

MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler
pavan@ime.usp.br

1º Semestre IME/2019

MAE5776

Vamos ver
😊

$$X_{n \times p} = (X_{ij}) \in \mathfrak{R}^{n \times p}$$

Matriz de Dados: Estatísticas descritivas multivariadas

- Definidas no espaço das colunas (p-vetores n-dimensionais): $\bar{X}_{p \times 1}, S_{p \times p}, R_{p \times p}, S_{p \times p}^{-1}$
- Definidas no espaço das linhas (n-vetores p-dimensionais): $D_{n \times n} = (d_{ij}^2), d_{Pij}^2, d_{Mij}^2$



Propriedades em Espaços Duais

$$X_{n \times p} \in \mathfrak{R}^{n \times p} \quad (X'X)_{p \times p} \in \mathfrak{R}^{p \times p} \quad (XX')_{n \times n} \in \mathfrak{R}^{n \times n}$$

Matriz Aleatória: Propriedades distribucionais

$$X_{n \times p} = (X_{ij}) \in \mathfrak{R}^{n \times p}; \quad X_{n \times p} \sim N_{n \times p}(\mathbf{1}_n \mu_{p \times 1}; \Omega_{np \times np} = \Psi_{n \times n} \otimes \Sigma_{p \times p}); \quad (\text{vec} X)_{np \times 1} \sim N_{np}(\mathbf{1}_n \otimes \mu_{p \times 1}; \Omega)$$

- Estimadores e Distribuições Amostrais sob $X_{i_{p \times 1}} \stackrel{iid}{\sim} N_p(\mu; \Sigma)$ $\bar{X}_{p \times 1}, S_{p \times p}$
- Episódios de Concentração e Regiões de Confiança
- MANOVA: Decomposição de Matrizes em $\mathfrak{R}^{p \times p}$

Intervalos de Confiança - Regiões de Confiança

(Everitt, 2007)

$$I.C(\mu_k) \text{ a } 100(1-\alpha)\%$$

$$= \left(\bar{X}_k \mp t_{n-1}(\alpha/2) \sqrt{\frac{s_{kk}}{n}} \right)$$

$$t^2 = \frac{(\bar{X} - \mu)^2}{s^2/n}$$

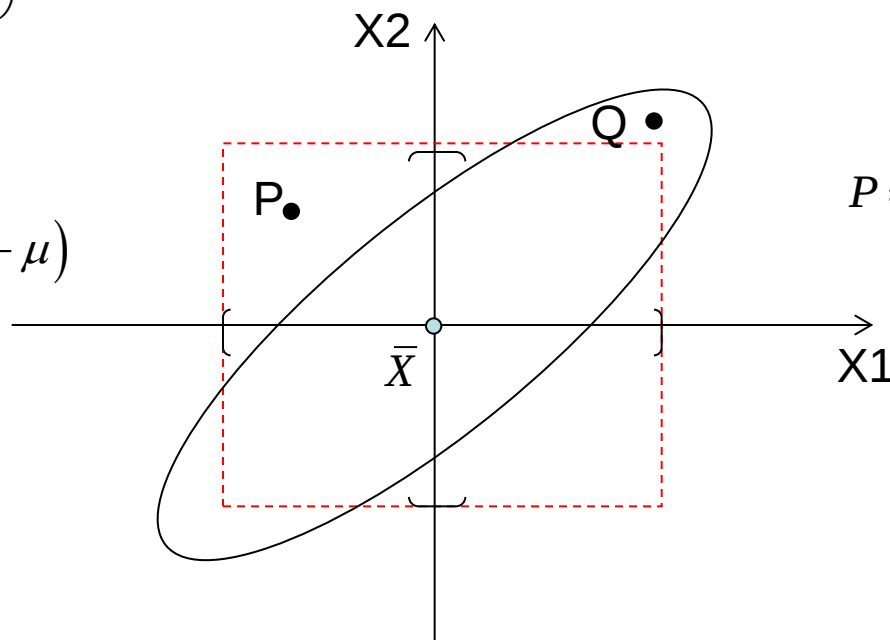
$$= n(\bar{X} - \mu)(s^2)^{-1}(\bar{X} - \mu)$$

