Introdução à Ciência de Dados para Engenheiros

Modelos de ordem reduzida

Regressão por Componentes Principais (PCR)

Regressão Parcial de Mínimos Quadrados (PLSR)

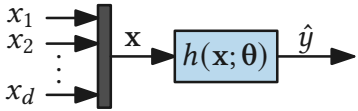
Renato Maia Matarazzo Orsino
reorsino@usp.br

Conteúdo

Introdução

Problemas de mínimos quadrados

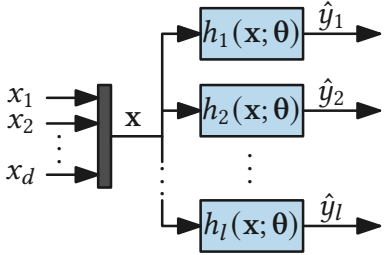
Para sistemas de interesse em que de um dado conjunto de *entradas* $\mathbf{x} \in \mathbb{R}^d$ produz um dado conjunto de *saídas* $y \in \mathbb{R}$ de tal forma que se possa admitir que tal relação é *determinística* (ou seja, repetidas as entradas, repetem-se as saídas a menos de diferenças ruidosas ou devidas a efeitos de segunda ordem), é possível propor o ajuste de um modelo matemático parametrizado $\hat{y} = h(\mathbf{x}; \boldsymbol{\theta})$ que fornece uma estimativa para y dadas novas entradas $\mathbf{x}$.



O ajuste do modelo baseia-se na determinação de um *vetor de parâmetros* $\boldsymbol{\theta} \in \mathbb{R}^m$ que minimiza de uma função objetivo, baseada em uma métrica do erro de aproximação e descrita por uma forma quadrática. Formulações deste tipo são genericamente denominados problemas de mínimos quadrados (LS).

Sistemas com múltiplas saídas

Caso o sistema possua múltiplas saídas, é sempre possível ajustar um modelo matemático para a previsão de cada saída (nestes casos, entretanto, todos os modelos terão as mesmas entradas). Assim, via de regra, o problema será formulado de forma a obter, via mínimos quadrados, o vetor de parâmetros $\boldsymbol{\theta} \in \mathbb{R}^m$ dado uma série de dados $(\mathbf{x}^{(i)}, y^{(i)}), i = 1, 2, \ldots, n$, com $y^{(i)} \in \mathbb{R}$ (escalar) e $\mathbf{x}^{(i)} \in \mathbb{R}^d$. Tais formulações são tipicamente conhecidas como MISO (do Inglês, *multiple input, single output*).



Problemas de regressão linear e não-linear

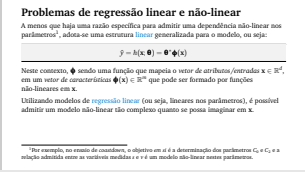
A menos que haja uma razão específica para admitir uma dependência não-linear nos parâmetros[1], adota-se uma estrutura linear generalizada para o modelo, ou seja:

$$\hat{y} = h(\mathbf{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x})$$

Neste contexto, $\boldsymbol{\phi}$ sendo uma função que mapeia o *vetor de atributos/entradas* $\mathbf{x} \in \mathbb{R}^d$, em um *vetor de características* $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^m$ que pode ser formado por funções não-lineares em $\mathbf{x}$.

Utilizando modelos de regressão linear (ou seja, lineares nos parâmetros), é possível admitir um modelo não-linear tão complexo quanto se possa imaginar em $\mathbf{x}$.

[1] Por exemplo, no ensaio de *coastdown*, o objetivo *em si* é a determinação dos parâmetros C_0 e C_2 e a relação admitida entre as variáveis medidas s e v é um modelo não-linear nestes parâmetros.



- Se a complexidade da função $\boldsymbol{\phi}$ for demasiada e a dimensão m for elevada, pode ser vantajoso admitir: $\boldsymbol{\theta} = \sum_{i=1}^{n} \beta_i \boldsymbol{\phi}(\mathbf{x}^{(i)})$ e utilizar a técnica da função núcleo (*kernel function*) para o ajuste de um novo vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^n$.
- Todo problema de regressão linear recai, em última instância, em um problema de solução de um sistema linear no respectivo vetor de parâmetros. Sempre é possível encontrar uma solução para este sistema utilizando matrizes pseudo-inversas.
- Uma abordagem alternativa, particularmente vantajosa do ponto de vista computacional quando o número de pontos de dados envolvidos for muito maior que o de parâmetros a estimar ($n \gg m$), e válida tanto para *regressão linear* quanto para *regressão não-linear* (ainda que neste último caso costume haver grande sensibilidade de convergência com respeito aos parâmetros de inicialização do algoritmo), é o uso do algoritmo recursivo de mínimos quadrados (RLS) ou de versões aproximadas, *não necessariamente convergentes*, como o algoritmo do gradiente descendente estocástico (SGD).

Modelos de ordem reduzida

Ainda, é comum que haja redundância na informação descrita nas entradas e saídas de um sistema, sendo possível a proposição de modelos de ordem reduzida. A seguir, veremos duas estratégias para tal:

- A regressão por componentes principais (PCR), que se baseia em reduzir a regressão apenas às componentes de entrada que se encontram melhor representadas na série de dados utilizada para o ajuste.
- A regressão parcial de mínimos quadrados (PLSR), que se baseia em reduzir a regressão às componentes de entrada e saída com maior covariância cruzada (ou seja, componentes de entrada que têm maior influência nas saída e componentes de saída mais fortemente influenciadas pelas entradas). Dentre todas as técnicas estudadas, esta é a única que exige que se considere o sistema como um modelo MIMO (*multiple inputs, multiple outputs*).

