SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

Aula 7a: **Análise de Componentes Principais**

Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele

Baseado em Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Prentice Hall.

Análise de componentes principais

A Análise de componentes principais (ACP ou PCA, de *principal component analysis*) é uma técnica que **transforma linearmente** um conjunto de p **variáveis correlacionadas** em um conjunto de k **variáveis não correlacionadas** (com $k < p$), que explicam uma parcela substancial das informações do conjunto original.

A ACP foi proposta por Karl Pearson em 1901 e estudada e nomeada por Harold Hotelling em 1933.

Se as variáveis originais tiverem distribuição normal multivariada, as componentes principais também terão distribuição normal multivariada e serão independentes.

Análise de componentes principais

A Análise de componentes principais (ACP ou PCA, de *principal component analysis*) é uma técnica que **transforma linearmente** um conjunto de p **variáveis correlacionadas** em um conjunto de k **variáveis não correlacionadas** (com $k < p$), que explicam uma parcela substancial das informações do conjunto original.

A ACP foi proposta por Karl Pearson em 1901 e estudada e nomeada por Harold Hotelling em 1933.

Se as variáveis originais tiverem distribuição normal multivariada, as componentes principais também terão distribuição normal multivariada e serão independentes.

Análise de componentes principais

A Análise de componentes principais (ACP ou PCA, de *principal component analysis*) é uma técnica que **transforma linearmente** um conjunto de p **variáveis correlacionadas** em um conjunto de k **variáveis não correlacionadas** (com $k < p$), que explicam uma parcela substancial das informações do conjunto original.

A ACP foi proposta por Karl Pearson em 1901 e estudada e nomeada por Harold Hotelling em 1933.

Se as variáveis originais tiverem distribuição normal multivariada, as componentes principais também terão distribuição normal multivariada e serão independentes.

Objetivos principais

- Reduzir a dimensionalidade dos dados.
- Obter combinações interpretáveis das variáveis originais.
- Descrever e compreender a estrutura de correlação das variáveis originais.

Começaremos com a obtenção das **componentes principais exatas**, extraídas da matriz de variâncias e covariâncias populacionais, Σ, e depois veremos as **componentes principais estimadas**, obtidas da matriz de variâncias e covariâncias amostrais, S, quando Σ é desconhecida.

Objetivos principais

- Reduzir a dimensionalidade dos dados.
- Obter combinações interpretáveis das variáveis originais.
- Descrever e compreender a estrutura de correlação das variáveis originais.

Começaremos com a obtenção das **componentes principais exatas**, extraídas da matriz de variâncias e covariâncias populacionais, Σ, e depois veremos as **componentes principais estimadas**, obtidas da matriz de variâncias e covariâncias amostrais, S, quando Σ é desconhecida.

Objetivos principais

- Reduzir a dimensionalidade dos dados.
- Obter combinações interpretáveis das variáveis originais.
- Descrever e compreender a estrutura de correlação das variáveis originais.

Começaremos com a obtenção das **componentes principais exatas**, extraídas da matriz de variâncias e covariâncias populacionais, Σ, e depois veremos as **componentes principais estimadas**, obtidas da matriz de variâncias e covariâncias amostrais, S, quando Σ é desconhecida.

Objetivos principais

- Reduzir a dimensionalidade dos dados.
- Obter combinações interpretáveis das variáveis originais.
- Descrever e compreender a estrutura de correlação das variáveis originais.

Começaremos com a obtenção das **componentes principais exatas**, extraídas da matriz de variâncias e covariâncias populacionais, Σ, e depois veremos as **componentes principais estimadas**, obtidas da matriz de variâncias e covariâncias amostrais, S, quando Σ é desconhecida.

Contexto

Seja $\underset{\sim}{X}$ um vetor aleatório de dimensão $p \times 1$ com vetor de médias (populacionais) $\underset{\sim}{\mu}_{p\times 1}$ e matriz de variâncias e covariâncias (populacionais) de $\Sigma_{p\times p}$.

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

Contexto

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.

Análise de componentes principais

Seja $\underset{\sim}{X} \sim (\underset{\sim}{\mu}, \Sigma)$.