$$t_{(n-1)}^2 = F_{1,(n-1)}$$

$$R(X) = \left\{ \mu; n(\bar{X} - \mu)' S^{-1}(\bar{X} - \mu) \leq \frac{(20-1)p}{(20-3)} F_{3,(17)}(\alpha) \right\}$$

$$T^2 = n(\bar{X} - \mu)' S^{-1}(\bar{X} - \mu)$$

$$P = (\mu_{P1}, \mu_{P2}); \quad Q = (\mu_{Q1}, \mu_{Q2})$$



⇒ Valores de $\mu_0 \in \mathbb{R}^2$ iguais a P e Q estão na região de aceitação ou de rejeição de possíveis valores do parâmetro μ ?

⇒ Elipse de concentração: $(X_i - \bar{X})' S^{-1}(X_i - \bar{X}) \leq \chi_2^2(\alpha = 0,10) = 4,61$

Outliers
Multivariados

MAE5776

$$X_{n \times p} \in \mathcal{R}^{n \times p}$$

n: unidade amostral
p: variáveis

Big/Little “n”

Big/Little “p”

Dimensão e Estrutura dos Dados

Casos Clássicos: $n > p$ e observações iid

Casos mais Gerais: $\left\{ \begin{array}{l} n \ll p \text{ e observações iid (soluções regularizadas, penalizadas)} \\ n \gg p \text{ e observações iid (soluções por amostragem, estatísticas} \right.$
suficientes, visualização de dados)

observações estruturadas: dependências (medidas repetidas, dados de famílias)