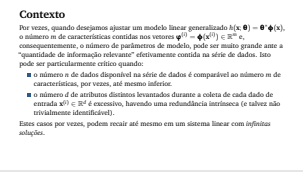
Regressão por Componentes Principais (PCR)

Contexto

Por vezes, quando desejamos ajustar um modelo linear generalizado $h(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\theta}^{*}\boldsymbol{\phi}(\mathbf{x})$, o número m de características contidas nos vetores $\boldsymbol{\varphi}^{(i)} = \boldsymbol{\phi}(\mathbf{x}^{(i)}) \in \mathbb{R}^{m}$ e, consequentemente, o número de parâmetros de modelo, pode ser muito grande ante a "quantidade de informação relevante" efetivamente contida na série de dados. Isto pode ser particularmente crítico quando:

- o número n de dados disponível na série de dados é comparável ao número m de características, por vezes, até mesmo inferior.
- o número d de atributos distintos levantados durante a coleta de cada dado de entrada $\mathbf{x}^{(i)} \in \mathbb{R}^{d}$ é excessivo, havendo uma redundância intrínseca (e talvez não trivialmente identificável).

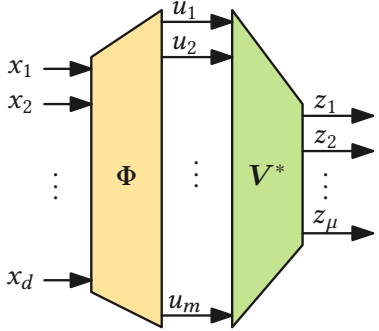
Estes casos por vezes, podem recair até mesmo em um sistema linear com *infinitas soluções*.



Assim como na formulação do problema de mínimos quadrados via pseudo-inversa, a metodologia da Análise de Componentes Principais (PCA) funciona igualmente para problemas que recaiam em qualquer um dos casos (I, II ou III) de solução de sistemas lineares, não exigindo que o usuário faça qualquer distinção *a priori* entre os casos para a aplicação dos algoritmos.

Análise de Componentes Principais (PCA)

O objetivo desta seção é apresentar uma metodologia, denominada Análise de Componentes Principais (*Principal Component Analysis* – PCA), utilizada para identificar *componentes significativas de sinal* em dados brutos de entrada, obtendo um *vetor de variáveis latentes* $\mathbf{z} \in \mathbb{R}^{\mu}$, $\mu \ll m$, que encapsule "a maior parte da informação relevante" originalmente disponível nos vetores de características $\boldsymbol{\varphi}^{(i)} = \boldsymbol{\phi}(\mathbf{x}^{(i)}) \in \mathbb{R}^{m}$.



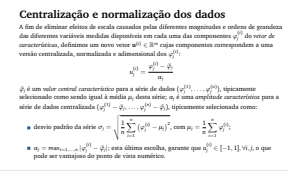
Centralização e normalização dos dados

A fim de eliminar efeitos de escala causados pelas diferentes magnitudes e ordens de grandeza das diferentes variáveis medidas disponíveis em cada uma das componentes $\varphi_j^{(i)}$ do *vetor de características*, definimos um novo vetor $\mathbf{u}^{(i)} \in \mathbb{R}^m$ cujas componentes correspondem a uma versão centralizada, normalizada e adimensional dos $\varphi_j^{(i)}$:

$$u_j^{(i)} = \frac{\varphi_j^{(i)} - \bar{\varphi}_j}{\alpha_j}$$

$\bar{\varphi}_j$ é um *valor central característico* para a série de dados $(\varphi_j^{(1)}, \ldots, \varphi_j^{(n)})$, tipicamente selecionado como sendo igual à média μ_j desta série; α_j é uma *amplitude característica* para a série de dados centralizada $(\varphi_j^{(1)} - \bar{\varphi}_j, \ldots, \varphi_j^{(n)} - \bar{\varphi}_j)$, tipicamente selecionada como:

- desvio padrão da série $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\varphi_j^{(i)} - \mu_j\right)^2}$, com $\mu_j = \frac{1}{n}\sum_{i=1}^{n}\varphi_j^{(i)}$;
- $\alpha_j = \max_{i=1,\ldots,n} |\varphi_j^{(i)} - \bar{\varphi}_j|$; esta última escolha, garante que $u_j^{(i)} \in [-1, 1], \forall i, j$, o que pode ser vantajoso do ponto de vista numérico.



Centralização e normalização dos dados

A fim de eliminar efeitos de escala causados pelas diferentes magnitudes e ordens de grandeza das diferentes variáveis medidas disponíveis em cada uma das componentes $\varphi_i^{(i)}$ do *vetor de características*, definimos um novo vetor $\boldsymbol{u}^{(i)} \in \mathbb{R}^m$ cujas componentes correspondem a uma versão centralizada, normalizada e adimensional dos $\varphi_i^{(i)}$:

$$u_i^{(i)} = \frac{\varphi_i^{(i)} - \bar{\varphi}_i}{a_i}$$

$\bar{\varphi}_i$ é um *valor central característico* para a série de dados $(\varphi_i^{(1)}, \ldots, \varphi_i^{(n)})$, tipicamente selecionado como sendo igual à média μ_i desta série; a_i é uma *amplitude característica* para a série de dados centralizada $(\varphi_i^{(1)} - \bar{\varphi}_i, \ldots, \varphi_i^{(n)} - \bar{\varphi}_i)$, tipicamente selecionada como:

- desvio padrão da série $\sigma_i = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\varphi_i^{(i)} - \mu_i)^2}$, com $\mu_i = \frac{1}{n}\sum_{i=1}^{n} \varphi_i^{(i)}$;
- $a_i = \max_{i=1,\ldots,n} |\varphi_i^{(i)} - \bar{\varphi}_i|$; esta última escolha, garante que $u_i^{(i)} \in [-1, 1]$, $\forall i, j$, o que pode ser vantajoso do ponto de vista numérico.

Para fins de regressão, também convém obter uma versão centralizada (e, eventualmente, normalizada) das saídas dada por:

$$s^{(i)} = \frac{y^{(i)} - \bar{y}}{\omega}$$

com $\bar{y}$ sendo um valor central característico e ω sendo uma amplitude característica, definidos de forma análoga ao que foi feito para cada componente do vetor de características.

