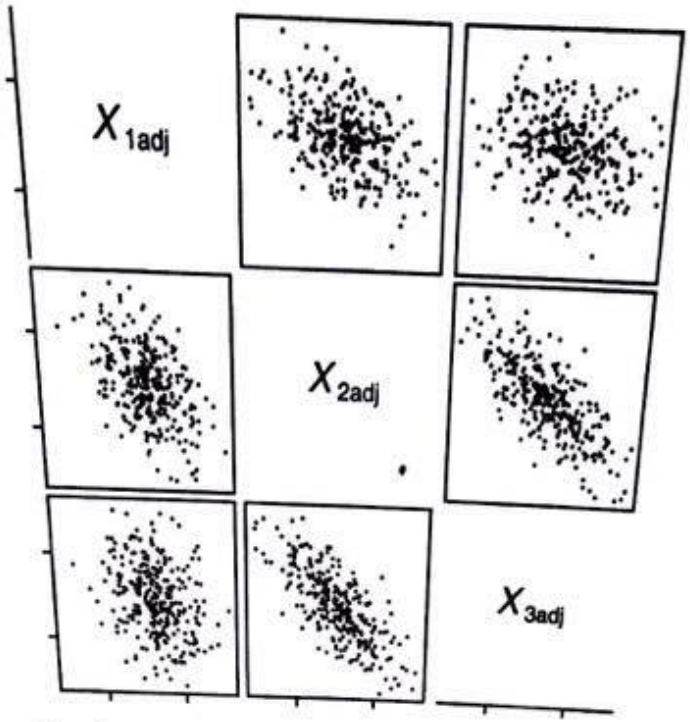


Figura 4.9 Diagramas de dispersão de pares X_1 *versus* X_2, X_1 *versus* X_3 e X_2 *versus* X_3, *após remoção das* informações em Z_1.

Os diagramas revelam algo que não era óbvio na matriz de correlação (que mostrava associação positiva entre as três variáveis. Depois de explicar Z_1, os diagramas de dispersão sugerem que os padrões de associação remanescentes refletem uma forte associação negativa entre X_2 e X_3. Isto diz que os CP além de reducir a dimensionalidade, provêm insights sobre os padrões de associação.
Escolhemos agora uma segunda combinação linear dos X_i's que explique a maior variabilidade possível restante e não explicada por Z_1: A variável resultante Z_2 .
Considere $\boldsymbol{u_2}$ um vetor de comprimento unitário na direção do eixo longo do corte transversal, de modo que

$$\boldsymbol{z_2 = Xu_2}$$

em que $\boldsymbol{z_2}$ é um vetor n-dimensional com os valores Z_2. Observe que $\boldsymbol{u_1}$ e $\boldsymbol{u_2}$ são não correlacionados. Assim como Z_1 e Z_2.

A variância de Z_2 (0,72) é menor que a de Z_1 . A variação residual restante (não explicada nem por Z_1 nem Z_2) é inteiramente explicada por Z_3.
As três novas variáveis estão resumidas na figura abaixo, junto com a representação estilizada dos Z_i's.
O diagrama de dispersão dos valores reais, mostram que a configuração original dos pontos não mudou, apenas sua orientação.
A transformação $\boldsymbol{Z} = \boldsymbol{XU}$ serviu para rodar os eixos do diagrama de dispersão original, enquanto preserva sua ortogonalidade (devido à ortogonalidade mútua dos $\boldsymbol{u_i}$'s

Tabela 4.6 Matriz de correlação para Z_1, Z_2 e Z_3

	Z_1	Z_2	Z_3
Z_1	1,000	0,000	0,000
Z_2	0,000	1,000	0,000
Z_3	0,000	0,000	1,000
	var(Z_1) = 2,05	var(Z_2) = 0,72	var(Z_3) = 0,23

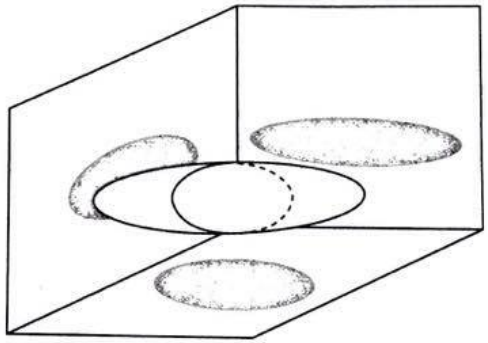


Figura 4.10 Visão tridimensional estilizada da forma de distribuição de Z_1, Z_2 e Z_3.