Representação BiPlot

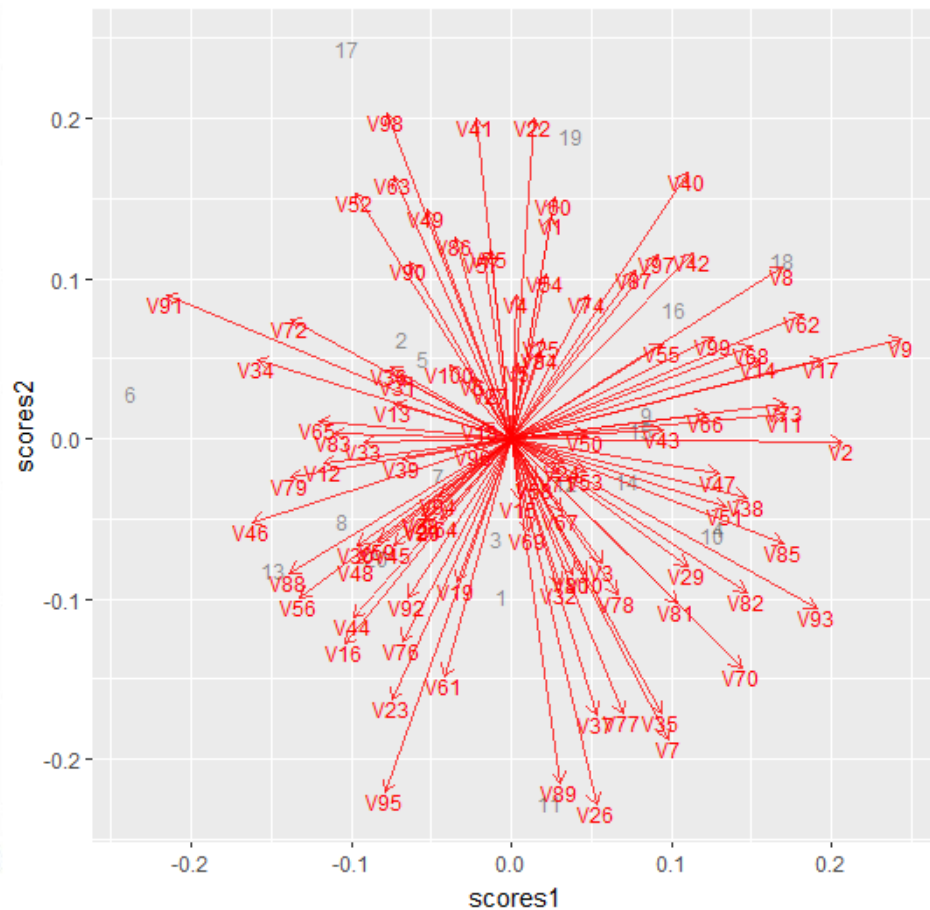
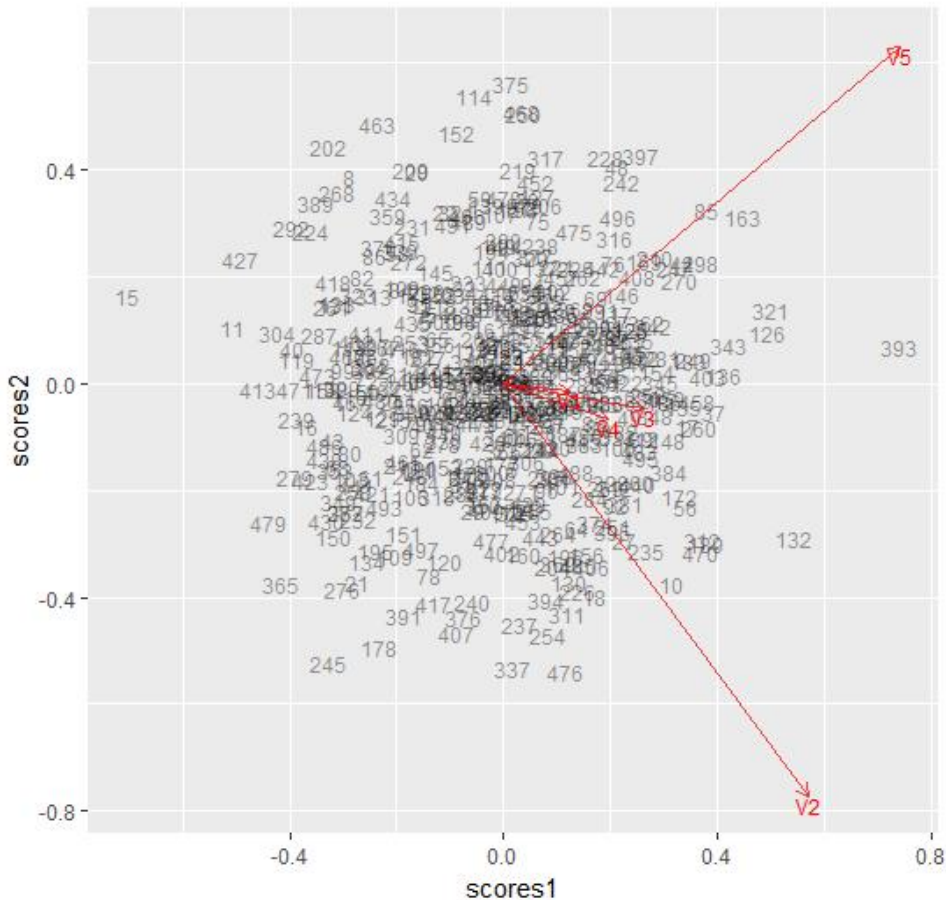
$$X_{n \times p} \in \mathbb{R}^{n \times p}$$

black screen problem
Esquemas de Re-Amostragem
Matloff, 2016 (Handbook of Big

Soluções Regularizadas e
Penalizadas
Hastie and Tibshirani (2015)

Big_n: n=500 p=5 (n >> p) Data)

Big_p: n=20 p=100 (n << p)



Redução de Dimensionalidade - $\mathbb{R}^{p \times p}$

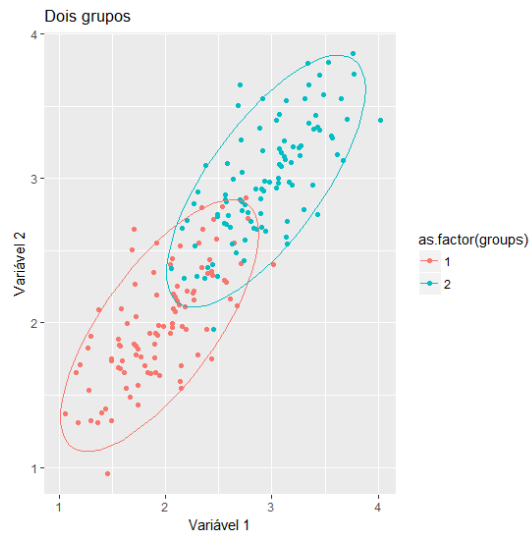
Obtenção de Eixos Principais (direções "ótimas")