Redução da dimensão do espaço de características

Objetivo: encontrar as direções do espaço de características, definidas por meio dos vetores unitários $\mathbf{v} \in \mathbb{R}^m$ (ou seja, com $\mathbf{v}^*\mathbf{v} = 1$), que estejam melhor representadas pelos dados de entrada centralizados e normalizados $\mathbf{u}^{(i)}$. Para tal, busquemos os pontos críticos da função objetivo:

$$L(\mathbf{v}, \lambda) = \frac{1}{2} \sum_{i=1}^{n} \left| \mathbf{v}^* \mathbf{u}^{(i)} \right|^2 - \frac{1}{2} \lambda (\mathbf{v}^* \mathbf{v} - 1)$$

O uso do multiplicador de Lagrange λ na função objetivo visa garantir a satisfação da condição $\mathbf{v}^*\mathbf{v} - 1 = 0$, necessária para que a busca por pontos críticos fique restrita a vetores unitários $\mathbf{v}$. Note que:

$$L(\mathbf{v}, \lambda) = \frac{1}{2} \mathbf{v}^* \left[\sum_{i=1}^{n} \mathbf{u}^{(i)} \mathbf{u}^{(i)*} \right] \mathbf{v} - \frac{1}{2} \lambda (\mathbf{v}^* \mathbf{v} - 1)$$

Redução da dimensão do espaço de características

Definindo:

$$\boldsymbol{U} = \begin{bmatrix} \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(n)} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

obtemos:

$$L(\mathbf{v}, \lambda) = \frac{1}{2}\mathbf{v}^* \boldsymbol{U}\boldsymbol{U}^* \mathbf{v} - \frac{1}{2}\lambda(\mathbf{v}^*\mathbf{v} - 1)$$

Os valores extremos de $L(\mathbf{v}, \lambda)$ ocorrem para:

$$\frac{\partial L}{\partial \mathbf{v}} = 0 \quad \Rightarrow \quad \boldsymbol{U}\boldsymbol{U}^*\mathbf{v} - \lambda\mathbf{v} = 0 \quad \Rightarrow \quad \boxed{(\boldsymbol{U}\boldsymbol{U}^* - \lambda \boldsymbol{I})\mathbf{v} = 0}$$

$$\frac{\partial L}{\partial \lambda} = 0 \quad \Rightarrow \quad -\frac{1}{2}(\mathbf{v}^*\mathbf{v} - 1) = 0 \quad \Rightarrow \quad \mathbf{v}^*\mathbf{v} = 1$$

Redução da dimensão do espaço de características

Temos assim que resolver o problema de autovalor da matriz $\boldsymbol{U}\boldsymbol{U}^*$ que é uma matriz simétrica e *semi-definida positiva* (de fato, $\mathbf{v}^*\boldsymbol{U}\boldsymbol{U}^*\mathbf{v} = (\boldsymbol{U}^*\mathbf{v})^*(\boldsymbol{U}^*\mathbf{v}) = \|\boldsymbol{U}^*\mathbf{v}\|^2 \geq 0, \forall \mathbf{v}$).

Em outras palavras, $\boldsymbol{U}\boldsymbol{U}^*$ não apenas é *diagonalizável*, como também todos os seus autovalores são *reais não-negativos* $\left(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0\right)$ e todos os autovetores unitários a eles associados constituem uma base ortonormal.

O posto de $\boldsymbol{U}\boldsymbol{U}^*$ corresponderá essencialmente ao número de autovalores não-nulos ($\boldsymbol{U}\boldsymbol{U}^*$ terá posto completo se, e somente se, seu menor autovalor for positivo, i.e. $\lambda_m > 0$). Assim, pode-se reescrever $\boldsymbol{U}\boldsymbol{U}^*$ na forma:

$$\boldsymbol{U}\boldsymbol{U}^* = \bar{\boldsymbol{V}}\boldsymbol{\Lambda}\bar{\boldsymbol{V}}^* = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_m \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^* \\ \mathbf{v}_2^* \\ \vdots \\ \mathbf{v}_m^* \end{bmatrix} = \sum_{j=1}^{m} \lambda_j \mathbf{v}_j \mathbf{v}_j^*$$

Redução da dimensão do espaço de características

Como os autovalores estão ordenados de forma descrescente, é possível truncar a soma acima após as μ primeiras parcelas, obtendo-se assim:

$$\boldsymbol{U}\boldsymbol{U}^* \approx \sum_{j=1}^{\mu} \lambda_j \mathbf{v}_j \mathbf{v}_j^*$$

A precisão desta aproximação pode ser avaliada a partir da medida de *resíduo*:

$$\kappa = 1 - \frac{\lambda_1 + \ldots + \lambda_\mu}{\lambda_1 + \ldots + \lambda_m} = 1 - \left[\sum_{j=1}^{\mu} \lambda_j\right] / \left[\sum_{j=1}^{m} \lambda_j\right]$$

Modelo de regressão de ordem reduzida

Começamos definindo o *vetor de variáveis latentes* $\mathbf{z} \in \mathbb{R}^{\mu}$ tal que $z_j = \mathbf{v}_j^* \mathbf{u}$, $j = 1, \ldots, \mu$, para todo *vetor de características* centralizado e normalizado $\mathbf{u} \in \mathbb{R}^m$, ou seja:

$$\mathbf{z} = \boldsymbol{V}^* \mathbf{u} \quad \text{com} \quad \boldsymbol{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_{\mu} \end{bmatrix} \in \mathbb{R}^{m \times \mu}$$

A ideia é obter um modelo de regressão linear de ordem reduzida em que a estimativa para $s^{(i)}$ possa ser obtida por meio de uma expressão da forma:

$$\hat{s}^{(i)} = \mathbf{b}^* \mathbf{z}^{(i)}$$

com $\mathbf{b} \in \mathbb{R}^{\mu}$ sendo o *vetor de parâmetros* que se deseja ajustar. Para tal, defina:

$$\boldsymbol{Z} = \boldsymbol{V}^* \boldsymbol{U}$$

com:

$$\boldsymbol{U} = \begin{bmatrix} \mathbf{u}^{(1)} & \ldots & \mathbf{u}^{(n)} \end{bmatrix} \in \mathbb{R}^{m \times n} \quad \text{e} \quad \boldsymbol{Z} = \begin{bmatrix} \mathbf{z}^{(1)} & \ldots & \mathbf{z}^{(n)} \end{bmatrix} \in \mathbb{R}^{\mu \times n}$$

