Motivação
Análise de componentes principais (PCA)

Prof. Cibele Russo

Contexto

$\underset{\sim}{X}$ é um vetor aleatório de dimensão $p \times 1$, isto é,

$$\underset{\sim}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix},$$

em que cada X_i é uma variável aleatória com média μ_i e variância σ_i^2, para $i = 1, \ldots, p$.

Contexto

A covariância e a correlação entre duas variáveis X_i e X_j, com $i, j = 1, \ldots, p$ e $i \neq j$ são dadas, respectivamente, por

$$\text{Cov}(X_i, X_j) = \sigma_{ij} \text{ e } \text{Cor}(X_i, X_j) = \rho_{ij}.$$

Obs: $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_i^2}\sqrt{\sigma_j^2}}$.

Contexto

O vetor de médias (populacionais) de $\underset{\sim}{X}$ é $\underset{\sim}{\mu}_{p\times 1}$ e a matriz de variâncias e covariâncias (populacionais) de $\underset{\sim}{X}$ é $\Sigma_{p\times p}$, em que

$$\underset{\sim}{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \text{ e } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_p^2 \end{bmatrix}$$

Obs: Note que Σ é simétrica pois $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, para $i, j = 1, \dots, p$.

Contexto

Estamos particularmente interessados no caso em que as variáveis $X_1, \ldots, X_p$ estão correlacionadas, isto é, algumas (ou muitas) das covariâncias $\mathrm{Cov}X_i, X_j, i, j = 1, \ldots, p$ e $i \neq j$ são não-nulas.

Quando isso ocorre, significa que existe **redundância entre dimensões**, e podemos ter interesse em **reduzir a dimensionalidade do problema**, construindo novas variáveis, não correlacionadas entre si, que sejam combinações lineares das variáveis originais.

Pode ser que poucas (k) novas variáveis expliquem grande parte da variabilidade existente nas (p) variáveis originais. Isso pode significar a **redução de custos** como tempo computacional e espaço para armazenamento de dados.

Contexto

Estamos particularmente interessados no caso em que as variáveis $X_1, \ldots, X_p$ estão correlacionadas, isto é, algumas (ou muitas) das covariâncias $\mathrm{Cov} X_i, X_j, i, j = 1, \ldots, p$ e $i \neq j$ são não-nulas.

Quando isso ocorre, significa que existe **redundância entre dimensões**, e podemos ter interesse em **reduzir a dimensionalidade do problema**, construindo novas variáveis, não correlacionadas entre si, que sejam combinações lineares das variáveis originais.

Pode ser que poucas (k) novas variáveis expliquem grande parte da variabilidade existente nas (p) variáveis originais. Isso pode significar a **redução de custos** como tempo computacional e espaço para armazenamento de dados.

Contexto

Estamos particularmente interessados no caso em que as variáveis $X_1, \ldots, X_p$ estão correlacionadas, isto é, algumas (ou muitas) das covariâncias $\mathrm{Cov} X_i, X_j, i, j = 1, \ldots, p$ e $i \neq j$ são não-nulas.

Quando isso ocorre, significa que existe **redundância entre dimensões**, e podemos ter interesse em **reduzir a dimensionalidade do problema**, construindo novas variáveis, não correlacionadas entre si, que sejam combinações lineares das variáveis originais.

Pode ser que poucas (k) novas variáveis expliquem grande parte da variabilidade existente nas (p) variáveis originais. Isso pode significar a **redução de custos** como tempo computacional e espaço para armazenamento de dados.

Componentes principais

Seja $\underset{\sim}{X}$ um vetor aleatório com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os autovalores de Σ, com autovetores correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, tais que

1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)^\top$.

Então $\underset{\sim}{Y}_{p \times 1} = O^\top \underset{\sim}{X}$ é o **vetor de componentes principais** de Σ.

Componentes principais

Seja $\underset{\sim}{X}$ um vetor aleatório com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os autovalores de Σ, com autovetores correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, tais que

1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)^\top$.

Então $\underset{\sim}{Y}_{p \times 1} = O^\top \underset{\sim}{X}$ é o **vetor de componentes principais** de Σ.

Componentes principais

Seja $\underset{\sim}{X}$ um vetor aleatório com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os autovalores de Σ, com autovetores correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, tais que

1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)^\top$.

Então $\underset{\sim}{Y}_{p \times 1} = O^\top \underset{\sim}{X}$ é o **vetor de componentes principais** de Σ.