$$X_{n \times p}; \quad \text{Cov}(X) = \Sigma_{p \times p}$$

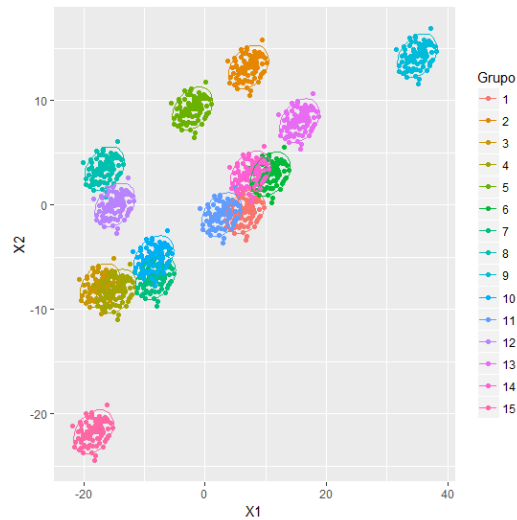
- Componentes Principais (PC): $\max_a f(\Sigma; a) = \max_a \frac{a' \Sigma a}{a' a}, \quad a' a = 1 \Rightarrow F_k = X a_k$
- Análise de Fatores Latentes: solução via PC $\Rightarrow Z_k = \lambda_k^{-1/2} F_k; \quad \Sigma = \Phi \Phi' + \Psi$
- Análise Discriminante (Linear de Fisher): $f(\Sigma_w^{-1} \Sigma_b; a) = \frac{a' \Sigma_b a}{a' \Sigma_w a}, \quad a' \Sigma_w a = 1$
- Análise de Correlação Canônica:
$$X_{n \times (p+q)} = \begin{pmatrix} X_{1 \ n \times p} & X_{2 \ n \times q} \end{pmatrix} \left\{ \begin{array}{l} f_1(\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}; a) = \frac{a' \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} a}{a' \Sigma_{11} a}, \quad a' \Sigma_{11} a = 1 \Rightarrow X_1 a \\ f_2(\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}; b) = \frac{b' \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} b}{b' \Sigma_{22} b}, \quad b' \Sigma_{22} b = 1 \Rightarrow X_2 b \end{array} \right.$$

Onde estão os vetores reducionistas?