Modelo de regressão de ordem reduzida

Recaímos assim na necessidade de encontrar uma solução para o problema definido pela função objetivo:

$$J_n(\mathbf{b}) = \frac{1}{2}\sum_{i=1}^{n}\left(s^{(i)} - \mathbf{b}^*\mathbf{z}^{(i)}\right)^2 + \frac{1}{2}\eta\,\mathbf{b}^*\mathbf{b} = \frac{1}{2}\|\mathbf{s} - \boldsymbol{Z}^*\mathbf{b}\|^2 + \frac{1}{2}\eta\|\mathbf{b}\|^2$$

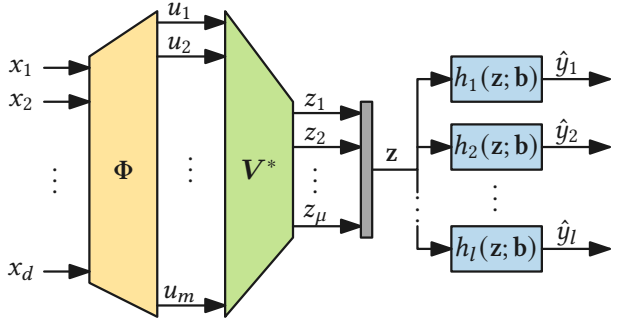
Dessa forma, o vetor de parâmetros b do modelo de regressão linear de ordem reduzida é dado por:

$$\mathbf{b} = \boldsymbol{Z}^{*\#}\mathbf{s}$$

Tipicamente, após a redução de ordem, o problema recai no caso I, em que *não há solução* $\mathbf{b}$ para o sistema linear $\boldsymbol{Z}^*\mathbf{b} = \mathbf{s}$. Neste caso, $\mathbf{b} = \boldsymbol{Z}^{*\#}\mathbf{s}$, com $\boldsymbol{Z}^{*\#} = (\boldsymbol{Z}\boldsymbol{Z}^*)^{-1}\boldsymbol{Z}$, é a solução que minimiza o erro quadrático $\|\mathbf{s} - \boldsymbol{Z}^*\mathbf{b}\|^2$ e, uma vez que $(\boldsymbol{Z}\boldsymbol{Z}^*)$ é invertível, pode-se tomar o *parâmetro de regularização* $\eta = 0$.

Modelo de regressão de ordem reduzida

Todos os algoritmos discutidos até o momento, baseiam-se em obter modelos matemáticos para série de dados $(\mathbf{x}^{(i)}, y^{(i)}), i = 1, 2, \ldots, n$ em que as saídas $y^{(i)}$ possam ser consideradas *escalares*. Neste caso, para um sistema com *múltiplas entradas* e *múltiplas saídas*, criam-se modelos matemáticos individualizados para cada uma das saídas cada um deles tendo as mesmas múltiplas entradas admitidas para o sistema. Para todos os algoritmos considerados até o presente momento, tal estratégia é válida.

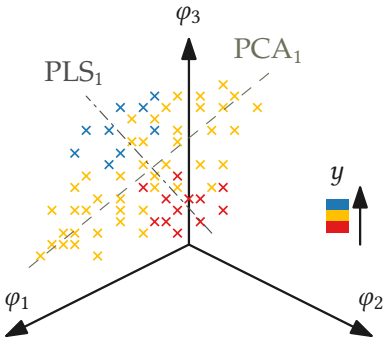


Regressão Parcial de Mínimos Quadrados (PLSR)

Contexto

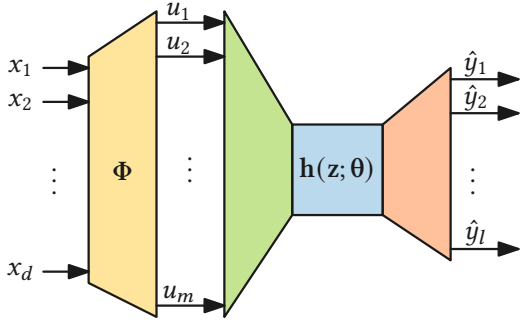
A Regressão Parcial de Mínimos Quadrados (*Partial Least Squares Regression* – PLSR) constitui uma estratégia alternativa para a obtenção de modelos matemáticos de ordem reduzida, aplicável a modelos MIMO, e que se baseia em identificar componentes de sinal de entrada que estejam mais fortemente correlacionadas com componentes de sinal observadas na saída do sistema.

Neste sentido, difere da estratégia da Regressão por Componentes Principais (PCR) tanto por atuar reduzindo a ordem do modelo simultaneamente nas entradas e nas saídas, quanto por priorizar as componentes de sinal entrada que terão maior influência na regressão, em vez de destacar as componentes principais que são as que se encontram melhor representadas na série de dados de entrada.



Objetivo

Dada uma série de dados $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}), i = 1, 2, \ldots, n$, em que $\mathbf{x}^{(i)} \in \mathbb{R}^d$ é o *vetor de entradas* ou *vetor de atributos*, $\boldsymbol{\varphi}^{(i)} = \boldsymbol{\phi}(\mathbf{x}^{(i)}) \in \mathbb{R}^m$ é o *vetor de características* e $\mathbf{y}^{(i)} \in \mathbb{R}^l$ é o *vetor de saídas*, obter um *modelo matemático de ordem reduzida* que seja capaz de estimar as saídas a partir do conhecimento das entradas, garantindo que o erro de estimativa seja mínimo sob uma métrica bem definida.