Componentes principais

Seja $\underset{\sim}{X}$ um vetor aleatório com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os autovalores de Σ, com autovetores correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, tais que

1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)^\top$.

Então $\underset{\sim}{Y}_{p \times 1} = O^\top \underset{\sim}{X}$ é o **vetor de componentes principais** de Σ.

Componentes principais

Propriedades:

1. A j-ésima componente principal de Σ é dada por
$$Y_j = \underset{\sim}{e_j}^\top \underset{\sim}{X}.$$
2. $\mathrm{E}(Y_j) = \underset{\sim}{e_j}^\top \underset{\sim}{\mu}$.
3. $\mathrm{Var}(Y_j) = \underset{\sim}{e_j}^\top \Sigma \underset{\sim}{e_j} = \lambda_j$.
4. $\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cor}(Y_i, Y_j) = 0$ para $i, j = 1, \ldots, p$ e $i \neq j$.
5. A proporção da variância total de $\underset{\sim}{X}$ que é explicada pela j-ésima componente principal é
$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

Componentes Principais

Observações:

1. Com a matriz de variâncias e covariâncias amostrais, estima-se os componentes principais da matriz de variâncias e covariâncias populacionais.

2. O mesmo desenvolvimento pode ser feito para a matriz de correlações, com a vantagem de que as componentes principais são menos influenciadas pela magnitude das variâncias.

Componentes Principais

Observações:

1. Com a matriz de variâncias e covariâncias amostrais, estima-se os componentes principais da matriz de variâncias e covariâncias populacionais.

2. O mesmo desenvolvimento pode ser feito para a matriz de correlações, com a vantagem de que as componentes principais são menos influenciadas pela magnitude das variâncias.

Motivação 1

Estudos mostram que grande parte de adultos e adolescentes norte-americanos usam regularmente substâncias psicoativas. Em um destes estudos (Huba et al. 1981, J. of Personality and Social Psychology), dados foram coletados de 1634 estudantes na área metropolitana de Los Angeles. Cada participante completou um questionário informando o número de vezes que cada item foi usado.

Os itens são os seguintes: cigarro, cerveja, vinho, licor, cocaína, tranquilizantes, medicamentos, heroína, maconha, haxixe, inalantes, alucinógenos e anfetaminas.

Motivação 1

Estudos mostram que grande parte de adultos e adolescentes norte-americanos usam regularmente substâncias psicoativas. Em um destes estudos (Huba et al. 1981, J. of Personality and Social Psychology), dados foram coletados de 1634 estudantes na área metropolitana de Los Angeles. Cada participante completou um questionário informando o número de vezes que cada item foi usado.

Os itens são os seguintes: cigarro, cerveja, vinho, licor, cocaína, tranquilizantes, medicamentos, heroína, maconha, haxixe, inalantes, alucinógenos e anfetaminas.

Motivação 1

As respostas foram registradas em uma escala de cinco pontos: 1. nunca experimentei, 2. apenas uma vez, 3. poucas vezes, 4. muitas vezes e 5. regularmente.

Os comandos para análises se encontram no arquivo Motivação1.R.

Deseja-se obter as componentes principais a matriz de correlações dos dados, e propor um “índice de utilização de substâncias psicoativas".

Motivação 1

As respostas foram registradas em uma escala de cinco pontos: 1. nunca experimentei, 2. apenas uma vez, 3. poucas vezes, 4. muitas vezes e 5. regularmente.

Os comandos para análises se encontram no arquivo Motivação1.R.

Deseja-se obter as componentes principais a matriz de correlações dos dados, e propor um “índice de utilização de substâncias psicoativas".

Motivação 1

As respostas foram registradas em uma escala de cinco pontos: 1. nunca experimentei, 2. apenas uma vez, 3. poucas vezes, 4. muitas vezes e 5. regularmente.

Os comandos para análises se encontram no arquivo Motivação1.R.

Deseja-se obter as componentes principais a matriz de correlações dos dados, e propor um “índice de utilização de substâncias psicoativas".

Motivação 2

O conjunto de dados "Educação"(Fonte: Ipeadata) mostra dados educacionais por estado de acordo com a descrição a seguir.