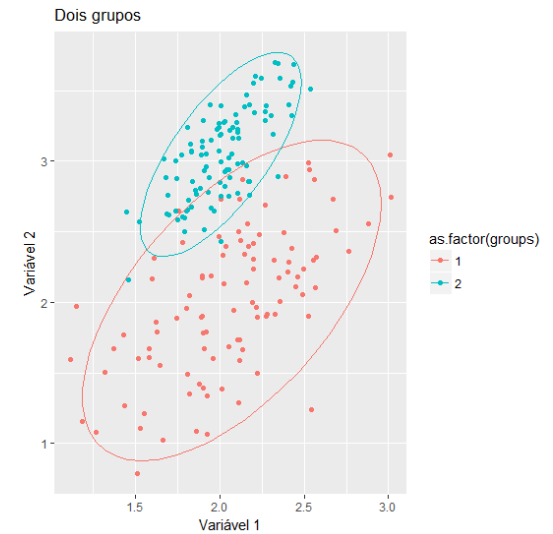
Exemplo 1



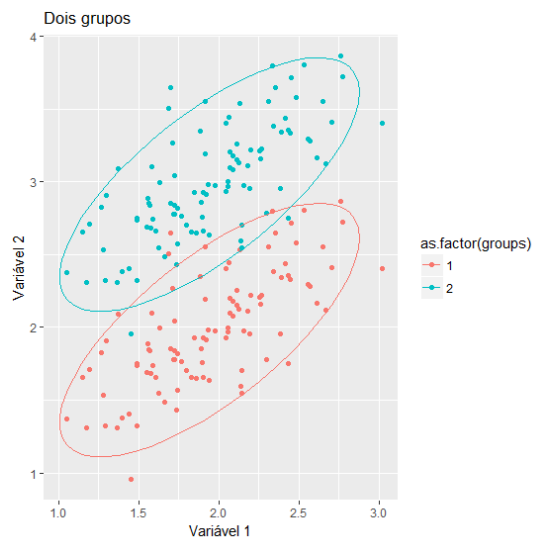
Exemplo 3



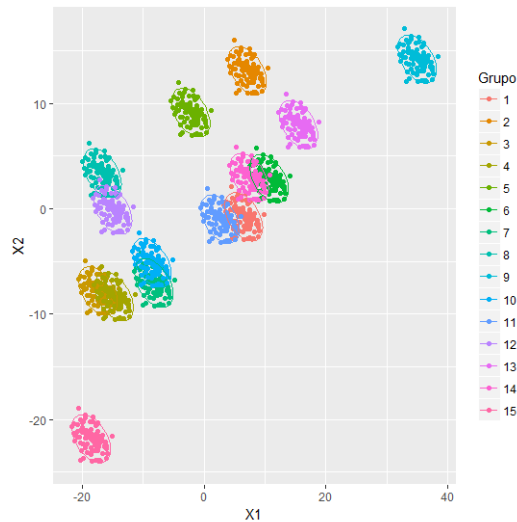
Exemplo 5



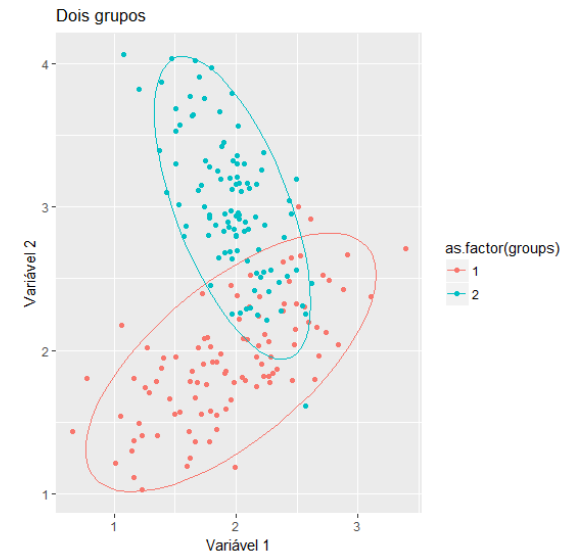
Exemplo 2



Exemplo 4



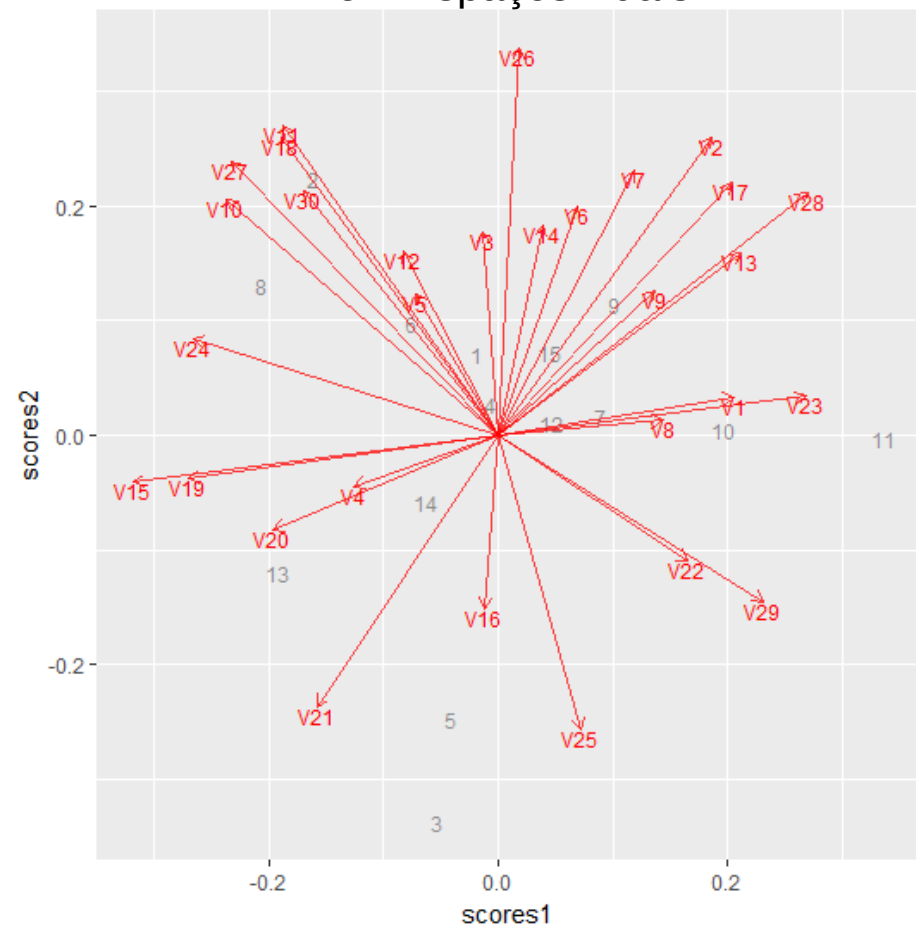
Exemplo 6



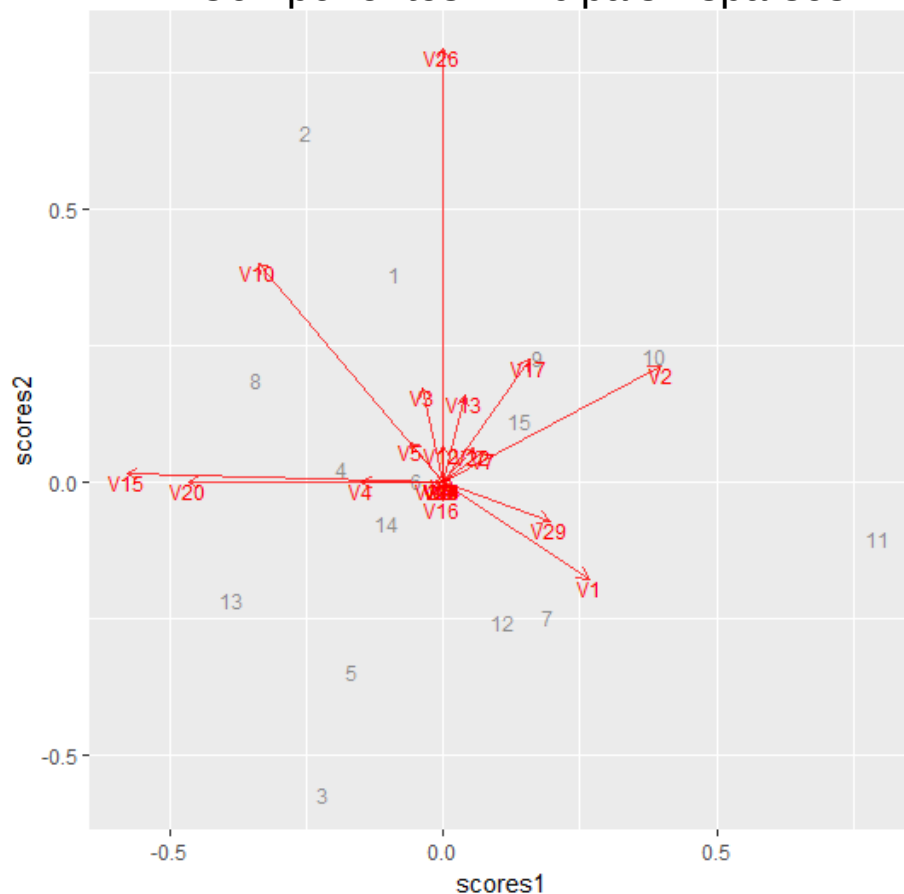
Componentes Principais – $n \ll p$

Representação Biplot: $n=15$ $p=30$

R-prcomp: Componentes Principais
em Espaços Duais



R-SPCA do pacote ElasticNet:
Componentes Principais Esparsos

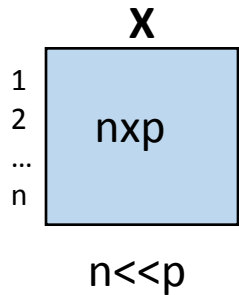


N-Integração de Bancos de Dados

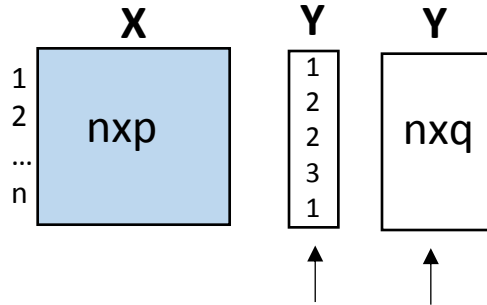
Integração ("interação") entre múltiplos níveis de informação avaliadas nas mesmas unidades amostrais

Um Único BD

Análise Não-Supervisionada



Análise Supervisionada

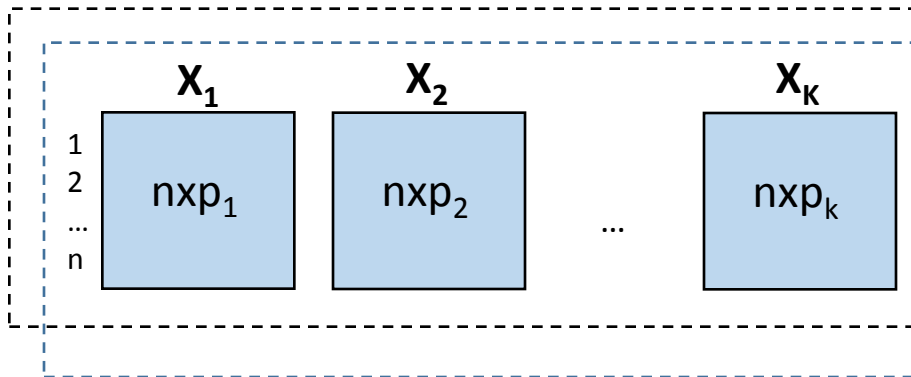


X: Matriz(es) de Dados

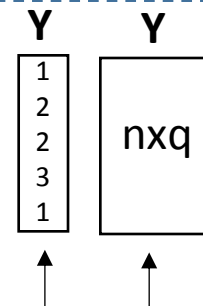
Y: Respostas de interesse
(Classes de doenças ou
Variáveis quantitativas)