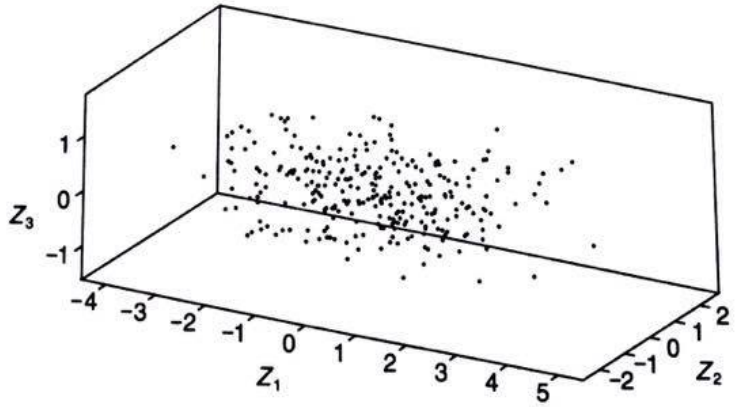


Figura 4.11 Diagrama de dispersão tridimensional dos valores reais de Z_1, Z_2 e Z_3.

Normalmente é instrutivo examinar a relação entre as componentes principais Z e as variáveis originais X. Uma maneira é olhar para a matriz de correlação (tabela abaixo). Embora Z_1, Z_2, Z_3 não sejam mutuamente correlacionadas, elas são relacionadas a X_1, X_2, X_3 em padrões consistentes com a análise precedente.
Essas correlações são conhecidas como cargas dos CP. Por exemplo, Z_1 é correlacionada positivamente a X_1, X_2, X_3, (consistente com a ideia de que Z_1 reflete um componente de variância compartilhada subjacente a todas as variáveis originais). O componente Z_2 é pouco correlacionado com X_1 , positivamente correlacionado com X_2 e negativamente com X_3
As cargas do CP são úteis para nos dizer quanto da variância em cada uma das variáveis originais X é explicado pelos componentes principais.

Tabela 4.7 Cargas do componente principal. Correlações entre os componentes principais e os dados originais

	Cargas do Componente Principal		
	Z_1	Z_2	Z_3
X_1	0,9279	−0,0798	−0,3641
X_2	0,7255	0,6696	0,1590
X_3	0,8222	−0,5008	0,2706

Podemos dizer da tabela que a proporção da variância em X_1 que é explicada pelo Z_1 é $0,93^2 = 0,86$. Da mesma maneira, a variância em X_1 explicada pelo segundo componente Z_2 é menor que 1% $(-0,08^2)$. Uma vez que Z_1 e Z_2 são não correlacionados, podemos somar os valores de R^2 para determinar o valor da viariância em X_1 explicada pelos primeiros dos componentes principais. Neste caso, a proporção é: $(0,93^2 + (-0,08)^2 = 0,87)$. I.e, os dois primeiros CP, explicam quase o 90% da variância em cada variável original.

Na tabela abaixo, mostramos os valores dos autovalores e autovetores para os dados X.

Tabela 4.8 Autovalores (λ) e autovetores ($\mathbf{u}$) de X_1, X_2 e X_3

$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$
0,65	0,09	−0,76
0,51	0,80	0,33
0,57	−0,59	0,56
$\lambda_1 = 2{,}05$	$\lambda_2 = 0{,}72$	$\lambda_3 = 0{,}23$

veja que os autovalores λ_1, λ_2 e λ_3 são iguais às variâncias para os três CP na tabela 4.6

Versão para matriz de covariâncias

A seguir, apresentamos a versão de ACP quando utilizarmos a matriz de covariâncias para a obtenção dos componentes.

Apresentação do problema

Supondo que estão disponíveis os valores de p variáveis em n elementos da população, por meio de uma matriz X $(n \times p)$.
Suponha que a cada variável foi subtraida sua média, de forma que a matriz X tem média zero e matriz de covariâncias dada por $\frac{1}{n}X^{`}X$.