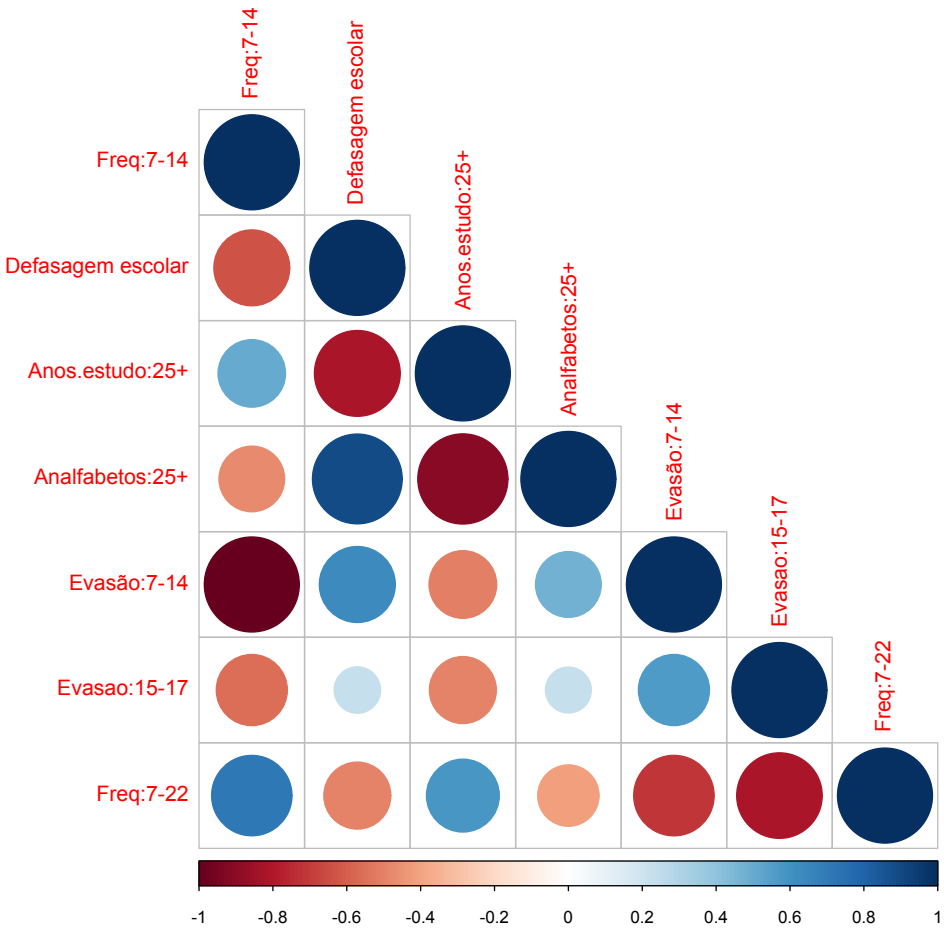
- Sigla (do estado)
- Estado
- X1: Frequência escolar de pessoas com 7 a 14 anos
- X2: Defasagem escolar em mais de 1 ano de atraso de pessoas com 7 a 14 anos (2000)
- X3: Anos de estudo de pessoas com 25 anos ou mais (2000)
- X4: Analfabetos com 25 anos ou mais (2000)
- X5: Evasão escolar de pessoas com 7 a 14 anos (2000)
- X6: Evasão escolar de pessoas com 15 a 17 anos (2000)
- X7: Frequência escolar de pessoas com 7 a 22 anos

Os comandos para análises se encontram no arquivo Motivação2.R

Matriz de correlações

	Freq:7-14	Defasagem escolar	Anos.estudo:25+	Analfabetos:25+	Evasão:7-14	Evasao:1
Freq:7-14	1.00	-0.64	0.51	-0.48	-1.00	-
Defasagem escolar	-0.64	1.00	-0.82	0.89	0.64	
Anos.estudo:25+	0.51	-0.82	1.00	-0.90	-0.51	-
Analfabetos:25+	-0.48	0.89	-0.90	1.00	0.48	
Evasão:7-14	-1.00	0.64	-0.51	0.48	1.00	
Evasao:15-17	-0.56	0.24	-0.50	0.23	0.56	
Freq:7-22	0.72	-0.50	0.59	-0.42	-0.72	-

Gráfico de correlação



Aplicação 2

1. Deseja-se criar uma ordenação pela primeira componente principal da matriz de variâncias e covariâncias.
2. Quanto da variabilidade total dos dados é explicado pela primeira componente principai?