Múltiplos BD (Multimodais, Multivisão)

Análise Não-Supervisionada



An. Supervisionada



P-Integração: mesmas variáveis em diferentes unidades amostrais (estudos multicentros)

Redução de Dimensionalidade e Integração de BD

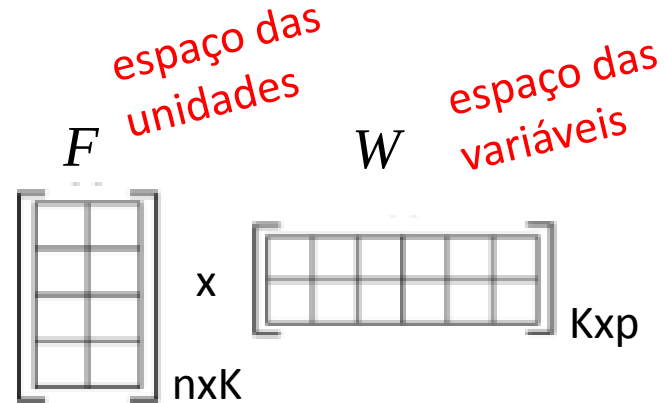
$X_{n \times p}$: Matriz de dados

Soluções via Fatoração do espaço $\mathcal{R}^{n \times p}$

$$X_{n \times p} = \boxed{U_{n \times n} D_r^{1/2} V'_{p \times p}} \quad \begin{matrix} = F \\ = W \end{matrix}$$

