Sensores espectroscópicos e modelos de regressão aplicados na análise de solos

Aula 1 – Métodos de regressão linear

Me. José Vinícius Ribeiro

Universidade Estadual de Londrina

PÓS GRADUAÇÃO
FÍSICA UEL

SUMÁRIO

- Regressão (estatística)
- Linearidade
- Regressão linear simples
- Regressão linear múltipla
- Regressão linear por mínimos quadrados parciais
- Prática no python (colab ou vscode)

REGRESSÃO (ESTATÍSTICA)

Por que e quando utilizar regressão?

- Conceito: técnica estatística que busca modelar a relação entre uma **variável dependente** (a que queremos prever, classificar ou explicar) e uma ou mais **variáveis independentes**

REGRESSÃO (ESTATÍSTICA)

A ideia é encontrar uma função que consiga conectar, da melhor forma possível, X com y

$$y \rightarrow X$$

REGRESSÃO (ESTATÍSTICA)

No nosso contexto, um **modelo** pode ser conceituado como uma ferramenta matemática que descreve a relação entre uma ou mais variáveis independentes e uma variável dependente ou alvo, fornecendo **predições**, identificando tendências e/ou analisando o efeito de diferentes fatores no parâmetro alvo.

Modelo estatisticos: modelos que assumem uma estrutura funcional predefinida de dados, ou seja, são baseados em hipóteses sobre a distribuição de erros ou/e as relações entre variáveis

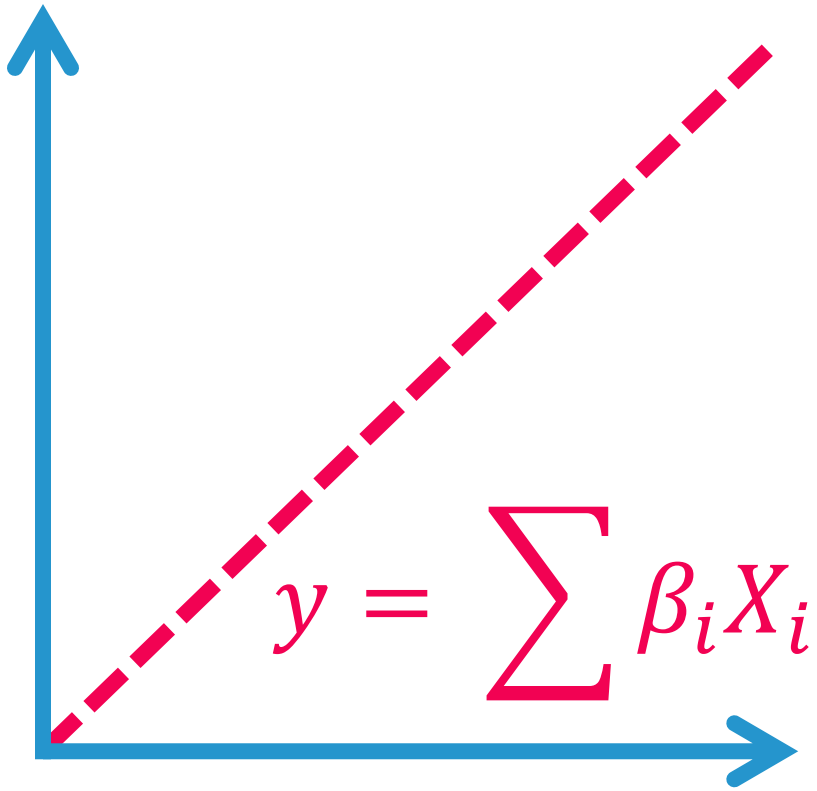
REGRESSÃO (ESTATÍSTICA)

Modelo de machine learning: modelos que aprendem a partir de **dados rotulados** e executam tarefas sem depender de suposições sobre a forma funcional das relações entre variáveis. Portanto, enfatizam o **desempenho preditivo** e se concentram na capacidade de generalização para novos dados, mesmo que às vezes as **custas da interpretabilidade.**

Deep learning: modelos que empregam **redes neurais artificiais** em sua estrutura de aprendizagem, que são projetados para descobrir padrões processando dados por meio de múltiplas camadas de unidades de processamento (**neurônios**)

LINEARIDADE

Modelo linear



$$Aditividade$$
$$f(x+y) = f(x) + f(y)$$

$$Homogeneidade$$
$$f(kx) = kf(x), \forall\, k \neq \mathbb{C}$$

A relação entre as variáveis independentes (matriz X) e as depententes (vetor y) pode ser expressa como uma combinação linear de variáveis. Apenas operações lineares.

PRINCIPAIS ALGORITMOS ATUALMENTE

Principais algoritmos de regressão linear

- Regressão Linear Simples
- Regressão Linear Múltipla
- Regressão Logística
- Regressão Linear por Mínimos Quadrados (PLS)
- Analise discriminante por Mínimos Quadrados (PLS-DA)
- Regressão de Ridge
- Análise Discriminante Linear (LDA)
- Máquina de Vetores de Suporte (kernel-linear)
- Redes Neurais Clássicas (função de ativação linear)

Regressão Linear Simples e Múltipla

Regressão Linear Simples

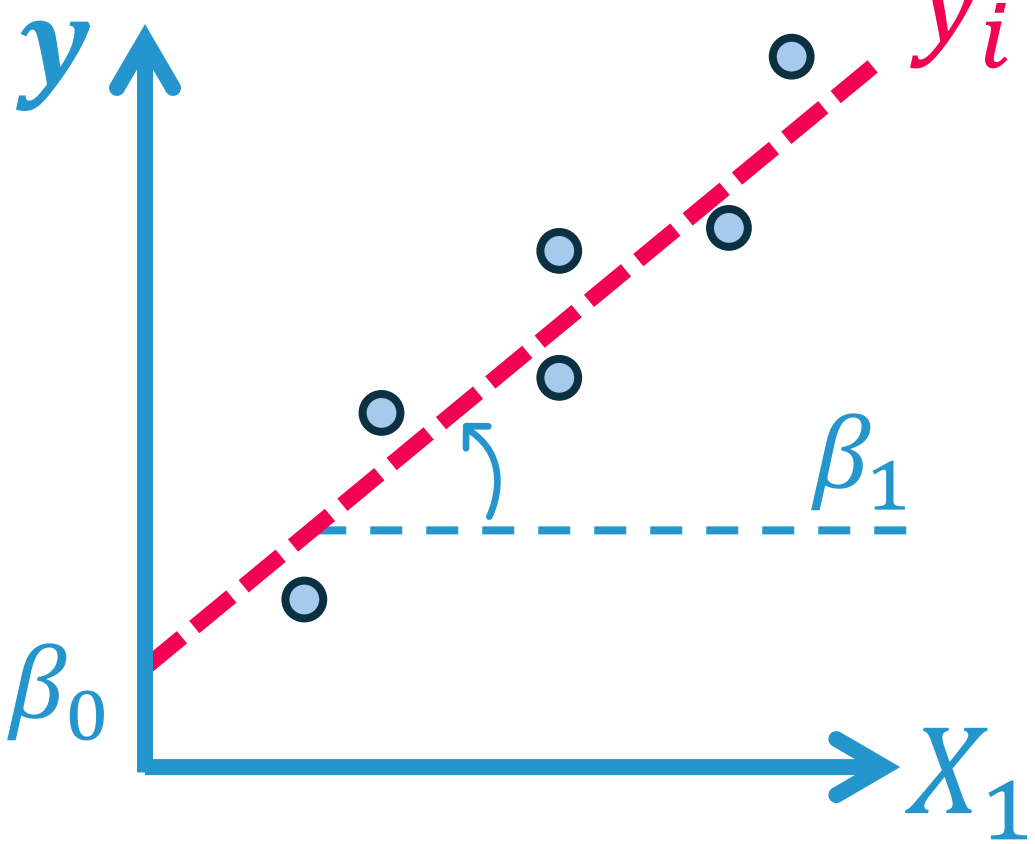
Notação: conjunto de *n* amostras medidas por *m* variáveis

$$y_i \rightarrow \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{bmatrix} \qquad X_{ij} \rightarrow \boldsymbol{X} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & X_{14} & \cdots & X_{1m} \\ X_{21} & X_{22} & X_{23} & X_{24} & \cdots & X_{2m} \\ X_{31} & X_{32} & X_{33} & X_{34} & \cdots & X_{3m} \\ \cdots & \cdots & \cdots & \cdots & & \cdots \\ X_{n1} & X_{n2} & X_{n3} & X_{n4} & \cdots & X_{nm} \end{bmatrix}$$

Regressão Linear

É um tipo de regressão que explora a linearidade entre um parâmetro de interesse e uma única variável preditora

$$\hat{y}_i = \beta_0 + \beta_1 X_{i1}$$



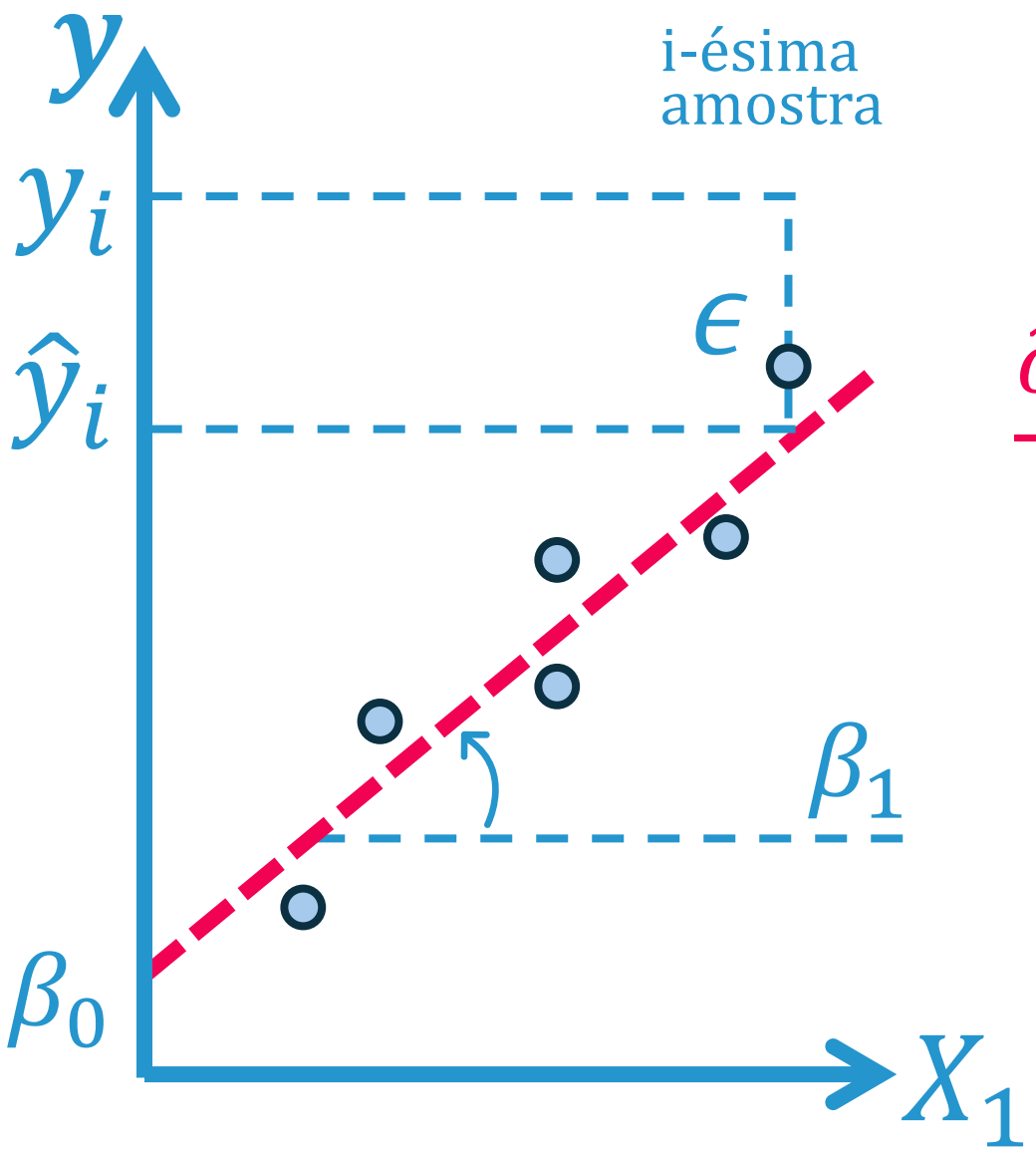
$\hat{y}_i \rightarrow$ $i-$ésima predição

$X_{i1} \rightarrow$ $i-$ésimo valor da var 1

$\beta_0 \rightarrow$ coef linear (intercepto)