Apresentação do problema

Deseja-se encontrar um sub-espaço de menor dimensão que p tal que ao projetar sobre ele os pontos, conservem a sua estrutura com menor distorção possível.
Se considerarmos um ponto x_i e uma direção $a_1 = (a_{11}, ..., a_{1p})'$ (com norma unitária), a projeção do ponto x_i sobre a_1 é o escalar:

$$z_i = a_{11}x_{i1} + ... + a_{1p}x_{ip} = a_1'x_i$$

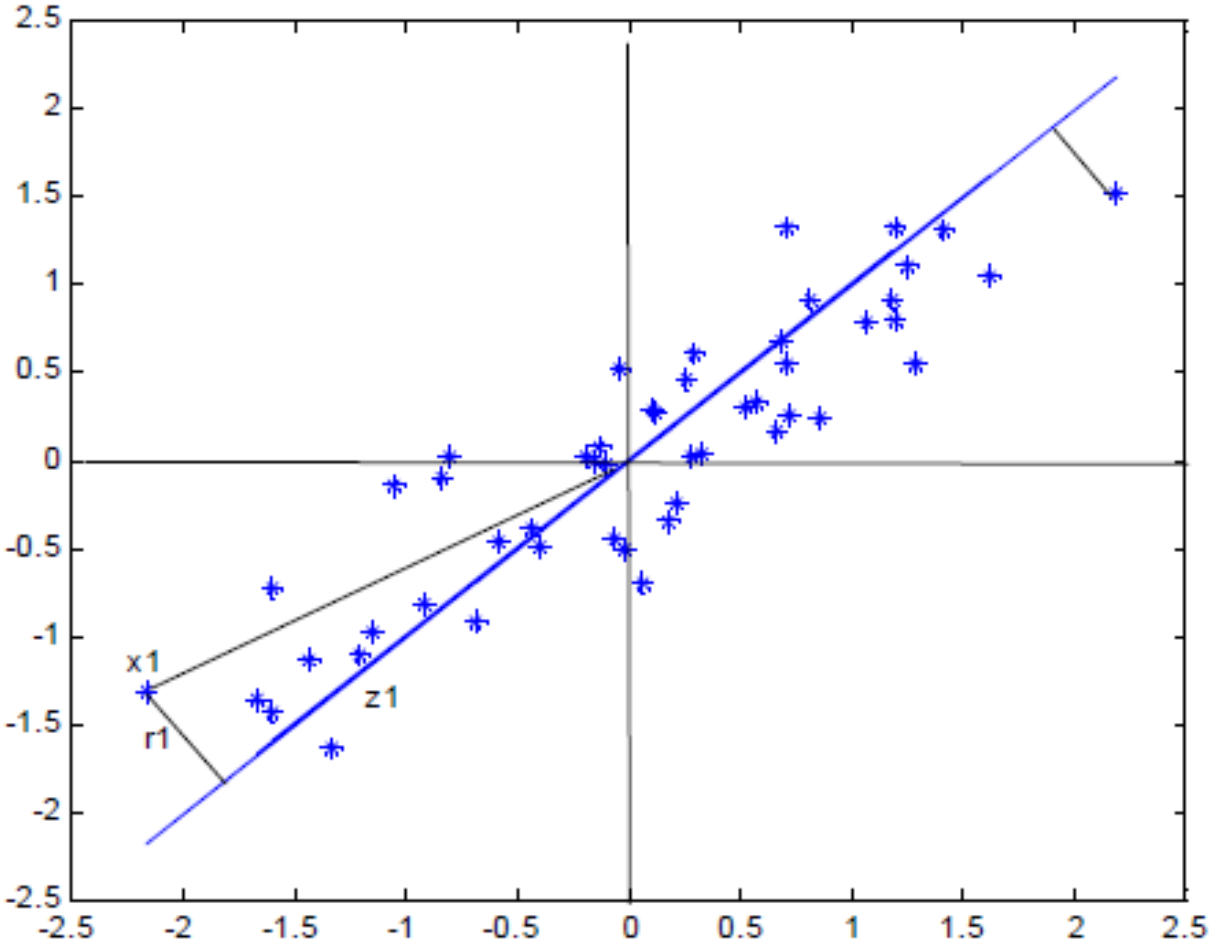
Apresentação do problema

O vetor que representa esta projeção será $z_i a_1$.
Se r_i representa a distância entre x_i e sua projeção sobre a_1, implica:

$$\text{minimizar} \sum_{i=1}^{n} r_i^2 = sum_{i=1}^{n} |x_i - z_i a_1|^2$$

onde $|u|$ é a norma euclideana do vetor u.