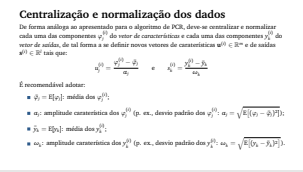
Centralização e normalização dos dados

De forma análoga ao apresentado para o algoritmo de PCR, deve-se centralizar e normalizar cada uma das componentes $\varphi_j^{(i)}$ do *vetor de características* e cada uma das componentes $y_k^{(i)}$ do *vetor de saídas*, de tal forma a se definir novos vetores de caraterísticas $\mathbf{u}^{(i)} \in \mathbb{R}^m$ e de saídas $\mathbf{s}^{(i)} \in \mathbb{R}^l$ tais que:

$$u_j^{(i)} = \frac{\varphi_j^{(i)} - \bar{\varphi}_j}{\alpha_j} \qquad \text{e} \qquad s_k^{(i)} = \frac{y_k^{(i)} - \bar{y}_k}{\omega_k}$$

É recomendável adotar:

- $\bar{\varphi}_j = \mathrm{E}[\varphi_j]$: média dos $\varphi_j^{(i)}$;
- α_j: amplitude caraterística dos $\varphi_j^{(i)}$ (p. ex., desvio padrão dos $\varphi_j^{(i)}$: $\alpha_j = \sqrt{\mathrm{E}\left[(\varphi_j - \bar{\varphi}_j)^2\right]}$);
- $\bar{y}_k = \mathrm{E}[y_k]$: média dos $y_k^{(i)}$;
- ω_k: amplitude caraterística dos $y_k^{(i)}$ (p. ex., desvio padrão dos $y_k^{(i)}$: $\omega_k = \sqrt{\mathrm{E}\left[(y_k - \bar{y}_k)^2\right]}$).


Centralização e normalização dos dados

De forma análoga ao apresentado para o algoritmo de PCR, deve-se centralizar e normalizar cada uma das componentes $\varphi_i^{(i)}$ do *vetor de características* e cada uma das componentes $y_k^{(i)}$ do vetor de saídas, de tal forma a se definir novos vetores de características $\mathbf{u}^{(i)} \in \mathbb{R}^m$ e de saídas $\mathbf{s}^{(i)} \in \mathbb{R}^l$ tais que:

$$u_i^{(i)} = \frac{\varphi_i^{(i)} - \bar{\varphi}_i}{a_i} \quad \text{e} \quad s_k^{(i)} = \frac{y_k^{(i)} - \bar{y}_k}{\omega_k}$$

É recomendável adotar:

- $\bar{\varphi}_i = \mathrm{E}[\varphi_i]$: média dos $\varphi_i^{(i)}$;
- a_i: amplitude característica dos $\varphi_i^{(i)}$ (p. ex., desvio padrão dos $\varphi_i^{(i)}$: $a_i = \sqrt{\mathrm{E}[(\varphi_i - \bar{\varphi}_i)^2]}$);
- $\bar{y}_k = \mathrm{E}[y_k]$: média dos $y_k^{(i)}$;
- ω_k: amplitude característica dos $y_k^{(i)}$ (p. ex., desvio padrão dos $y_k^{(i)}$: $\omega_k = \sqrt{\mathrm{E}[(y_k - \bar{y}_k)^2]}$).

Nesta seção, frequentemente usaremos a notação $\mathrm{E}[\cdot]$, que indica o valor esperado de uma *variável aleatória*. Do ponto de vista computacional, no entanto, se estamos baseando nossas estimativas em uma série de n dados, devemos recorrer à interpretação estatística:

$$\mathrm{E}[\#] = \frac{1}{n} \sum_{i=1}^{n} \#^{(i)}$$

Direções de máxima correlação

Identificar os vetores unitários $\mathbf{v} \in \mathbb{R}^m$, no espaço das características, e $\mathbf{w} \in \mathbb{R}^l$, no espaço das saídas, tais que seja máximo o valor esperado para a correlação entre as projeções $r \in \mathbb{R}$ e $z \in \mathbb{R}$ (também denominadas *scores*), com:

$$z^{(i)} = \mathbf{v}^* \mathbf{u}^{(i)} \qquad \text{e} \qquad r^{(i)} = \mathbf{w}^* \mathbf{s}^{(i)}$$

Notando que:

$$\mathrm{E}[zr] = \mathrm{E}[\mathbf{v}^* \mathbf{u} \mathbf{s}^* \mathbf{w}] = \mathbf{v}^* \mathrm{E}[\mathbf{u}\mathbf{s}^*] \mathbf{w}$$

Uma vez que $\mathbf{u}$ e $\mathbf{s}$ têm, por definição, média nula, então:

$$\boldsymbol{C} = \mathrm{E}[\mathbf{u}\mathbf{s}^*] = \begin{bmatrix} \mathrm{E}[u_1 s_1] & \dots & \mathrm{E}[u_1 s_l] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[u_m s_1] & \dots & \mathrm{E}[u_m s_l] \end{bmatrix} = \mathrm{cov}(\mathbf{u}, \mathbf{s}) \in \mathbb{R}^{m \times l}$$

é a matriz de covariância cruzada entre $\mathbf{u}$ e $\mathbf{s}$.

Direções de máxima correlação

Dessa forma, para encontrar os vetores unitários $\mathbf{v}$ e $\mathbf{w}$ que maximizam a correlação entre as projeções r e z, pode-se formular o problema a partir da seguinte função objetivo munida de multiplicadores de Lagrange λ_v e λ_w:

$$L(\mathbf{v}, \mathbf{w}, \lambda_v, \lambda_w) = \mathbf{v}^* \boldsymbol{C} \mathbf{w} - \frac{1}{2}\lambda_v(\mathbf{v}^*\mathbf{v} - 1) - \frac{1}{2}\lambda_w(\mathbf{w}^*\mathbf{w} - 1)$$