$\beta_1 \rightarrow$ coef angular

Regressão Linear



$$\epsilon_i = y_i - \hat{y}_i$$

$$\epsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_{i1})^2$$

$$\frac{\partial \epsilon^2}{\beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 X_{i1}) = 0$$

$$\sum y_i = \sum \beta_0 + \sum \beta_1 X_{i1}$$

$$\sum y_i = n\beta_0 + \beta_0 \sum X_{i1}$$

Regressão Linear

$$\epsilon_i = y_i - \hat{y}_i \qquad \epsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_{i1})^2$$

$$\frac{\partial \epsilon^2}{\beta_1} = -2 \sum X_{i1}(y_i - \beta_0 - \beta_1 X_{i1}) = 0$$

$$\sum y_i X_{i1} = \sum \beta_0 X_{i1} + \sum \beta_1 X_{i1} X_{i1}$$

Regressão Linear

$$y_i \rightarrow \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \\ \dots \\ y_n \end{bmatrix} \qquad X_{i1} \rightarrow \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ 1 & X_{31} \\ \dots & \dots \\ 1 & X_{n1} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{bmatrix} n & \sum X_{i1} \\ \sum X_{i1} & \sum X_{i1}X_{i1} \end{bmatrix} \qquad \boldsymbol{X}^t\boldsymbol{y} = \begin{bmatrix} \sum y_i \\ \sum X_{i1}y_i \end{bmatrix}$$

Regressão Linear

$$\boldsymbol{X}^t\boldsymbol{y} = \begin{bmatrix} \sum y_i \\ \sum X_{i1} y_i \end{bmatrix} \qquad \boldsymbol{X}^t\boldsymbol{X} = \begin{bmatrix} n & \sum X_{i1} \\ \sum X_{i1} & \sum X_{i1} X_{i1} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\sum y_i = n\beta_0 + \beta_0 \sum X_{i1} \qquad \sum y_i X_{i1} = \sum \beta_1 X_{i1} + \sum \beta_1 X_{i1} X_{i1}$$

$$\boldsymbol{X}^t\boldsymbol{y} = \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}$$

Regressão Linear

$$X^t y = X^t X \beta$$

$$(X^t X)^{-1} X^t y = (X^t X)^{-1} (X^t X) \beta$$

$$\beta = (X^t X)^{-1} X^t y$$

Perceba que essa equação não faz referencia a quantidade de variáveis em (X) , logo, ela funciona para regressão linear simples (1 variável) e múltipla (várias variáveis)

Regressão Linear

$$\boldsymbol{\beta} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

$$\boldsymbol{y} = \boldsymbol{\beta}\boldsymbol{X}$$

Simples

$$y_i = \beta_0 + \beta_1 X_{i1}$$

Múltipla

$$y_i = \sum_{i=1}^{n} \beta_j X_{ij}$$

Regressão Linear

$$\boldsymbol{\beta} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

Perceba que o cálculo da inversa $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ impõe restrições para o caso da regressão linear múltipla ($\boldsymbol{X}$ com várias variáveis)

- $\boldsymbol{X}^t\boldsymbol{X}$ deve ser não-singular ($\det \neq 0$)

- Na prática isso é garantido se $\boldsymbol{X}$ tem 2^{m-1} amostras

Algumas métricas de performance

$$R^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

$$RMSE = \sqrt{\sum \frac{(\hat{y}-y)^2}{n}}$$

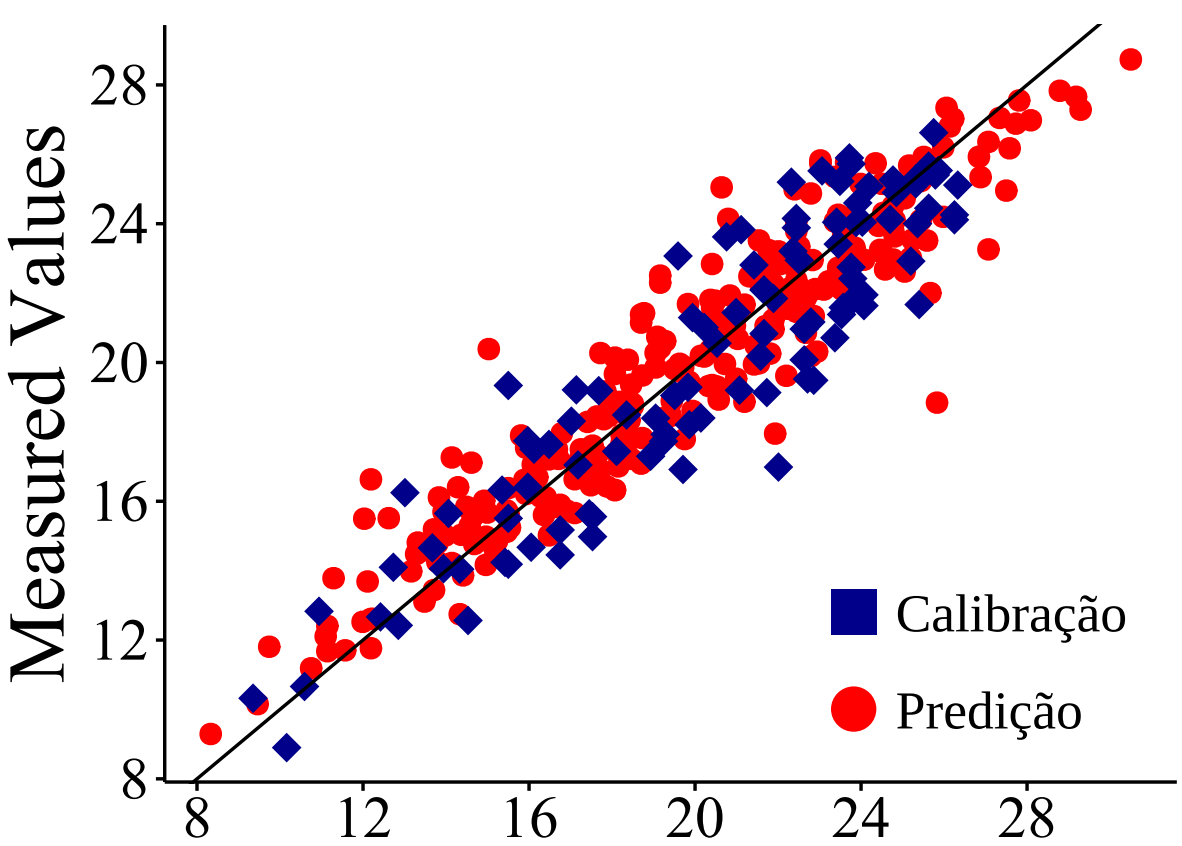
$$RPD = \frac{SD(y)}{RMSEP}$$

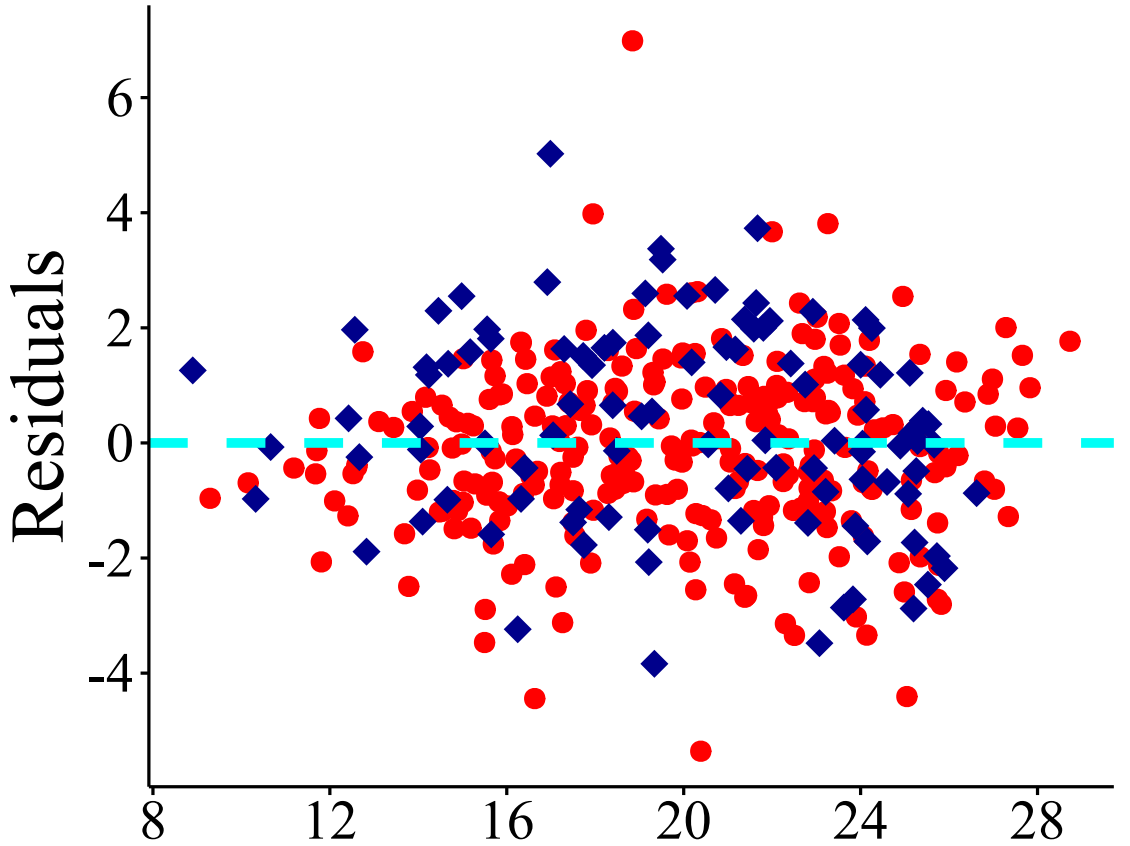
$$RPIQ = \frac{IQR(y)}{RMSEP}$$

$$Bias = \frac{\sum_{n_p}(y-\hat{y})}{n_p}$$

$$t_{Bias}$$

Algumas métricas de performance





Reference Values

Suposições assumidas

Linearidade: A relação entre variáveis dependentes e independentes deve ser linear.

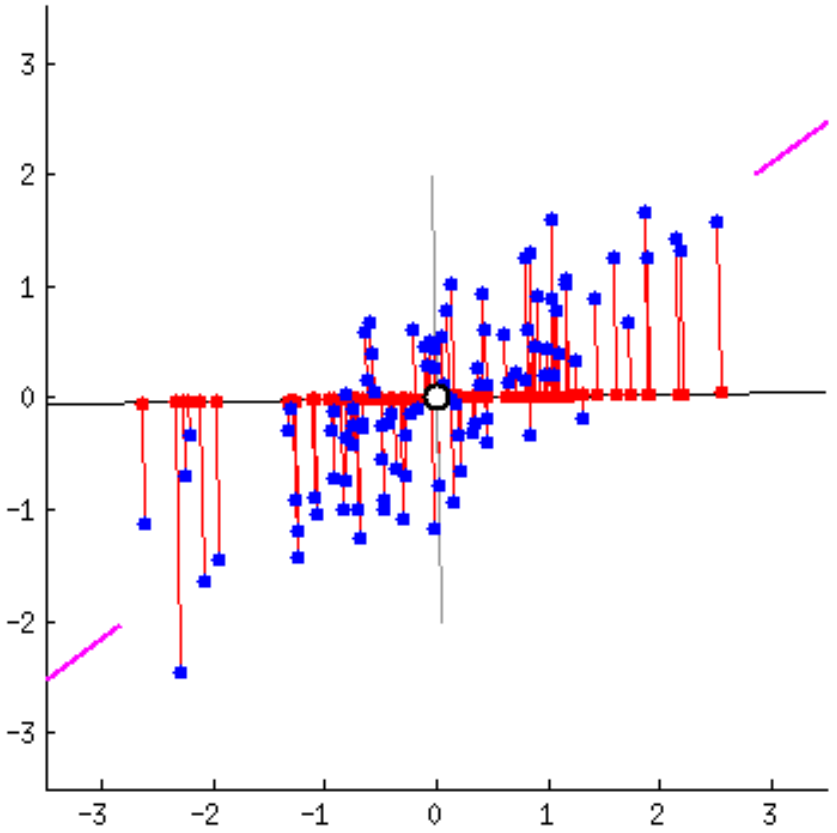
Homocedasticidade: A variância constante dos erros deve ser mantida.