Exemplo da reta que minimiza as distâncias ortogonais



Exemplo

Da figura vemos que ao projetar cada ponto sobre a reta forma-se um triângulo retângulo, onde a hipotenusa é a distância do ponto à origem $(x^{`}x)^{1/2}$ e os catetos a projeção do ponto sobre a reta (z_i) e a distância entre o ponto e a sua projeção (r_i).

$$x_i^{`}x_i = z_i^2 + r_i^2$$

somando para todos os pontos

$$\sum_{i=1}^{n} x_i^{`}x_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2$$

Exemplo

como $\sum_{i=1}^{n} x_i^{`} x_i$ é constante, minimizar $\sum_{i=1}^{n} r_i^2$ é equivalente a maximizar $\sum_{i=1}^{n} z_i^2$

enfoque estatístico

Representar pontos p dimensionales com perda de informação mínima é equivalente a substituir as p variáveis originais por uma nova variável z_1. Isto supõe que a nova variável terá maxima correlação com as originais, ou em outras palavras, permitir prever as variáveis originais com a máxima precisão.
Na figura, a linha desenhada não é a linha de regressão de alguma das variáveis com respeito de outra (obtida ao minimizar as distâncias verticais ou horizontais) senão a linha que minimiza as distâncias ortogonais ou entre os pontos e a reta.
Este enfoque extende-se para obter o melhor subespaço resumem (dimensão 2), obtida calculando-se o plano que melhor aproxima os pontos.

enfoque estatístico

O problema consiste em encontrar uma nova direção definida por um vetor unitário a_2 (ortogonal a a_1) e que verifique a condição de que a projeção de um ponto sobre o eixo maximize as distâncias entre os pontos projetados.
Isto equivale a encontrar uma segunda variável z_2 não correlacionada com z_1 e que tenha variância máxima.
Em geral, a componente $z_r (r < p)$ terá variância máxima entre todas as combinações lineares com a condição de estar não correlacionada com as $z_1, \ldots, z_{r-1}$ anteriores

enfoque geométrico

Se observarmos a nuvem de pontos, é possível ver que os pontos sitúam-se seguindo uma elipse e podem ser descritos pela projeção na direção do maior eixo da elipse.

Outline

Primeira componente

É definida como a C.L. das variáveis originais que tem a maior variância.

$$z_1 = Xa_1$$

z_1 tem média nula e variância:

$$\frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X' X a_1 = a_1' S a_1$$

Sendo S a matriz de variâncias e covariâncias.
a solução das equações para obter a primeira componente, passam por definições matemáticas (que não pertencem ao escopo desta disciplina). A representação para a variância de z_1 é:

$$Sa_1 = \lambda a_1$$

com

$$a_1' S a_1 = \lambda a_1' a_1 = \lambda$$

Segunda componente

Obtem-se o melhor plano da projeção das variáveis X. O objetivo é maximizar a soma das variâncias de $z_1 = Xa_1$ e $z_2 = Xa_2$, com a_i vetores que definem o plano.
A solução do sistema é:

$$Sa_1 = \lambda_1 a_1$$

$$Sa_2 = \lambda_2 a_2$$

Generalização

para um espaço de dimensão r, que melhor represente os pontos, calculamos os r autovetores associados aos r maiores autovalores.

$$|S - \lambda I| = 0$$

associados aos vetores:

$$(S - \lambda_i I)a_i = 0$$

Outline

Propriedades

- 1. Conservam a variabilidade inicial: a soma das variâncias das componentes é igual à soma das variâncias das variáveis originais
- 2. A proporção da variabilidade explicada por um componente é o quociente entre a sua variância, o autovalor associado ao autovetor que o define e a soma dos autovalores da matriz.
- 3. As covariâncias entre cada componente e as variáveis X são dadas pelo produto das coordenadas do autovetor que define a componente pelo seu autovalor:

$$Cov(z_i; x_1, ..., x_p) = \lambda_i a_1 = (\lambda_i a_{i1}, ..., \lambda_i a_{ip})$$

Propriedades

- 4. A correlação entre um componente principal e uma variável X é proporcional ao coeficiente dessa variável na definição da componente.

$$Corr(z_i; x_j) = a_{ij} \frac{\sqrt{\lambda_i}}{s_j}$$

- 5. As r componentes principais ($r < p$) proporcionam a predição linear ótima com r variáveis do conjunto de variáveis X.
- 6. Se padronizarmos as componentes principais, dividindo cada uma delas pelo desvio padrão, obtemos a padronização multivariada dos dados originais.