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

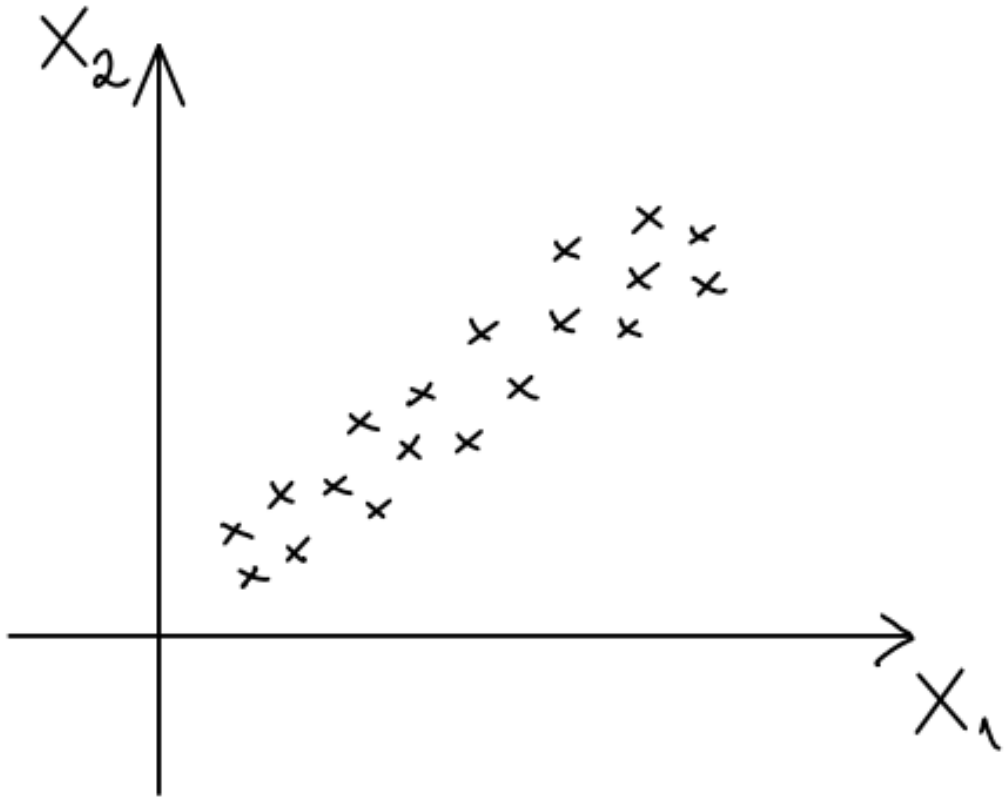
1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

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

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

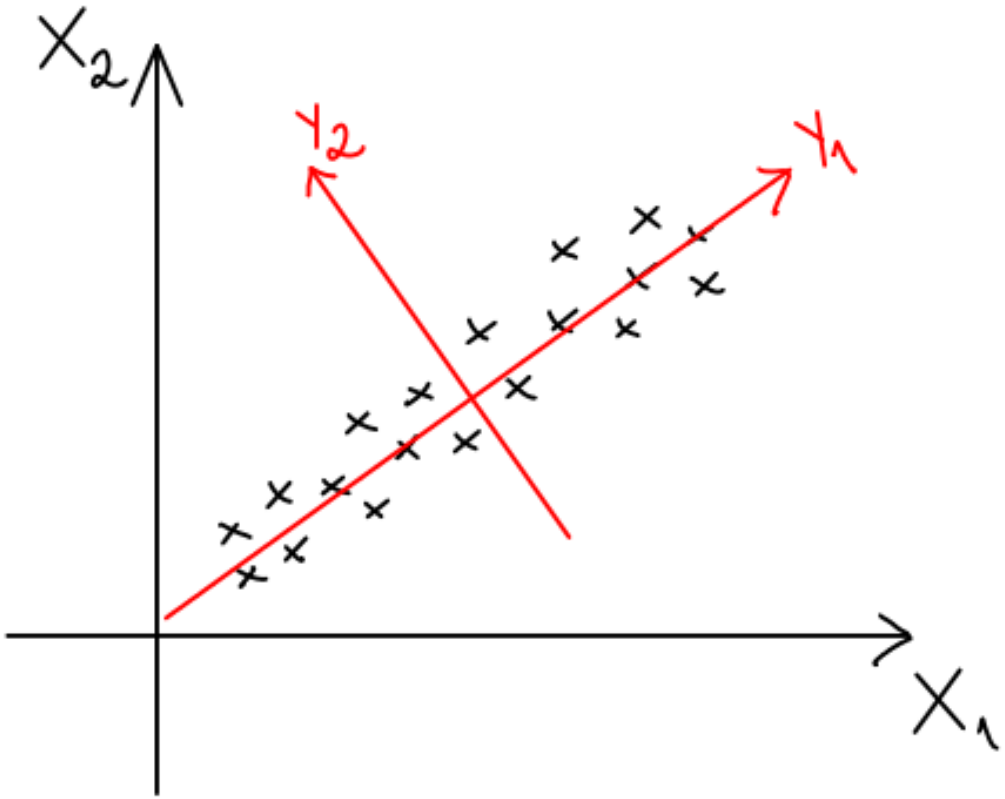
Análise de componentes principais

Interpretação geométrica ($p = 2$)



Análise de componentes principais

Interpretação geométrica ($p = 2$)



Análise de 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

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}.$$

Estimação das componentes principais

Como em geral a matriz Σ é desconhecida, utiliza-se a matriz S, de variâncias e covariâncias amostrais, para estimar as componentes principais.