Normalidade dos resíduos: a regressão múltipla assume que os resíduos são distribuídos normalmente

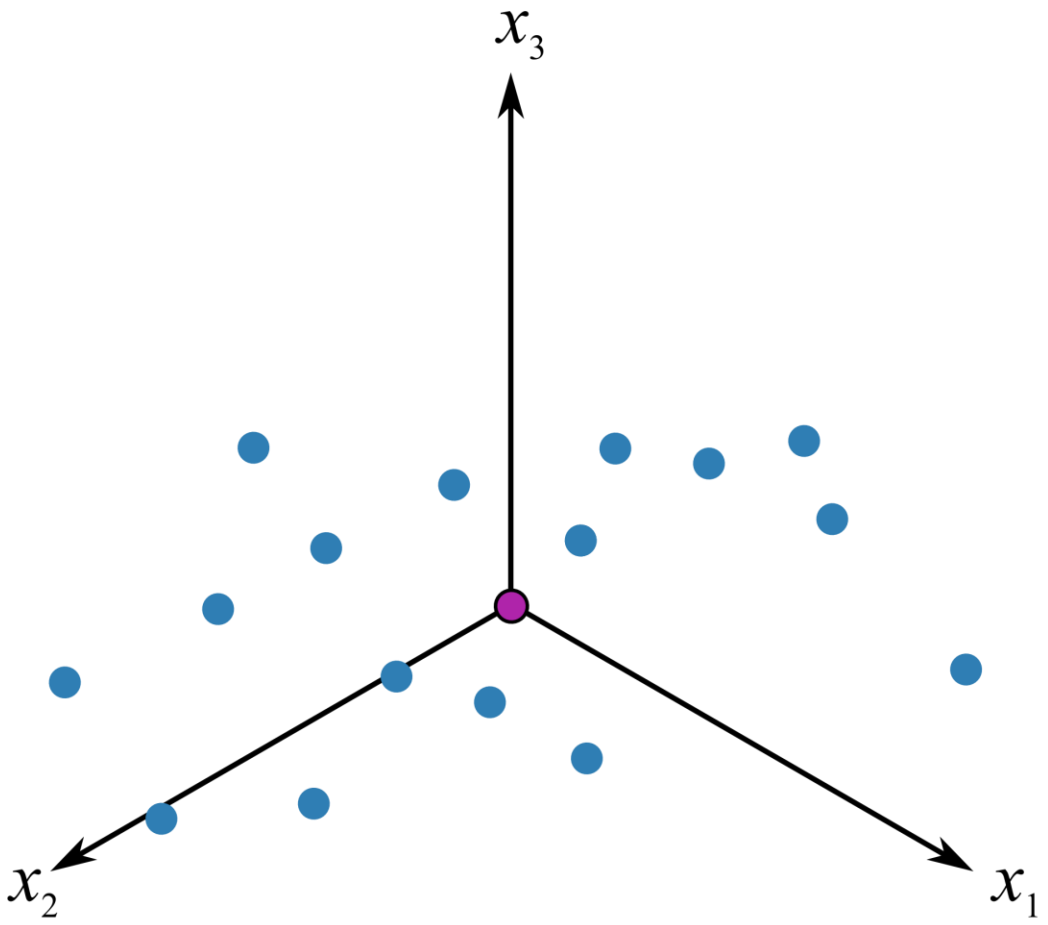
Regressão via Mínimos Quadrados Parciais

Mínimos Quadrados Parciais

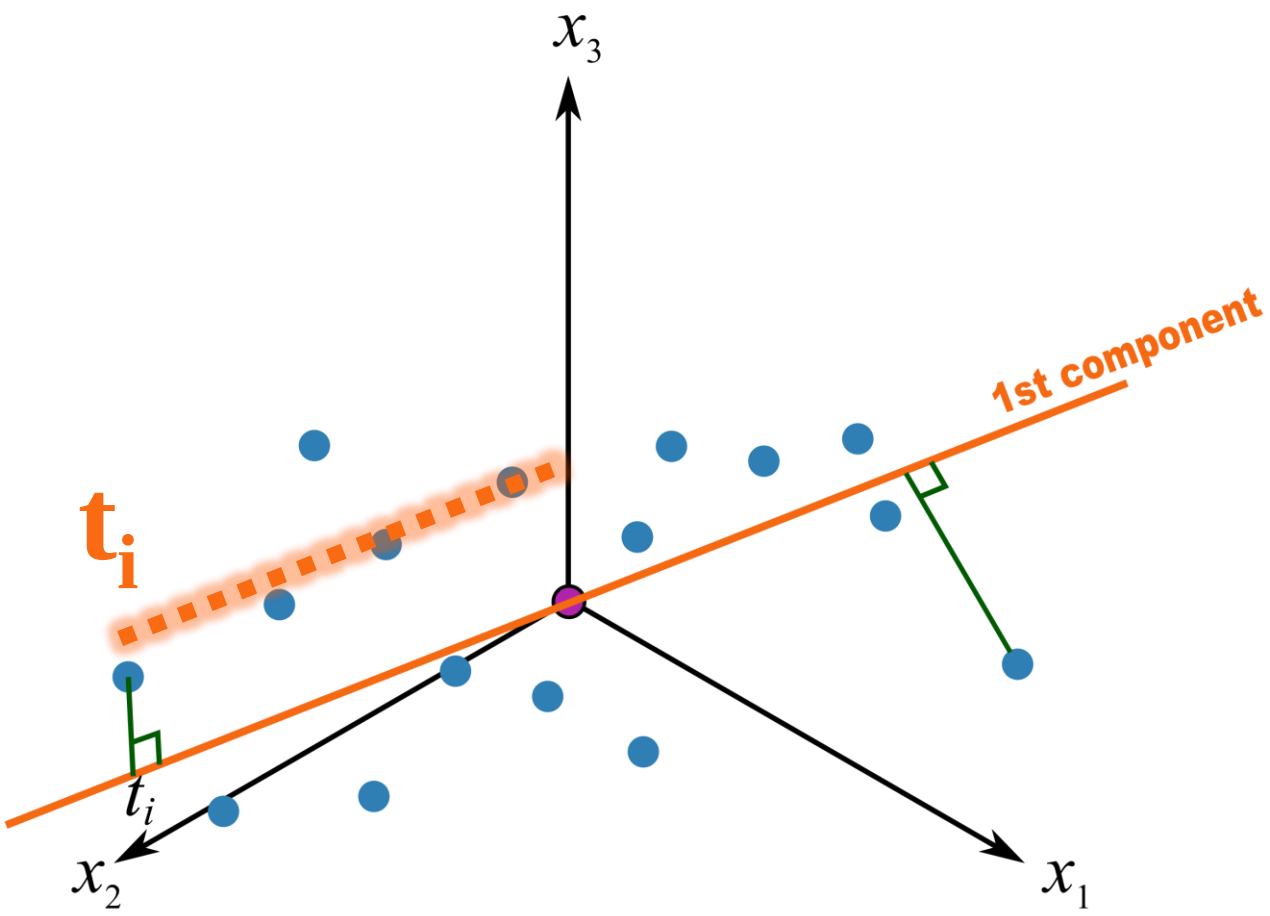
- O PLS resolve o problema do número de variáveis/amostras aplicando uma redução de dimensionalidade baseada em PCA



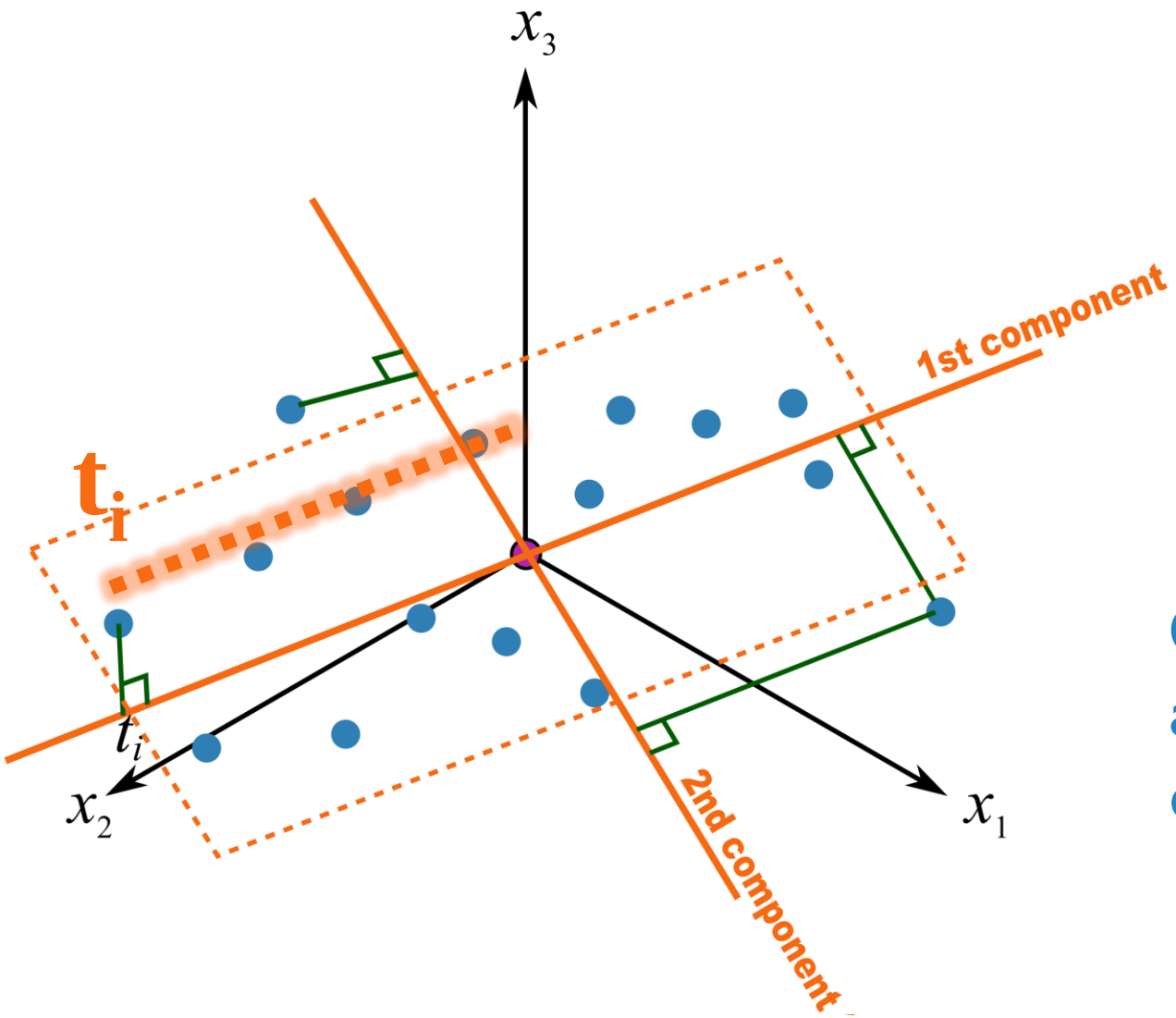
Mínimos Quadrados Parciais



Mínimos Quadrados Parciais



Mínimos Quadrados Parciais

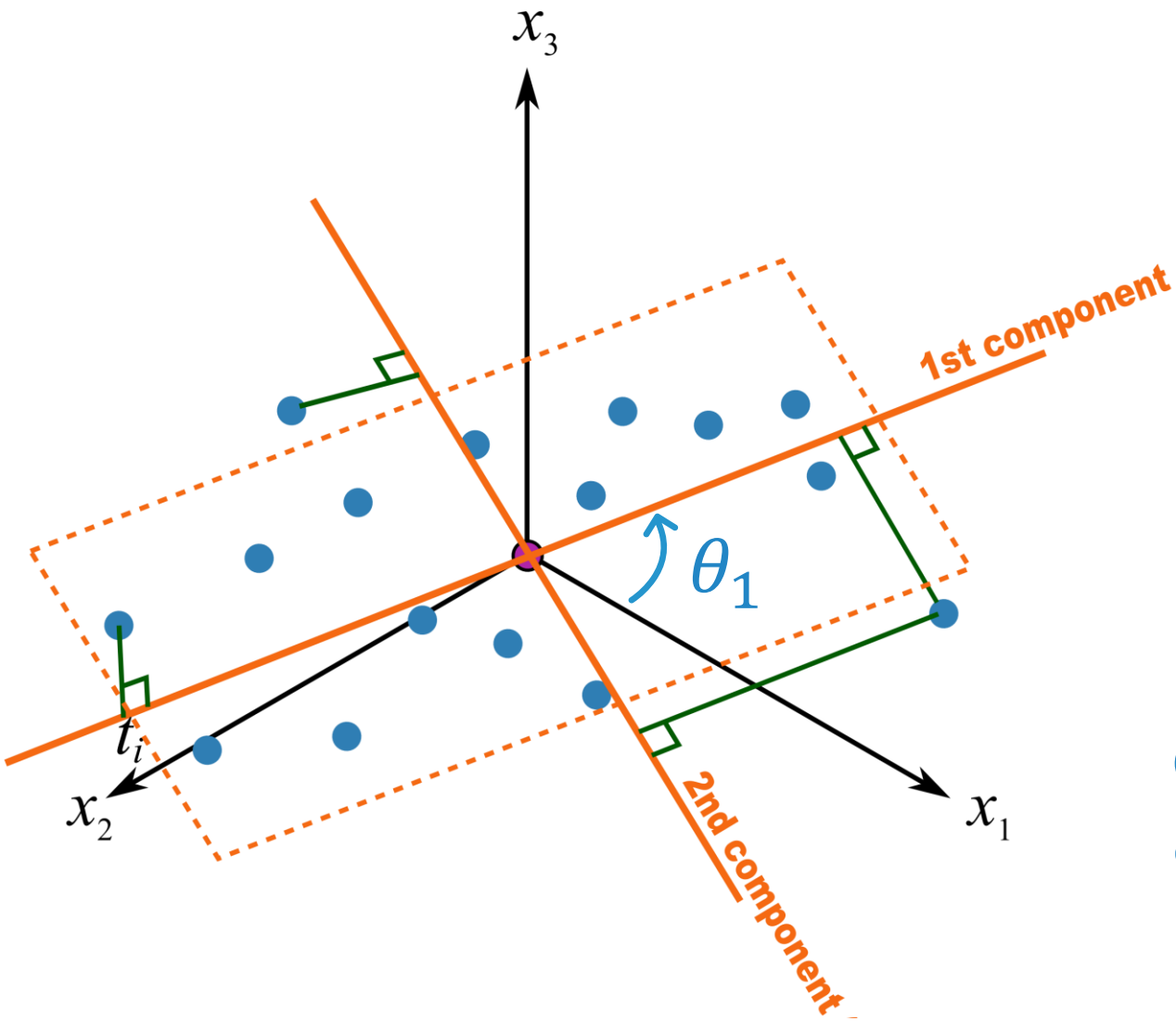


$$\boldsymbol{X = TP^t + E}$$

$$X = \sum_{i=1}^{A} t_i p_i + e_i$$

Os t_1 scores exibem relações entre as amostras (projeção delas no espaço das PCs)