Os valores extremos de $L(\mathbf{v}, \mathbf{w}, \lambda_v, \lambda_w)$ ocorrem para:

$$\frac{\partial L}{\partial \mathbf{v}} = 0 \quad \Rightarrow \quad \boldsymbol{C}\mathbf{w} - \lambda_v \mathbf{v} = 0 \quad \Rightarrow \quad \mathbf{v} = \frac{1}{\lambda_v}\boldsymbol{C}\mathbf{w} \tag{1}$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \boldsymbol{C}^*\mathbf{v} - \lambda_w \mathbf{w} = 0 \quad \Rightarrow \quad \mathbf{w} = \frac{1}{\lambda_w}\boldsymbol{C}^*\mathbf{v} \tag{2}$$

$$\frac{\partial L}{\partial \lambda_v} = 0 \quad \Rightarrow \quad -\frac{1}{2}(\mathbf{v}^*\mathbf{v} - 1) = 0 \quad \Rightarrow \quad \mathbf{v}^*\mathbf{v} = 1 \tag{3}$$

$$\frac{\partial L}{\partial \lambda_w} = 0 \quad \Rightarrow \quad -\frac{1}{2}(\mathbf{w}^*\mathbf{w} - 1) = 0 \quad \Rightarrow \quad \mathbf{w}^*\mathbf{w} = 1 \tag{4}$$

Direções de máxima correlação

Das duas primeiras equações, recorrendo ao fato de que, para um escalar real qualquer, $\lambda^* = \lambda$, conclui-se que:

$$\mathbf{v} = \frac{1}{\lambda_v}\boldsymbol{C}\mathbf{w} \quad \Rightarrow \quad \lambda_v = \mathbf{v}^*\boldsymbol{C}\mathbf{w} \tag{5}$$

$$\mathbf{w} = \frac{1}{\lambda_w}\boldsymbol{C}^*\mathbf{v} \quad \Rightarrow \quad \lambda_w = \mathbf{w}^*\boldsymbol{C}^*\mathbf{v} = (\mathbf{v}^*\boldsymbol{C}\mathbf{w})^* = \lambda_v^* = \lambda_v \quad \Rightarrow \quad \boxed{\lambda_v = \lambda_w = \lambda} \tag{6}$$

Ainda, destas mesmas equações:

$$\boldsymbol{C}\boldsymbol{C}^*\mathbf{v} = \lambda_v\lambda_w\mathbf{v} = \lambda^2\mathbf{v} \tag{7}$$

$$\boldsymbol{C}^*\boldsymbol{C}\mathbf{w} = \lambda_v\lambda_w\mathbf{w} = \lambda^2\mathbf{w} \tag{8}$$

Em outras palavras:

- $\mathbf{v}$ deve ser um autovetor de $\boldsymbol{C}\boldsymbol{C}^* \in \mathbb{R}^{m\times m}$;
- $\mathbf{w}$ deve ser um autovetor de $\boldsymbol{C}^*\boldsymbol{C} \in \mathbb{R}^{l\times l}$.

Decomposição em valores singulares (SVD)

A decomposição em valores singulares (*Singular Value Decomposition* – SVD) permite descrever uma matriz $\boldsymbol{C} \in \mathbb{R}^{m \times l}$ na forma:

$$\boldsymbol{C} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{W}^*$$

com:

- $\boldsymbol{D} \in \mathbb{R}^{m \times l}$: matriz diagonal formada pelos *valores singulares* de $\boldsymbol{C}$, ou seja, pelas raízes quadradas λ não-negativas dos autovalores de $\boldsymbol{C}\boldsymbol{C}^* \in \mathbb{R}^{m \times m}$ ou $\boldsymbol{C}^*\boldsymbol{C} \in \mathbb{R}^{l \times l}$.
- $\boldsymbol{V} = [\mathbf{v}_1 \ \ldots \ \mathbf{v}_m] \in \mathbb{R}^{m \times m}$: matriz quadrada formada pelos vetores singulares à esquerda de $\boldsymbol{C}$ (ou seja, autovetores de $\boldsymbol{C}\boldsymbol{C}^*$)
- $\boldsymbol{W} = [\mathbf{w}_1 \ \ldots \ \mathbf{w}_l] \in \mathbb{R}^{l \times l}$: matriz quadrada formada pelos vetores singulares à direita de $\boldsymbol{C}$ (ou seja, autovetores de $\boldsymbol{C}^*\boldsymbol{C}$).

Teorema: admitindo que os valores singulares de $\boldsymbol{C}$ estejam ordenados de forma *decrescente*, então $L(\mathbf{v}, \mathbf{w}, \lambda_v, \lambda_w)$ será máxima para $\mathbf{v} = \mathbf{v}_1$ e $\mathbf{w} = \mathbf{w}_1$.

Predição a partir de uma variável latente

Baseado no escalar $z^{(i)} = \mathbf{v}_1^* \mathbf{u}^{(i)}$, obter uma estimativa $\hat{\mathbf{s}}^{(i)} = z^{(i)}\mathbf{q} = \mathbf{q}\mathbf{v}_1^*\mathbf{u}^{(i)}$ para $\mathbf{s}^{(i)}$ de forma a minimizar $\mathrm{E}\left[\|\mathbf{s} - z\mathbf{q}\|^2\right]$. Notanto que:

$$\begin{aligned}\mathrm{E}\left[\|\mathbf{s} - z\mathbf{q}\|^2\right] = \mathrm{E}\left[(\mathbf{s} - z\mathbf{q})^*(\mathbf{s} - z\mathbf{q})\right] &= \mathrm{E}\left[\mathbf{s}^*\mathbf{s}\right] - 2\mathbf{q}^*\mathrm{E}\left[\mathbf{s}z\right] + \mathbf{q}^*\mathbf{q}\mathrm{E}\left[zz\right] \\ &= \mathrm{E}\left[\mathbf{s}^*\mathbf{s}\right] - 2\mathbf{q}^*\mathrm{E}\left[\mathbf{s}\mathbf{u}^*\right]\mathbf{v}_1 + \mathbf{q}^*\mathbf{q}\mathbf{v}_1^*\mathrm{E}\left[\mathbf{u}\mathbf{u}^*\right]\mathbf{v}_1\end{aligned}$$