Considere $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ os autovalores de S, com autovetores correspondentes padronizados $\hat{\underset{\sim}{e}}_1, \ldots, \hat{\underset{\sim}{e}}_p$.

A j-ésima componente principal amostral é dada por

$$\hat{Y}_j = \hat{\underset{\sim}{e}}_j^\top \underset{\sim}{X}.$$

Estimação das componentes principais

Propriedades:

1. $\widehat{\text{Var}}(\hat{Y}_j) = \hat{\lambda}_j$.
2. $\text{Cov}(\hat{Y}_i, \hat{Y}_j) = \text{Cor}(\hat{Y}_i, \hat{Y}_j) = 0$ para $i, j = 1, \ldots, p$ e $i \neq j$.
3. A proporção da variância total explicada pela j-ésima componente principal amostral é
$$\frac{\hat{\lambda}_j}{\sum_{i=1}^{p} \hat{\lambda}_i}$$
4. $\widehat{\text{Cor}}(\hat{Y}_j, X_i) = \dfrac{\hat{e}_{ji}\sqrt{\hat{\lambda}_j}}{\sqrt{s_{jj}}}$ (Exercício)

Estimação das componentes principais

5. Pelo teorema da decomposição espectral,

$$S_{p\times p} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{e}_j \hat{e}_j^\top$$

pode ser aproximada por

$$S_{p\times p} \approx \sum_{j=1}^{k} \hat{\lambda}_j \hat{e}_j \hat{e}_j^\top$$

Escores das componentes

É comum utilizar os **escores das componentes** para análises estatísticas ou ordenação (*ranking*) dos elementos amostrais. Esses escores são obtidos ordenando os valores obtidos de

$$\hat{Y}_j = \hat{e}_j^\top \underset{\sim}{X}.$$

Análise de componentes principais via matriz de correlação

Seja $\underset{\sim}{X}_{p\times 1} \sim (\underset{\sim}{\mu}, \Sigma)$, $\underset{\sim}{X} = (X_1, \ldots, X_p)^\top$.

Seja $\underset{\sim}{Z} = (Z_1, \ldots, Z_p)^\top$ tal que

$$Z_i = \frac{X_i - \mu_i}{\sigma_i},$$

em que $\mu_i = \mathsf{E}(X_i)$ e $\sigma_i^2 = \mathsf{Var}(X_i)$.

Temos que $\mathsf{Cov}(\underset{\sim}{Z}) = P = \mathsf{Cor}(\underset{\sim}{X})$.

Análise de componentes principais via matriz de correlação

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os autovalores de P, 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 $O_{p \times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)$.

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

Análise de componentes principais via matriz de correlação

Propriedades:

1. A j-ésima componente principal de P é dada por

$$Y_j = \underset{\sim}{e_j}^\top \underset{\sim}{Z}.$$

2. $\mathsf{E}(Y_j) = \underset{\sim}{0}$.
3. $\mathsf{Var}(Y_j) = \underset{\sim}{e_j}^\top P \underset{\sim}{e_j} = \lambda_j$.
4. $\mathsf{Cov}(Y_i, Y_j) = \mathsf{Cor}(Y_i, Y_j) = 0$ para $i, j = 1, \ldots, p$ e $i \neq j$.
5. A proporção da variância total de $\underset{\sim}{Z}$ que é explicada pela j-ésima componente principal é λ_j / p

Estimação das componentes principais

Como em geral a matriz P é desconhecida, utiliza-se a matriz R, de correlações amostrais, para estimar as componentes principais.

Considere $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ os autovalores de R, com autovetores correspondentes padronizados $\hat{\underset{\sim}{e}}_1, \ldots, \hat{\underset{\sim}{e}}_p$.