Mínimos Quadrados Parciais



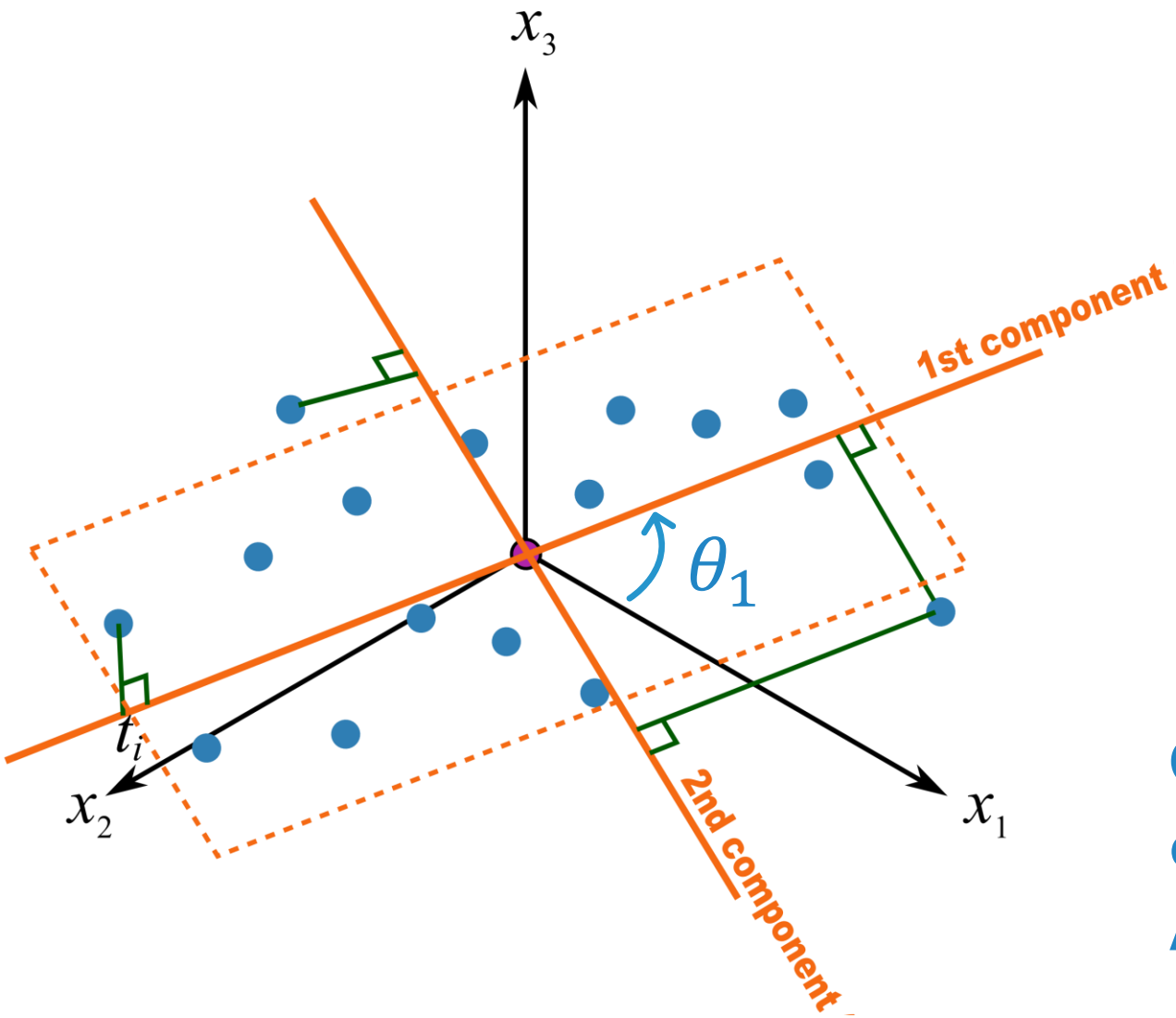
$$\boldsymbol{X = TP^t + E}$$

$$X = \sum_{i=1}^{A} t_i p_i + e_i$$

$$p_{11} = cos\theta_1$$

O p_{11} loading indica a contribuição da primeira variável para a PC1

Mínimos Quadrados Parciais



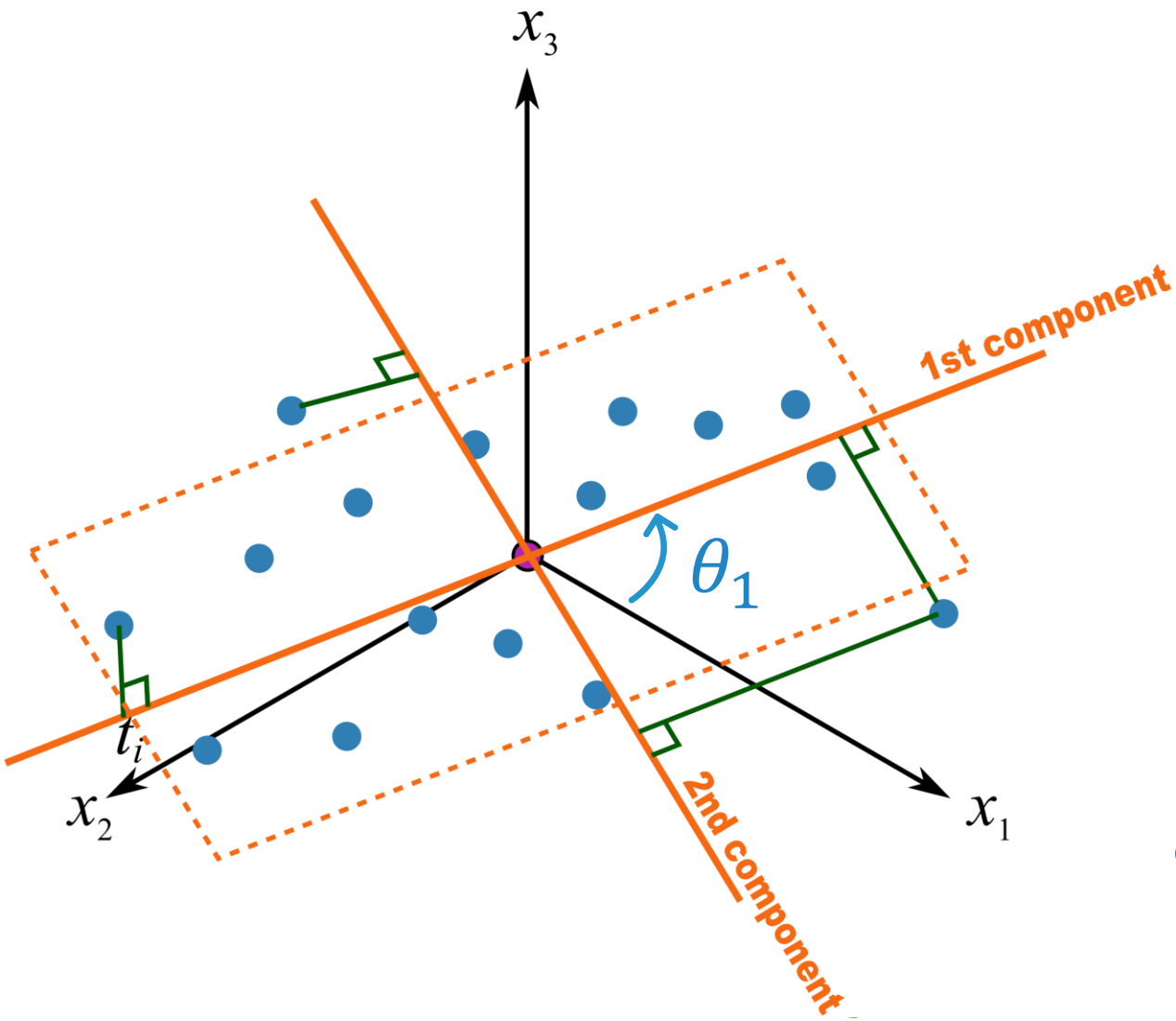
$$\boldsymbol{X = TP^t + E}$$

$$X = \sum_{i=1}^{A} t_i p_i + e_i$$

$$p_{1i} = cos\theta_i$$

O p_{1i} loading indica a contribuição da i-ésima variável para a PC1. Analogamente para PC2, PC3, ...