Exemplo de padronização das componentes

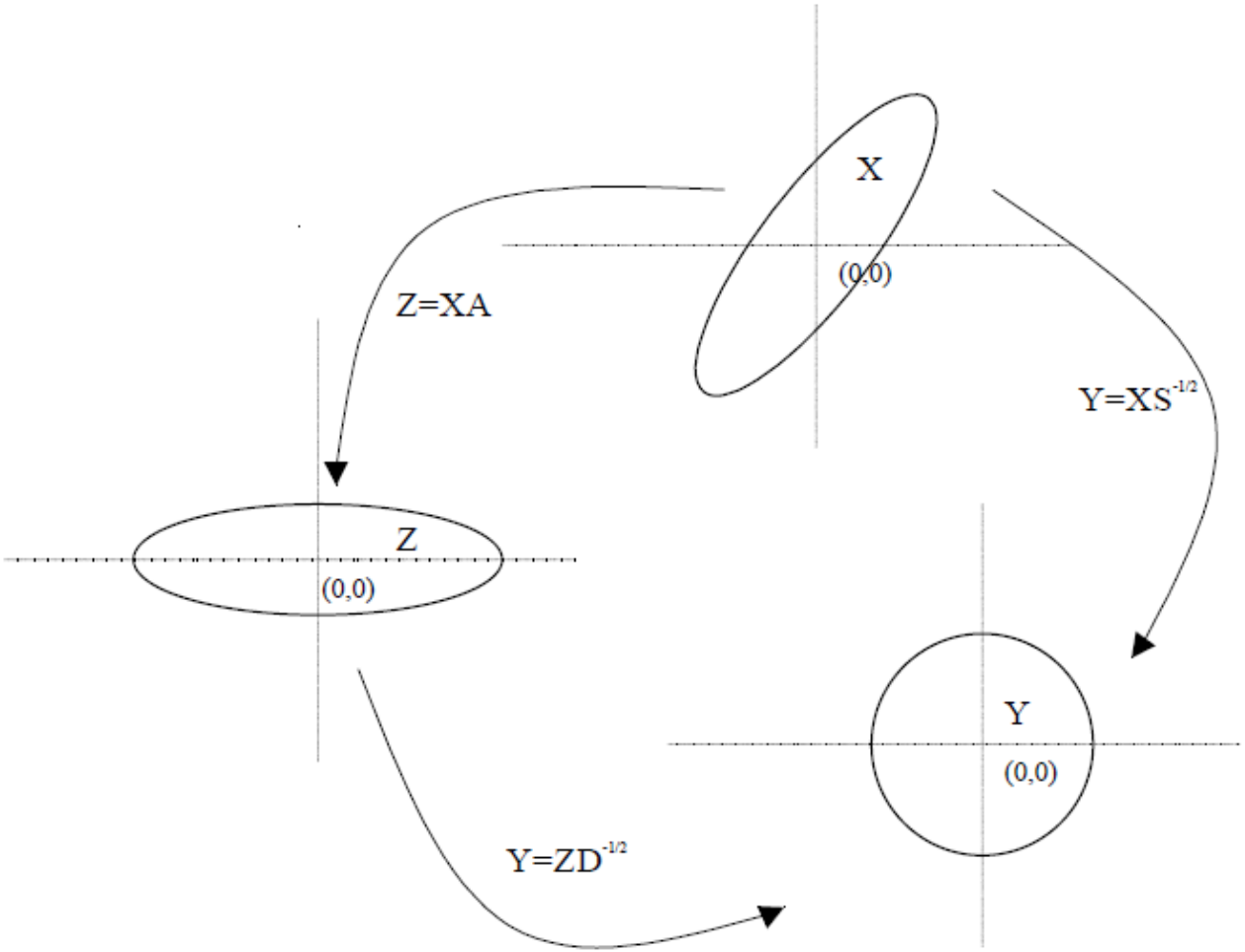


Figura: componentes

Interpretação das componentes

Cuando existe alta correlação positiva entre as variáveis, a primeira componente terá todas as coordenadas com o mesmo sinal (pode ser interpretada como uma média ponderada de todas as variáveis).
As componentes restantes são os fatores de "forma", tendo coordenadas positivas e negativas, que implicam que contrapõem uns grupos de variáveis frente a outros.

Seleção do número de componentes

- 1. desenhar um gráfico de λ_i vs i. Começe selecionando componentes até que as restantes tenham os valores de λ_i aproximadamente iguais.
- 2. Selecionar componentes até cubrir uma proporção determinada da variância (80 ou 90 % por exemplo).
- 3. Rejeitar as componentes associados a autovalores inferiores a um certo limite, que normalmente é dado por $\sum \lambda_i/p$. Quando trabalha-se com a matriz de correlação, o valor médio das componentes é 1.

Dados atípicos

Antes de obter as componentes principais, é conveniente garantir a não existência de dados atípicos, pois estes podem distorcionar a matriz de covariâncias.
A ACP pode ser também utilizada para detetar dados atípicos multivariados, dado que um valor muito extremo levará uma componente principal e aparecerá extremo nessa componente.