A j-ésima componente principal amostral é dada por

$$\hat{Y}_j = \hat{\underset{\sim}{e}}_j^\top \underset{\sim}{Z}.$$

Estimação das componentes principais

Propriedades:

1. $\widehat{\text{Var}}(\hat{Y}_j) = \hat{\lambda}_j$.
2. $\text{Cov}(\hat{Y}_i, \hat{Y}_j) = \text{Cor}(\hat{Y}_i, \hat{Y}_j) = 0$ para $i, j = 1, \ldots, p$ e $i \neq j$.
3. A proporção da variância total explicada pela j-ésima componente principal amostral é
$$\frac{\hat{\lambda}_j}{p}$$
4. $\widehat{\text{Cor}}(\hat{Y}_j, Z_i) = \hat{e}_{ji}\sqrt{\hat{\lambda}_j}$ (Exercício)

Estimação das componentes principais

5. Pelo teorema da decomposição espectral,

$$R_{p\times p} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{e}_j \hat{e}_j^\top$$

pode ser aproximada por

$$R_{p\times p} \approx \sum_{j=1}^{k} \hat{\lambda}_j \hat{e}_j \hat{e}_j^\top$$

Escores das componentes

É comum utilizar os **escores das componentes** para análises estatísticas ou ordenação (*ranking*) dos elementos amostrais. Esses escores são obtidos ordenando os valores obtidos de

$$\hat{Y}_j = \hat{e}_j^\top \underset{\sim}{Z}.$$

Determinação do número de componentes principais

Alguns métodos:

1. Proporção da variância total explicada.
2. Análise gráfica da variância explicada (scree-plot).
3. Aproximação da matriz S (ou R).
4. Análise prática das componentes principais.

Inferência assintótica sobre as componentes principais

Sejam $\underset{\sim}{X} \sim (\underset{\sim}{\mu}, \Sigma)$, $\lambda_1 \geq \ldots \geq \lambda_p > 0$ os autovalores de Σ, com autovetores correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_p > 0$ os autovalores de S, com autovetores correspondentes $\widehat{\underset{\sim}{e}}_1, \ldots, \widehat{\underset{\sim}{e}}_p$.

Sejam $\underset{\sim}{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$ e $\widehat{\underset{\sim}{\lambda}} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)^\top$.

Inferência assintótica sobre as componentes principais

Resultados:

1. Se $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$, então, para n suficientemente grande, temos

$$\sqrt{n}(\widehat{\underset{\sim}{\lambda}} - \underset{\sim}{\lambda}) \approx N_p(\underset{\sim}{0}, 2\Lambda^2).$$

2. Se $E_i = \lambda_i \sum_{k=1, k\neq i}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \underset{\sim}{e}_k \underset{\sim}{e}_k^\top$, então

$$\sqrt{n}(\widehat{\underset{\sim}{e}}_i - \underset{\sim}{e}_i) \approx N_p(\underset{\sim}{0}, E_i).$$

3. A distribuição de cada $\widehat{\lambda}_i$ não depende dos elementos de $\widehat{\underset{\sim}{e}}_i$ correspondentes.

Inferência assintótica sobre as componentes principais

Segue desses resultados que

$$\widehat{\lambda}_i \overset{ind}{\sim} N\left(\lambda_i, 2\frac{\lambda_i^2}{n}\right).$$

Assim,

$$P\left(|\widehat{\lambda}_i - \lambda_i| \leq z_{\alpha/2}\lambda_i\sqrt{\frac{2}{n}}\right) = 1 - \alpha.$$

Inferência assintótica sobre as componentes principais

Segue desses resultados que

$$\widehat{\lambda}_i \overset{ind}{\sim} N\left(\lambda_i, 2\frac{\lambda_i^2}{n}\right).$$

Assim,

$$P\left(|\widehat{\lambda}_i - \lambda_i| \leq z_{\alpha/2}\lambda_i\sqrt{\frac{2}{n}}\right) = 1 - \alpha.$$

Inferência assintótica sobre as componentes principais

Segue desses resultados que

$$\widehat{\lambda}_i \overset{ind}{\sim} N\left(\lambda_i, 2\frac{\lambda_i^2}{n}\right).$$

Assim,

$$P\left(|\widehat{\lambda}_i - \lambda_i| \leq z_{\alpha/2}\lambda_i\sqrt{\frac{2}{n}}\right) = 1 - \alpha.$$

Inferência assintótica sobre as componentes principais

E podemos então construir um intervalo com $100(1-\alpha)\%$ de confiança para λ_i usando propriedades da distribuição normal:

$$IC_{100(1-\alpha)\%}(\lambda_i) = \left(\frac{\widehat{\lambda}_i}{1 + z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n}}}, \frac{\widehat{\lambda}_i}{1 - z_{\frac{\alpha}{2}}\sqrt{\frac{2}{n}}} \right),$$

que também pode ser usado para fazer inferência sobre λ_i (por exemplo, avaliar se $H_0 : \lambda_i = \lambda_0$ contra $H_1 : \lambda_i \neq \lambda_0$).

Exemplo

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.

Exemplo

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.

A matriz de correlações das respostas encontra-se no arquivo Substancias.txt.

Foram obtidas as componentes principais a matriz de correlações dos dados. Como propor um “índice de utilização de substâncias psicoativas"?