Mínimos Quadrados Parciais



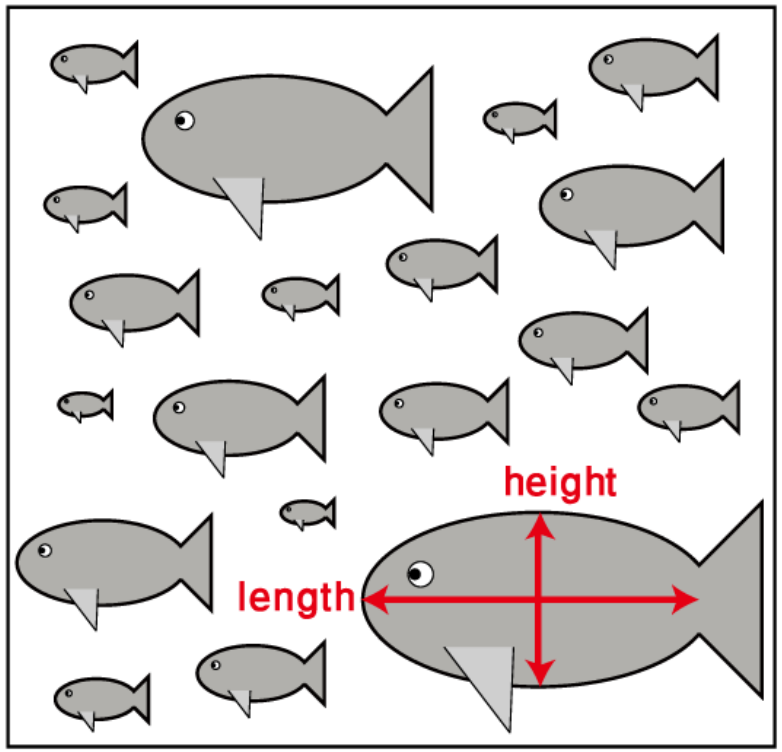
$$\boldsymbol{X = TP^t + E}$$

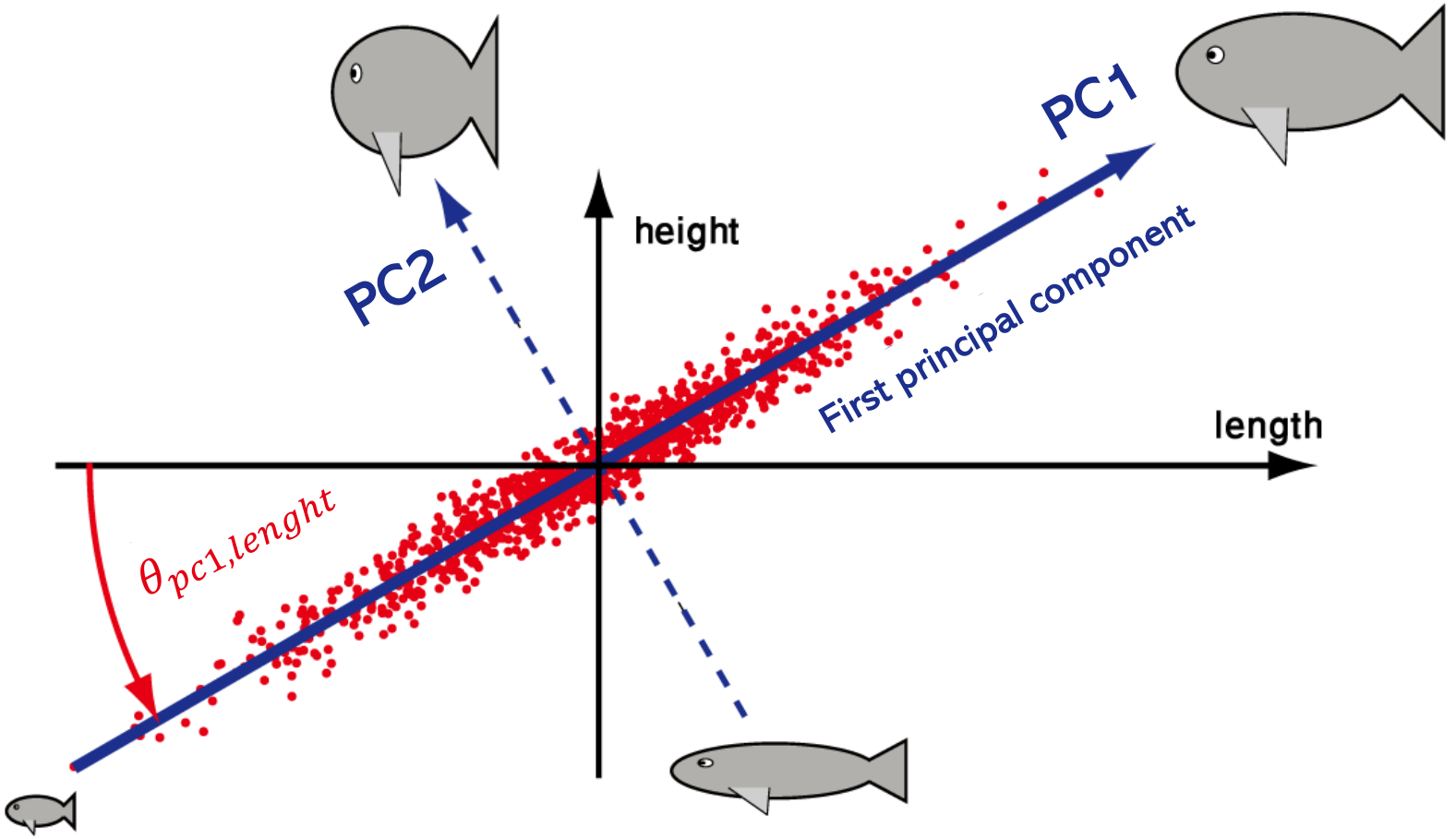
$$X = \sum_{i=1}^{A} t_i p_i + e_i$$

$$p_{1i} = cos\theta_i$$

Dessa forma, os loadings indicam as contribuições das variáveis originais para a construção das PCs

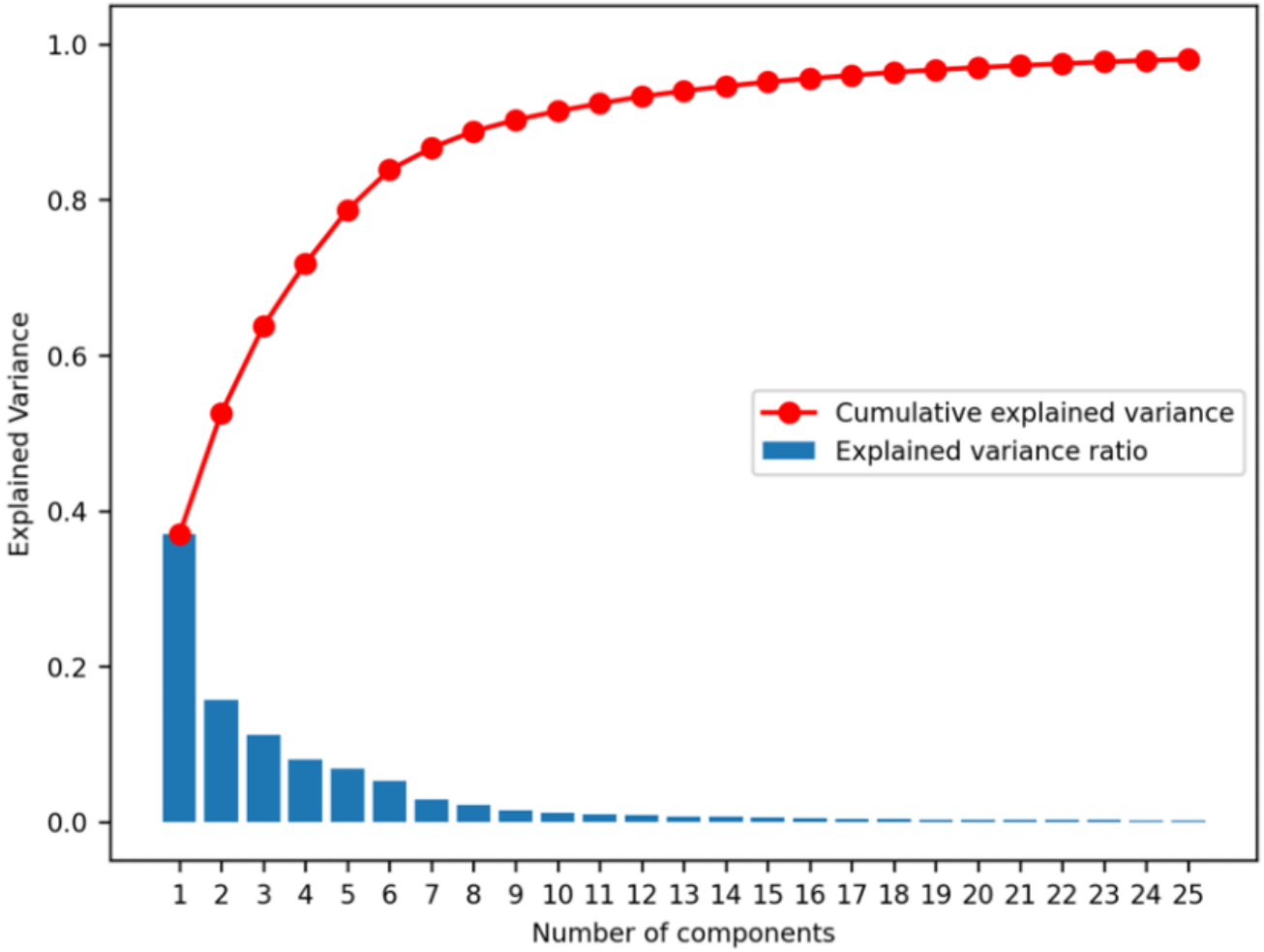
Mínimos Quadrados Parciais





Mínimos Quadrados Parciais

Como saber o numero adequado de PCs? **Variância explicada!**



Mínimos Quadrados Parciais

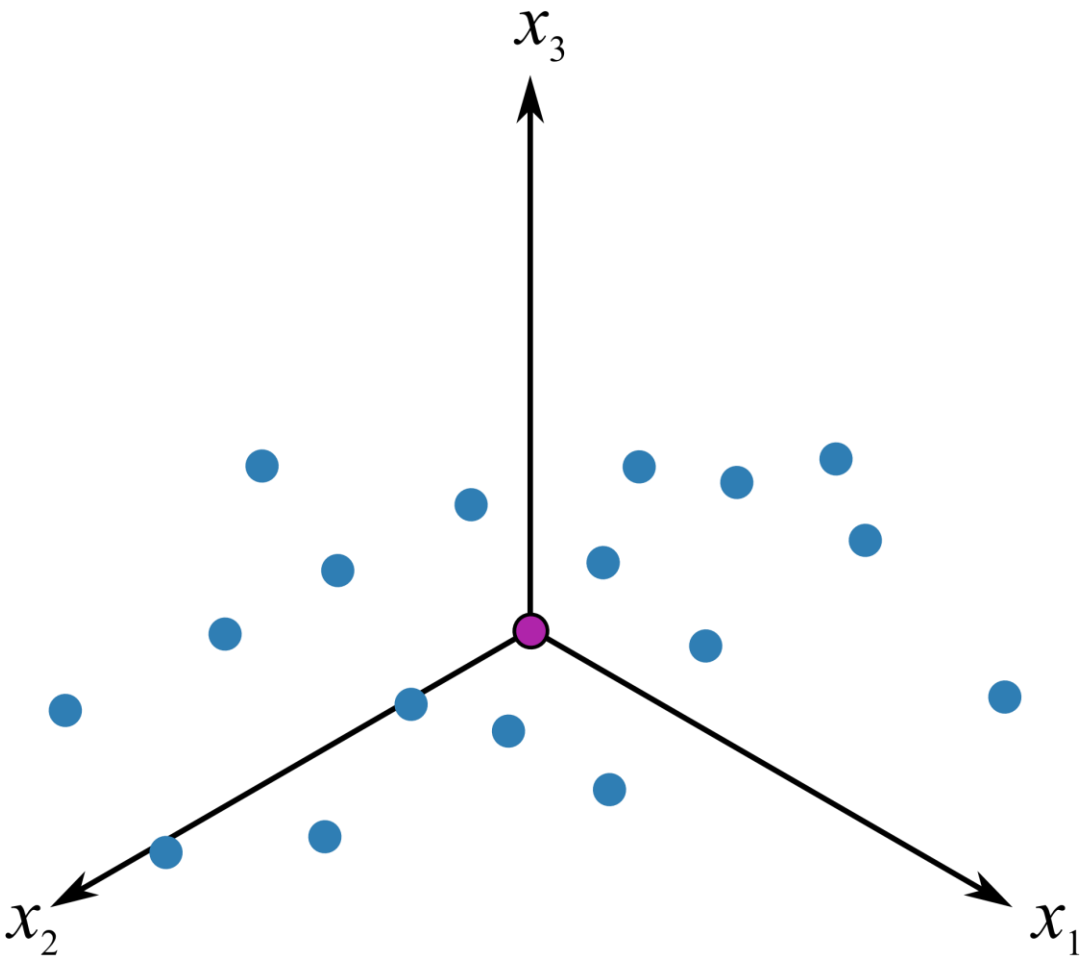
O PLS aplica uma decomposição em PCA em $\boldsymbol{X}$ e $\boldsymbol{y}$ para relacionar **linearmente** os scores de ambas as matrizes

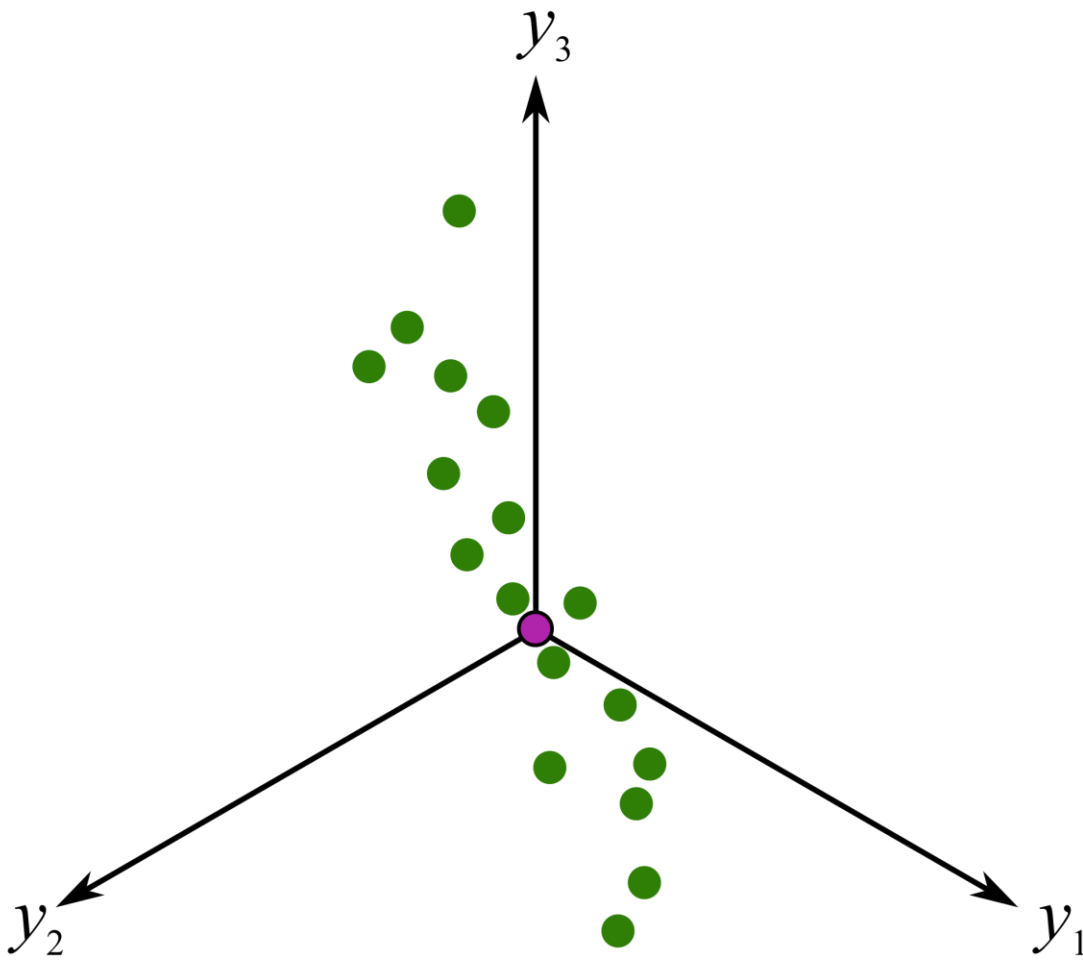
$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E} \rightarrow \sum_{i=1}^{A} \boldsymbol{t}_{ij}\boldsymbol{p}_{ij} + \boldsymbol{e}_{ij} \qquad \mathbf{y} = \boldsymbol{U}\boldsymbol{Q}^{\boldsymbol{T}} + \boldsymbol{F} \rightarrow \sum_{i=1}^{A} \boldsymbol{u}_i\boldsymbol{q}_i + \boldsymbol{f}_i$$

$$\boldsymbol{b}_i = \frac{\boldsymbol{u}_i^{\boldsymbol{T}}\boldsymbol{t}_i}{\boldsymbol{t}_i^{\boldsymbol{T}}\boldsymbol{t}_i}\ ; \quad \boldsymbol{u}_i = \boldsymbol{b}_i\boldsymbol{t}_i$$