Assim:

$$\frac{\partial}{\partial \mathbf{q}}\mathrm{E}\left[\|\mathbf{s} - z\mathbf{q}\|^2\right] = 0 \quad \Rightarrow \quad -2\mathrm{E}\left[\mathbf{s}\mathbf{u}^*\right]\mathbf{v}_1 + 2\mathbf{q}\mathbf{v}_1^*\mathrm{E}\left[\mathbf{u}\mathbf{u}^*\right]\mathbf{v}_1 = 0$$

Notando que:

$$\mathrm{E}\left[\mathbf{s}\mathbf{u}^*\right] = \mathrm{cov}(\mathbf{s}, \mathbf{u}) = \boldsymbol{C}^* \in \mathbb{R}^{l \times m} \quad \text{e} \quad \mathrm{E}\left[\mathbf{u}\mathbf{u}^*\right] = \mathrm{cov}(\mathbf{u}, \mathbf{u}) = \boldsymbol{K} \in \mathbb{R}^{m \times m}$$

obtém-se então a expressão para o vetor $\mathbf{q} \in \mathbb{R}^l$, denominado carga de saída (*output loading*):

$$\mathbf{q} = \frac{1}{\mathbf{v}_1^* \boldsymbol{K} \mathbf{v}_1}\boldsymbol{C}^*\mathbf{v}_1 \quad \Rightarrow \quad \hat{\mathbf{s}} = z\mathbf{q} = \frac{1}{\mathbf{v}_1^* \boldsymbol{K} \mathbf{v}_1}\boldsymbol{C}^*\mathbf{v}_1\mathbf{v}_1^*\mathbf{u}$$

Redução dos dados

O par de vetores singulares $(\mathbf{v}_2, \mathbf{w}_2)$ não corresponde à segunda variável latente do sistema (ou seja, não fornecem as segundas componentes de sinais de entrada/saída com maior correlação cruzada).

Assim, o procedimento de decomposição em valores singulares deve ser repetido para um novo conjunto de dados, obtido a partir do original por meio da remoção das componentes de carga associadas à variável latente já obtida, tanto nos dados de entrada como nos de saída. Tal procedimento é denominado redução dos dados (*deflation*).

Para a obtenção do vetor $\mathbf{p} \in \mathbb{R}^m$, denominado carga de entrada (*input loading*), deve-se minimizar:

$$\begin{aligned} \mathrm{E}\left[\|\mathbf{u} - z\mathbf{p}\|^2\right] &= \mathrm{E}\left[(\mathbf{u} - z\mathbf{p})^*(\mathbf{u} - z\mathbf{p})\right] \\ &= \mathrm{E}\left[\mathbf{u}^*\mathbf{u}\right] - 2\mathbf{p}^*\mathrm{E}\left[\mathbf{u}\mathbf{u}^*\right]\mathbf{v}_1 + \mathbf{p}^*\mathbf{p}\mathbf{v}_1^*\mathrm{E}\left[\mathbf{u}\mathbf{u}^*\right]\mathbf{v}_1 \end{aligned}$$

Redução dos dados

Tomando a derivada desta última expressão com respeito a p e igualando a zero, obtém-se:

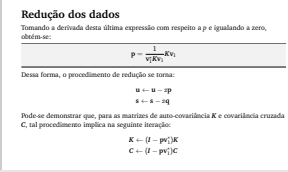
$$\mathbf{p} = \frac{1}{\mathbf{v}_1^* \boldsymbol{K} \mathbf{v}_1} \boldsymbol{K} \mathbf{v}_1$$

Dessa forma, o procedimento de redução se torna:

$$\mathbf{u} \leftarrow \mathbf{u} - z\mathbf{p}$$
$$\mathbf{s} \leftarrow \mathbf{s} - z\mathbf{q}$$

Pode-se demonstrar que, para as matrizes de auto-covariância $\boldsymbol{K}$ e covariância cruzada $\boldsymbol{C}$, tal procedimento implica na seguinte iteração:

$$\boldsymbol{K} \leftarrow (\boldsymbol{I} - \mathbf{p}\mathbf{v}_1^*)\boldsymbol{K}$$
$$\boldsymbol{C} \leftarrow (\boldsymbol{I} - \mathbf{p}\mathbf{v}_1^*)\boldsymbol{C}$$


Redução dos dados

Tomando a derivada desta última expressão com respeito a p e igualando a zero, obtém-se:

$$\mathbf{p} = \frac{1}{\mathbf{v}_1^* \boldsymbol{K} \mathbf{v}_1} \boldsymbol{K} \mathbf{v}_1$$

Dessa forma, o procedimento de redução se torna:

$$\mathbf{u} \leftarrow \mathbf{u} - z\mathbf{p}$$
$$\mathbf{s} \leftarrow \mathbf{s} - z\mathbf{q}$$

Pode-se demonstrar que, para as matrizes de auto-covariância $\boldsymbol{K}$ e covariância cruzada $\boldsymbol{C}$, tal procedimento implica na seguinte iteração:

$$\boldsymbol{K} \leftarrow (I - \mathbf{p}\mathbf{v}_1^*)\boldsymbol{K}$$
$$\mathbf{c} \leftarrow (I - \mathbf{p}\mathbf{v}_1^*)\mathbf{c}$$

A nova expressão para $\boldsymbol{K}$ após a redução é dada por:

$$\begin{aligned}
\mathrm{E}\big[(\mathbf{u} - z\mathbf{p})(\mathbf{u} - z\mathbf{p})^*\big] &= \mathrm{E}\big[\mathbf{u}\mathbf{u}^*\big] - \mathbf{p}\mathrm{E}\big[z\mathbf{u}^*\big] - \mathrm{E}\big[\mathbf{u}z\big]\mathbf{p}^* + \mathbf{p}\mathrm{E}\big[zz\big]\mathbf{p}^* \\
&= \mathrm{E}\big[\mathbf{u}\mathbf{u}^*\big] - \mathbf{p}\mathbf{v}_1^*\mathrm{E}\big[\mathbf{u}\mathbf{u}^*\big] - \mathrm{E}\big[\mathbf{u}\mathbf{u}^*\big]\mathbf{v}_1\mathbf{p}^* + \mathbf{p}\mathbf{v}_1^*\mathrm{E}\big[\mathbf{u}\mathbf{u}^*\big]\mathbf{v}_1\mathbf{p}^* \\
&= \boldsymbol{K} - \mathbf{p}\mathbf{v}_1^*\boldsymbol{K} - \boldsymbol{K}\mathbf{v}_1\mathbf{p}^* + \underbrace{\mathbf{p}\mathbf{v}_1^*\boldsymbol{K}\mathbf{v}_1}_{\boldsymbol{K}\mathbf{v}_1}\mathbf{p}^* = (I - \mathbf{p}\mathbf{v}_1^*)\boldsymbol{K}
\end{aligned}$$