Matriz de Escores
(valores latentes)

Matriz de Cargas
(direções de projeção)



Soluções via Decomposição do Espaço do espaço $\mathcal{R}^{n \times p}$

Tabela de MANOVA: Decomposição em $\mathcal{R}^{p \times p}$

F.V.	g.l.	Matriz de SQPC
Trat	G-1	$H_{p \times p} = \sum_{g=1}^G n_g (\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'$
Resíduo	n-G	$E_{p \times p} = \sum_{g=1}^G \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_g)(y_{gi} - \bar{y}_g)'$
TOTAL	n-1	$H + E = \sum_{g=1}^G \sum_{i=1}^{n_g} (y_{gi} - \bar{y})(y_{gi} - \bar{y})'$

Decomposição em $\mathcal{R}^{n \times p}$

$$X_{n \times p} = H + E$$

Realizar as análises (de redução ou integração) em components de X

Aprendizado de Estruturas – Inferência Causal

$$X_{n \times p} \in \mathcal{R}^{n \times p}$$

Aprender/Estimar a Estrutura de Dependência
entre Variáveis $\mathcal{R}^{p \times p}$

Dados multivariados ($n > p$)

$$X_i \in \mathcal{R}^p \stackrel{iid}{\sim} (\mu; \Sigma)$$

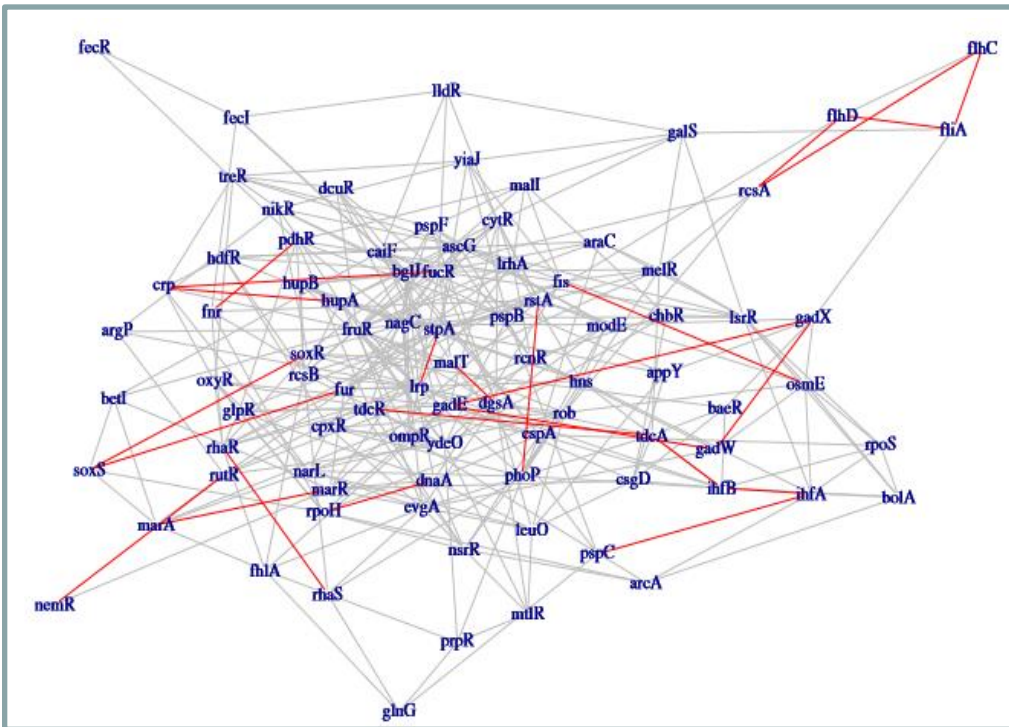
Σ diagonal
(independência)



Σ
Não Estruturada

“Aprender” a estrutura de dependência
entre variáveis:

- Grafo Não-Direcionado (UDG): Testes de correlação parcial
- Grafo Direcionado (DAG): orientar é descobrir V-estruturas ($X \rightarrow Z \leftarrow Y$)



Grafo Gaussiano de Independência Condicional
($p=87$ variáveis em *E. coli*.)