Mínimos Quadrados Parciais

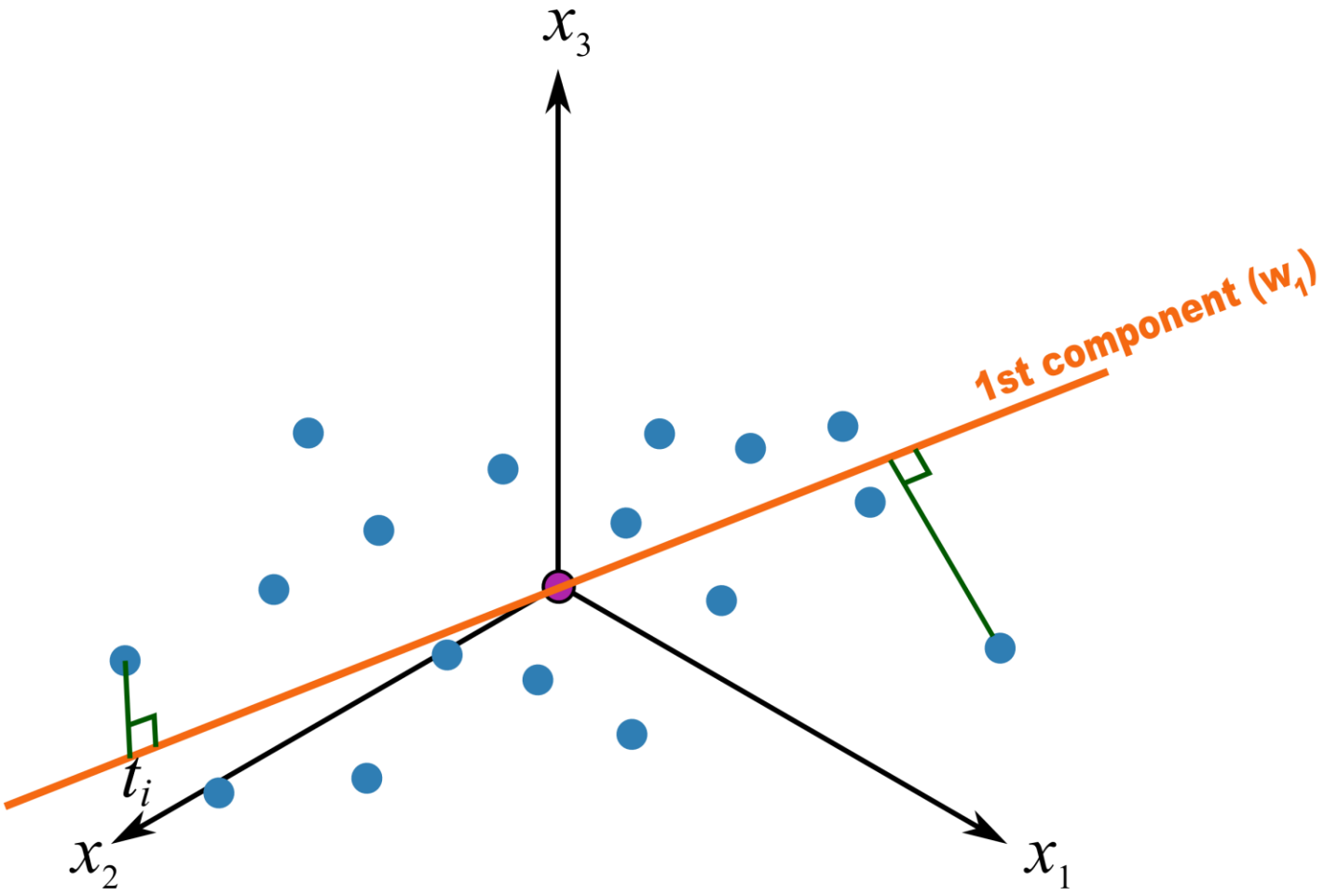
O PLS aplica uma decomposição em PCA em X e y para relacionar os scores de ambas as matrizes

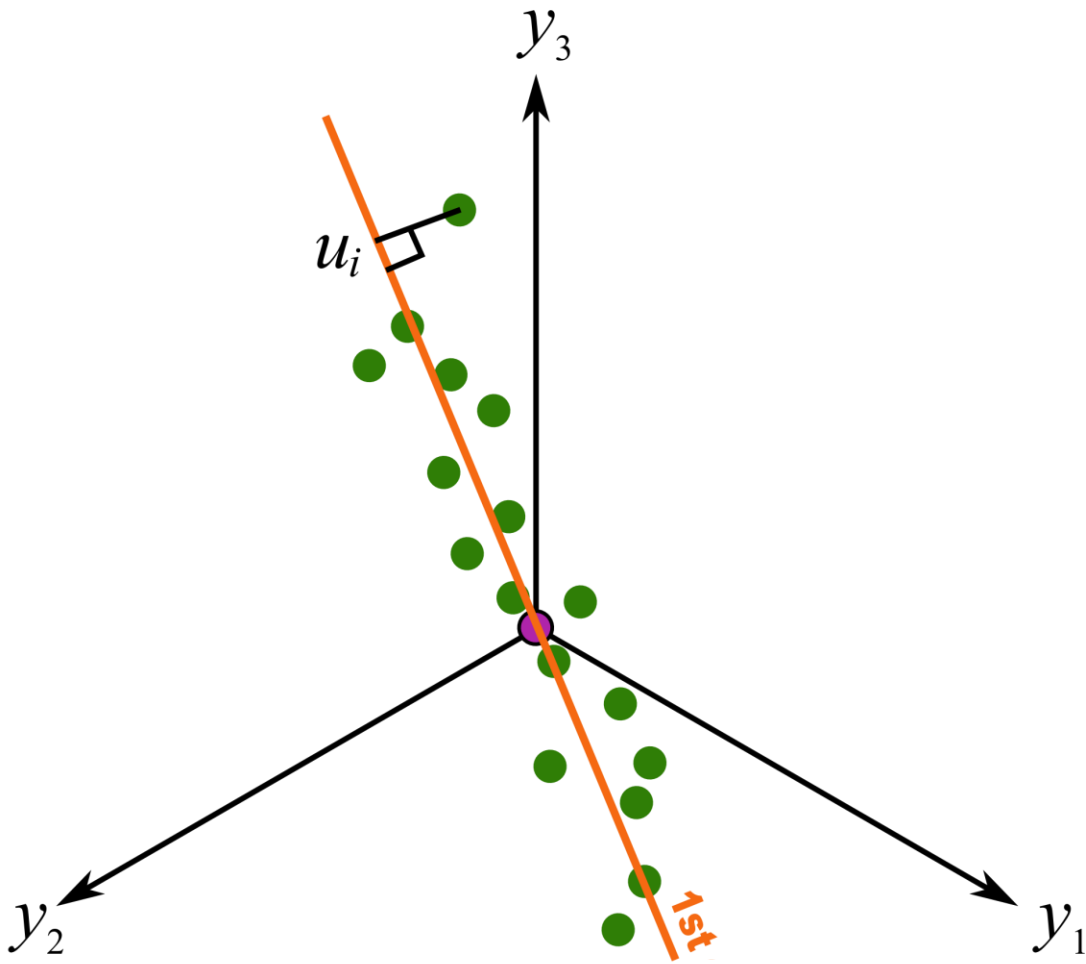




Mínimos Quadrados Parciais

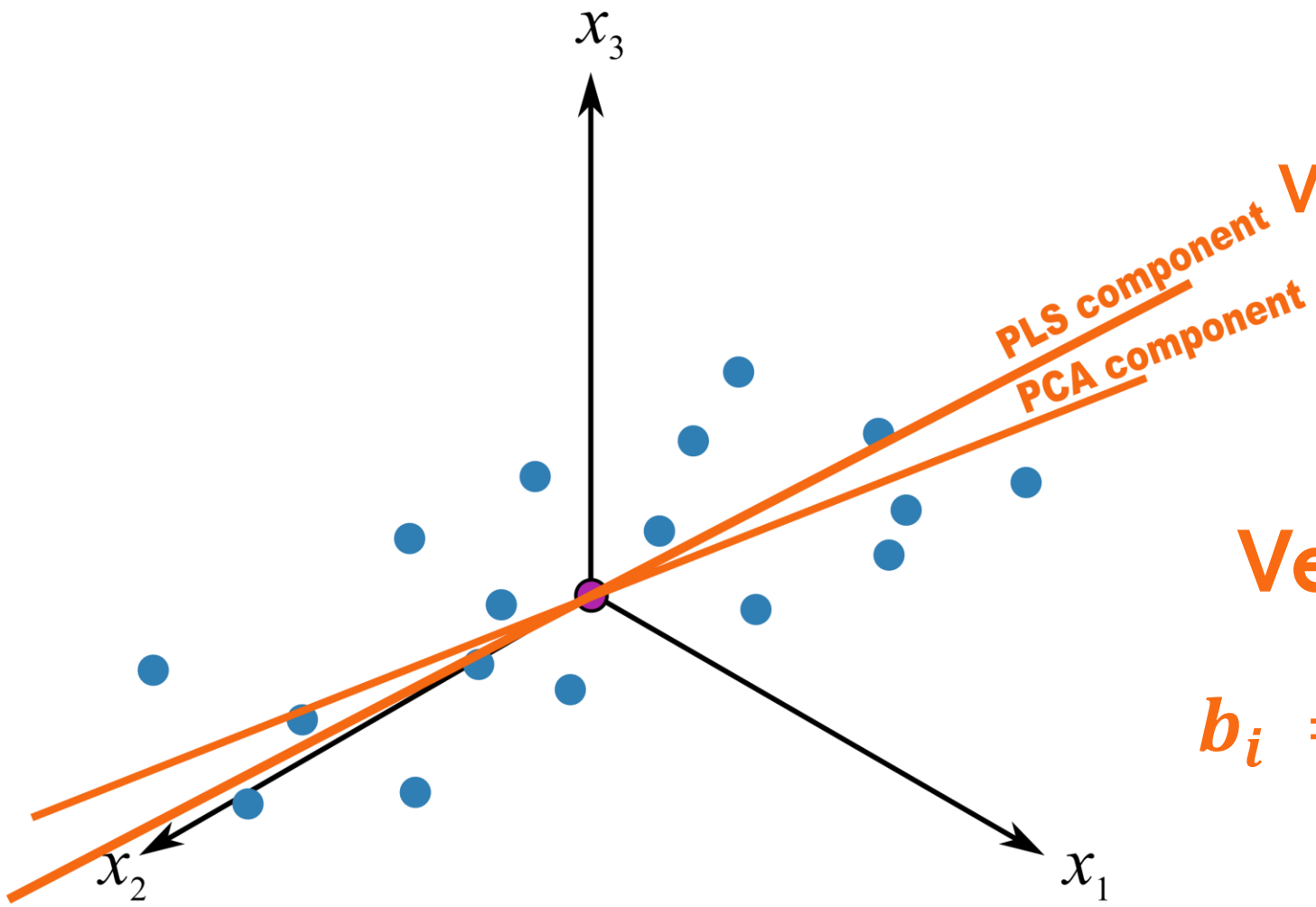
O PLS aplica uma decomposição em PCA em X e y para relacionar os scores de ambas as matrizes





Mínimos Quadrados Parciais

Os scores de X são rotacionados de forma a atingir a máxima covariância com y (perda de ortogonalidade).



Variável Latente (LV)

Vetor de regressão

$$b_i = \frac{u_i^T t_i}{t_i^T t_i} \; ; \quad u_i = b_i t_i$$

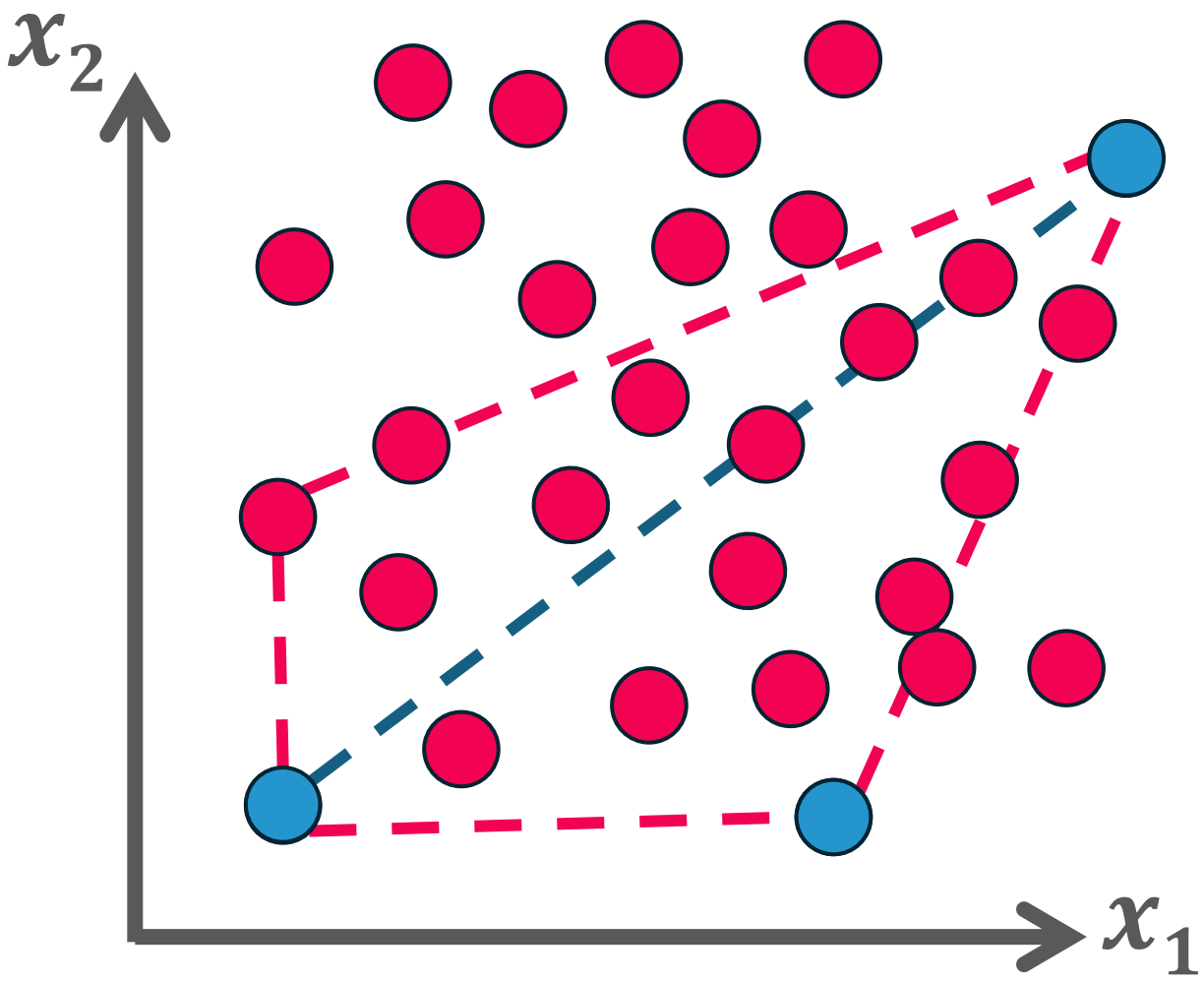
Conceitos

Conjunto de calibração (treinamento): Parcela dos dados sistematicamente escolhida para calibrar (ou treinar) o modelo através de um **aprendizado supervisionado**.

Conjunto de validação (predição): Parcela dos dados utilizada para validação independente da performance do modelo. Revela o **poder de generalização** do aprendizado para dados semelhantes, porém, que não foram explorados em nenhuma etapa pelo modelo.

Overfitting: tendência dos modelos de se especializarem demais nas características do conjunto de treinamento (calibração) a ponto de não serem generalizados para outros conjuntos independentes (mesmo que similares)

Kennard-Stone

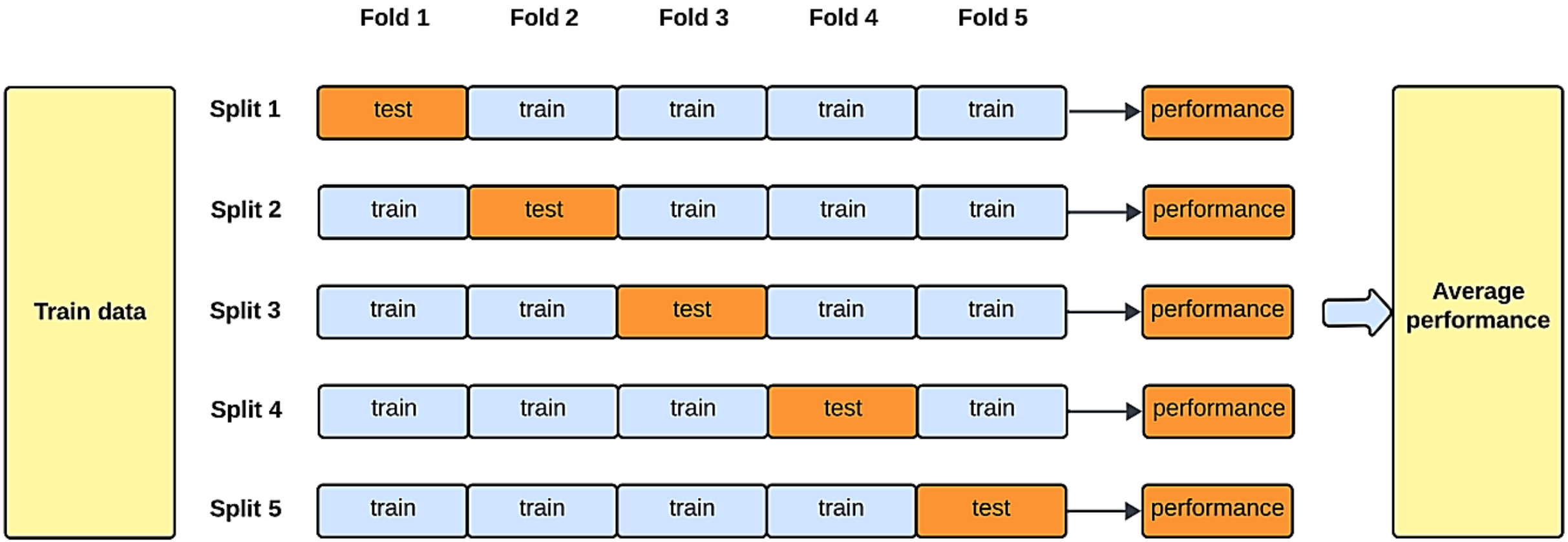


$$D_{ij} = \sqrt{\sum_{\substack{n=1 \\ i \neq j}}^{N} (x_{in} - x_{jn})^2}$$

Entre as outras amostras, aquelas com as **maiores distâncias mínimas** são englobadas no sub-conjunto

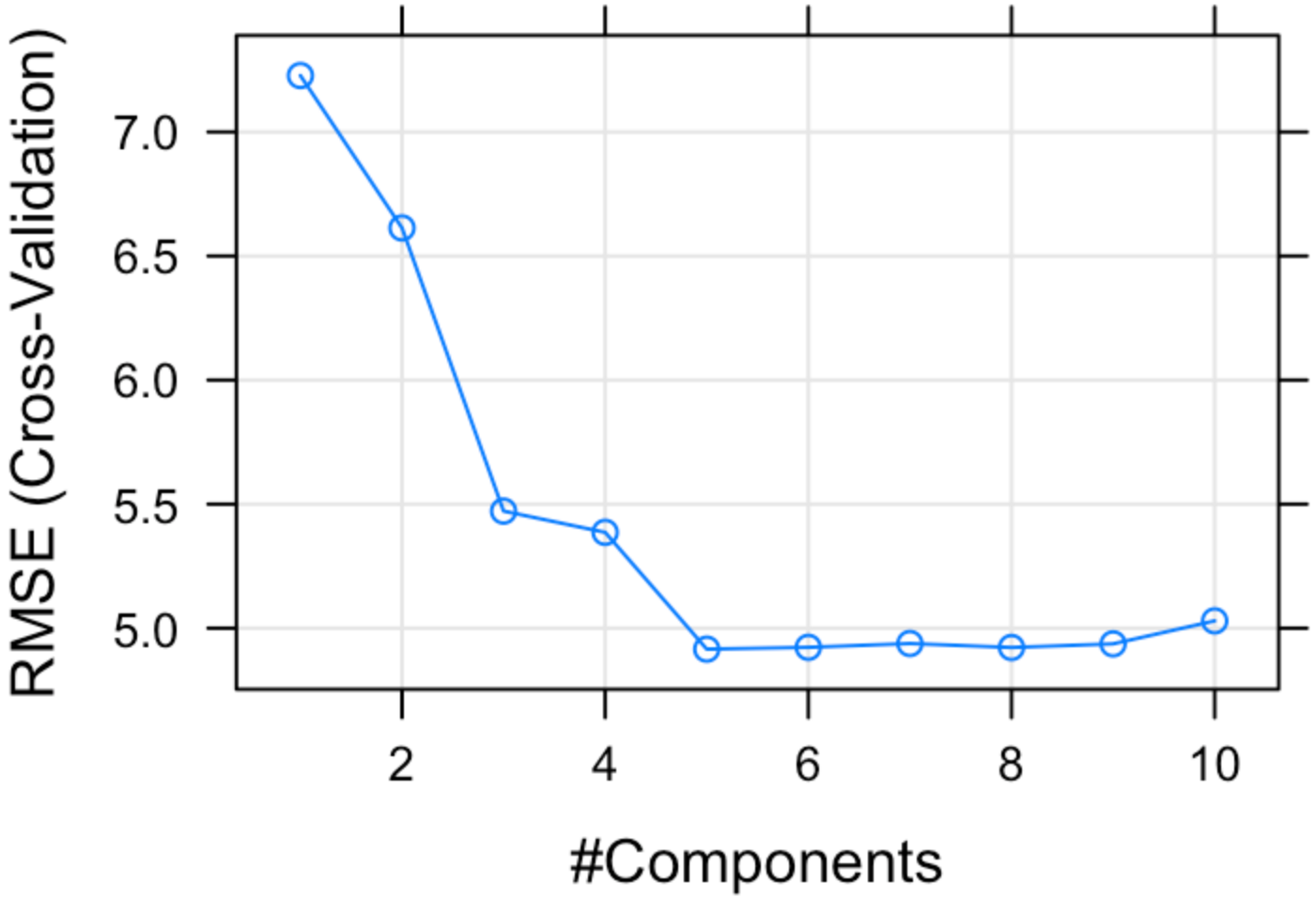
Cross-Validation

Como escolher o número adequado de LVs? **Cross-validation!**



Mínimos Quadrados Parciais

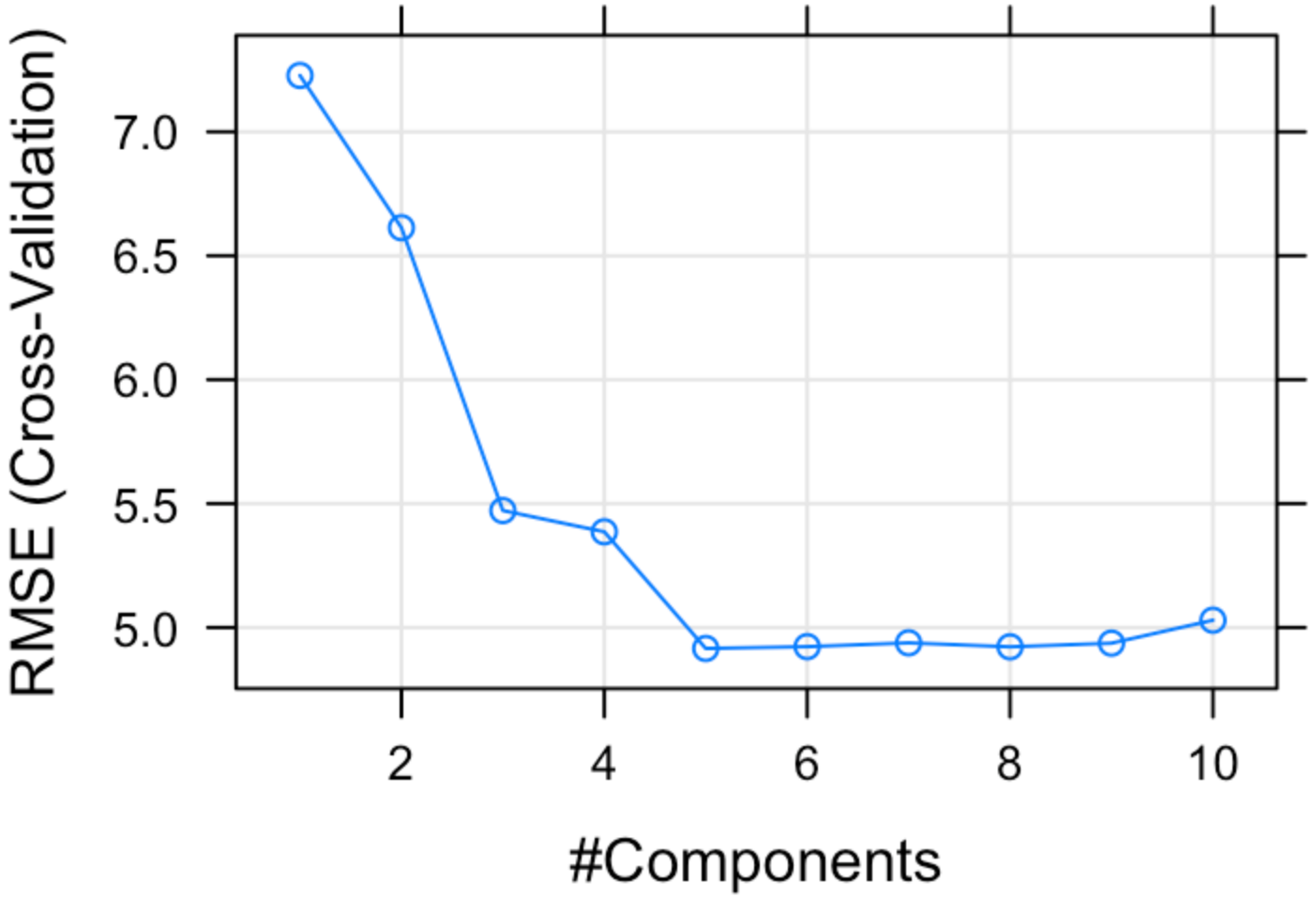
No PLS o **RMSECV** é comumente utilizado para a escolha da LV



Mínimos Quadrados Parciais

↑LVs (↑complexidade) → **Overfitting**
(modelo englobando ruído)

↓LVs (↓complexidade) → **Underfitting**
(modelo não englobando padrões relevantes)



Interpretabilidade
e
Explicabilidade

Interpretabilidade: quão bem um modelo pode ser compreendido por humanos. Ele aborda a clareza do processo de decisão interna do modelo e quão facilmente um humano pode entender as etapas que ele toma para tomar decisões

Explicabilidade: refere-se à clareza por trás dos mecanismos que geram as saída de um modelo. Ele visa aumentar a transparência e a compreensibilidade do processo de tomada de decisão do modelo, incluindo ferramentas e técnicas que buscam dividir resultados complexos em componentes compreensíveis

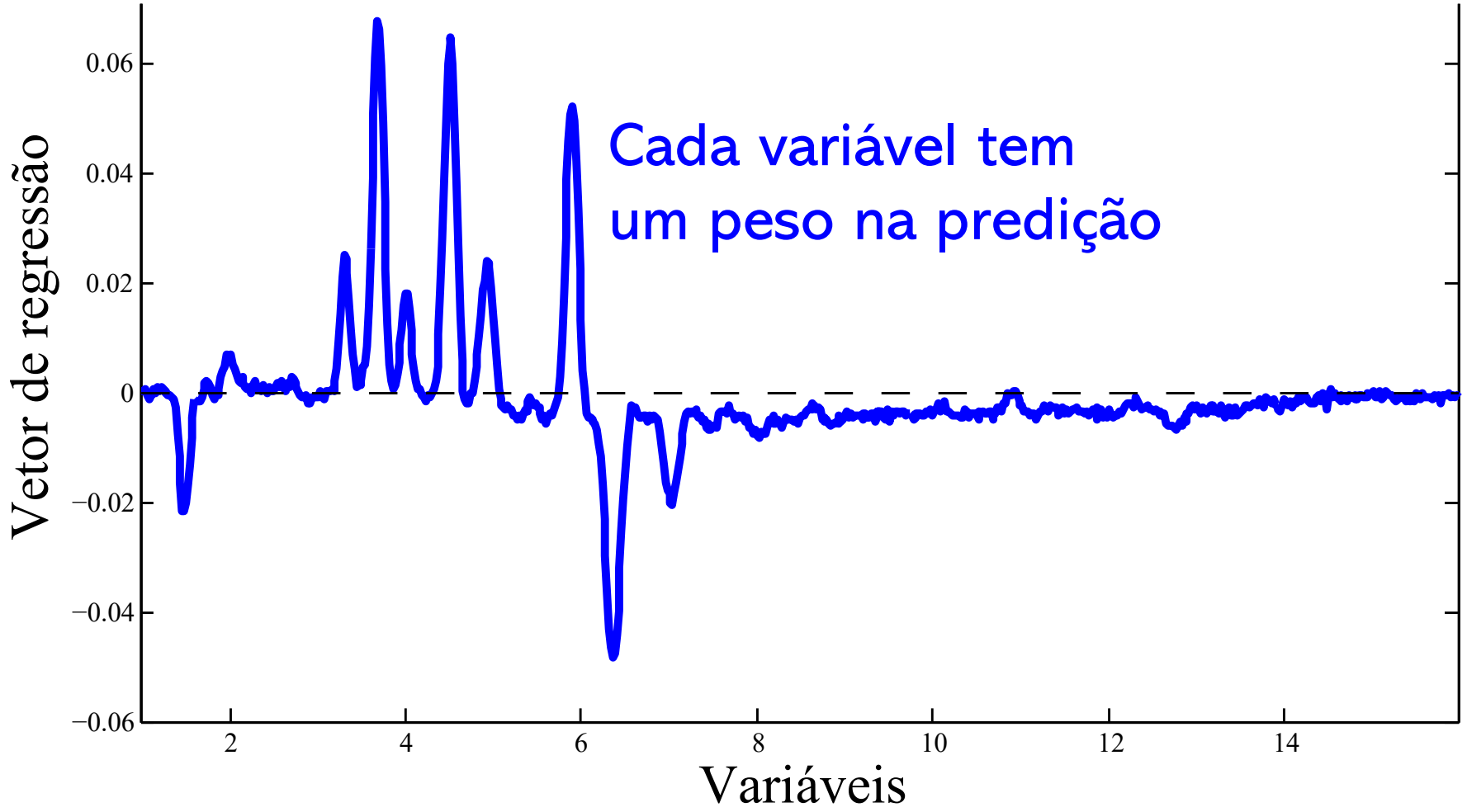
Vetores de regressão em modelos lineares

Modelos lineares como esses são tidos como “interpretáveis” e “explicáveis”, diferente de alguns outros que ainda veremos

I.e., além do sabermos muito bem como o processo de aprendizagem ocorre, temos acesso a parâmetros informativos internos: vetores de regressão

Vetores de regressão em modelos lineares

A natureza linear dos métodos utiliza vetores de regressão para gerar as predições



VIP scores

VIP scores também são comumente extraídos dos modelos PLS. Eles são obtidos através da matriz de pesos $\boldsymbol{W}$ $(\boldsymbol{X} = \boldsymbol{TW})$

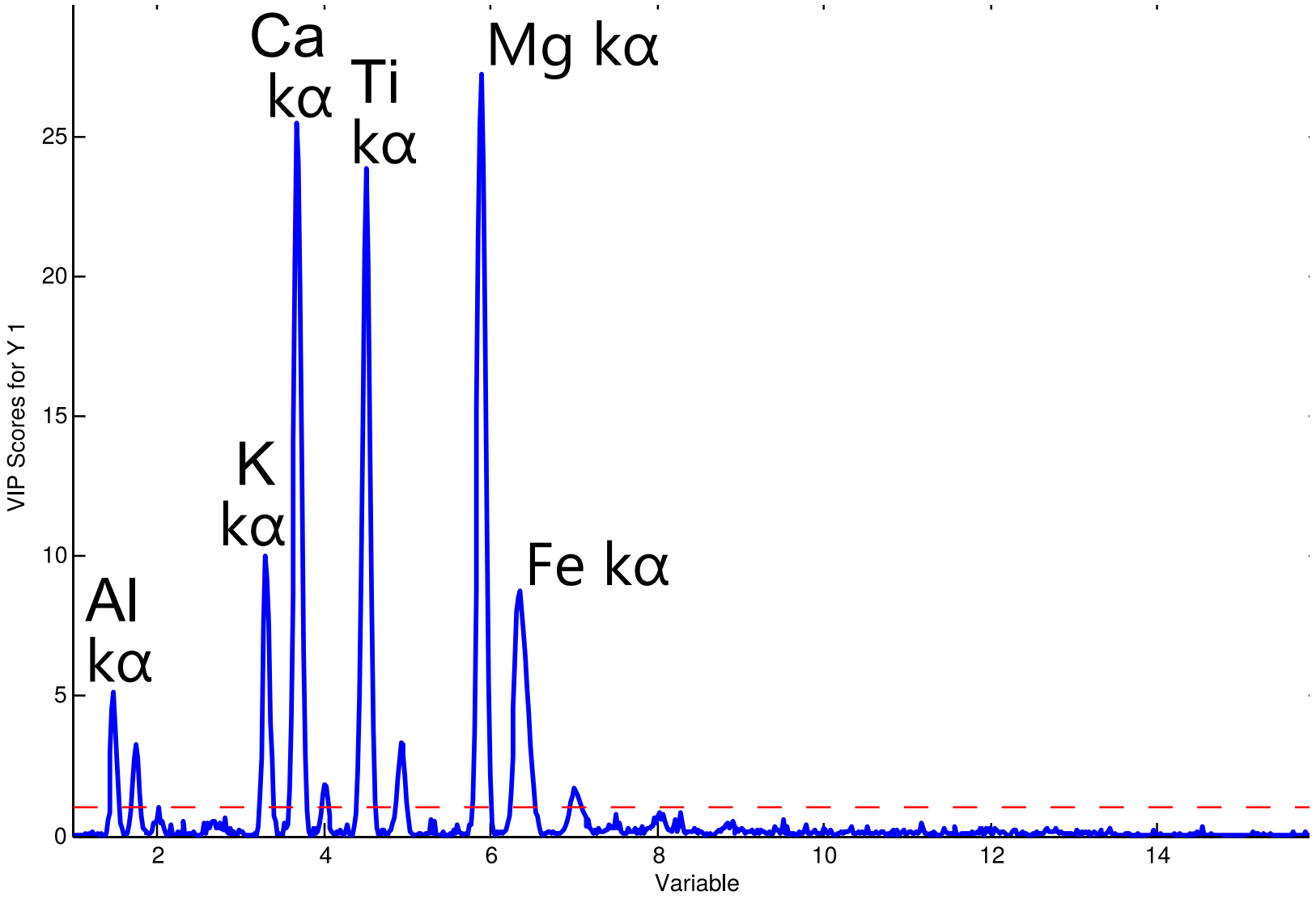
$$VIP_j = \sqrt{p \frac{\sum_{a=1}^{A} SSY_a \left(\frac{w_{ja}}{\|\mathbf{w}_a\|} \right)^2}{\sum_{a=1}^{A} SSY_a}}$$

- p é o número total de variáveis preditoras,
- A é o número total de componentes latentes do modelo,
- w_{ja} é a ponderação da variável j na componente a (elemento na j-ésima linha e a-ésima coluna da matriz $\mathbf{W}$),
- $\|\mathbf{w}_a\| = \sqrt{\sum_{j=1}^{p} w_{ja}^2}$ é a norma (ou comprimento) do vetor de pesos da componente a,
- SSY_a é a soma dos quadrados (ou "variância") da resposta $\mathbf{Y}$ que é explicada pela componente a.

Dessa forma, os VIPs combinam as contribuições de todas as LVs construídas durante a regressão. Em termos gerais, eles usado como um critério para estabelecer para as variáveis significativas para modelagem (quantificando o quanto cada variável foi importante para o modelo PLS considerando pela variância de $\mathbf{Y}$ cada LV)

Como a média dos VIP scores ao quadrado é 1, maior que 1 é

VIP scores


Ca kα
Ti kα
Mg kα
K kα
Al kα
Fe kα
VIP Scores for Y 1
Variable
0
5
10
15
20
25
2
4
6
8
10
12
14

Prática no VS Code