Analogamente, a nova expressão para $\boldsymbol{C}$ após a redução é:

$$\begin{aligned}
\mathrm{E}\big[(\mathbf{u} - z\mathbf{p})(\mathbf{s} - z\mathbf{q})^*\big] &= \mathrm{E}\big[\mathbf{u}\mathbf{s}^*\big] - \mathbf{p}\mathrm{E}\big[z\mathbf{s}^*\big] - \mathrm{E}\big[\mathbf{u}z\big]\mathbf{q}^* + \mathbf{p}\mathrm{E}\big[zz\big]\mathbf{q}^* \\
&= \mathrm{E}\big[\mathbf{u}\mathbf{s}^*\big] - \mathbf{p}\mathbf{v}_1^*\mathrm{E}\big[\mathbf{u}\mathbf{s}^*\big] - \mathrm{E}\big[\mathbf{u}\mathbf{u}^*\big]\mathbf{v}_1\mathbf{q}^* + \mathbf{p}\mathbf{v}_1^*\mathrm{E}\big[\mathbf{u}\mathbf{u}^*\big]\mathbf{v}_1\mathbf{q}^* \\
&= \boldsymbol{C} - \mathbf{p}\mathbf{v}_1^*\boldsymbol{C} - \boldsymbol{K}\mathbf{v}_1\mathbf{q}^* + \underbrace{\mathbf{p}\mathbf{v}_1^*\boldsymbol{K}\mathbf{v}_1}_{\boldsymbol{K}\mathbf{v}_1}\mathbf{q}^* = (I - \mathbf{p}\mathbf{v}_1^*)\boldsymbol{C}
\end{aligned}$$

Avaliação do resíduo e iteração

É possível avaliar os resíduos a cada iteração a partir dos valores dos traços da matriz de auto-covariância $\boldsymbol{K}$ das entradas, uma vez que:

$$R = \mathrm{E}[\|\mathbf{u}\|^2] = \mathrm{E}[\mathbf{u}^*\mathbf{u}] = \mathrm{tr}\big(\mathrm{E}[\mathbf{u}\mathbf{u}^*]\big) = \mathrm{tr}\big(\boldsymbol{K}\big)$$

Em particular, a métrica para o resíduo consiste em medir a razão entre o valor de R após k iterações e o valor no início do problema R_0:

$$\kappa = \frac{R}{R_0}$$

O monitoramento da variação de κ em função do número de passos de iteração k nos permite concluir se há ou não necessidade de iterar. Se houver, voltamos ao primeiro passo, encontrando as novas direções de máxima correlação, porém desta vez trocando tanto os dados como as respectivas matrizes de covariância pelas suas versões reduzidas, obtidas no procedimento de redução.

Avaliação do resíduo e iteração

Em geral, conclui-se que não é mais viável proceder com iterações se κ atingir um valor estacionário após um dado número μ de iterações. O número μ será então a ordem do modelo reduzido e os respectivos resíduos em $\mathbf{u}$ e $\mathbf{s}$ após μ iterações:

$$\mathbf{u}^{(i)}[\mu] \leftarrow \mathbf{u}^{(i)} - \sum_{k=1}^{\mu} z^{(i)}[k]\mathbf{p}[k] \qquad \text{e} \qquad \mathbf{s}^{(i)}[\mu] \leftarrow \mathbf{s}^{(i)} - \sum_{k=1}^{\mu} z^{(i)}[k]\mathbf{q}[k]$$

podem ser interpretados como *ruídos* nos dados, de tal forma que o modelo de ordem reduzida para o sistema se torna:

$$\mathbf{u}^{(i)} = \sum_{k=1}^{\mu} z^{(i)}[k]\mathbf{p}[k] + \mathbf{e}^{(i)} \qquad \text{e} \qquad \mathbf{s}^{(i)} = \sum_{k=1}^{\mu} z^{(i)}[k]\mathbf{q}[k] + \mathbf{f}^{(i)}$$

com $z^{(i)}[k] = (\mathbf{v}_1[k])^* \mathbf{u}^{(i)}$ e com $\mathbf{e}^{(i)}$ e $\mathbf{f}^{(i)}$ representando os respectivos erros de aproximação.

Representação matricial

Defina-se:

$$\boldsymbol{U} = \begin{bmatrix} \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(n)} \end{bmatrix} \in \mathbb{R}^{m \times n} \qquad \boldsymbol{S} = \begin{bmatrix} \mathbf{s}^{(1)} & \dots & \mathbf{s}^{(n)} \end{bmatrix} \in \mathbb{R}^{l \times n}$$

$$\boldsymbol{E} = \begin{bmatrix} \mathbf{e}^{(1)} & \dots & \mathbf{e}^{(n)} \end{bmatrix} \in \mathbb{R}^{m \times n} \qquad \boldsymbol{F} = \begin{bmatrix} \mathbf{f}^{(1)} & \dots & \mathbf{f}^{(n)} \end{bmatrix} \in \mathbb{R}^{l \times n}$$

$$\boldsymbol{P} = \begin{bmatrix} \mathbf{p}[1] & \dots & \mathbf{p}[\mu] \end{bmatrix} \in \mathbb{R}^{m \times \mu} \qquad \boldsymbol{Q} = \begin{bmatrix} \mathbf{q}[1] & \dots & \mathbf{q}[\mu] \end{bmatrix} \in \mathbb{R}^{l \times \mu}$$

$$\boldsymbol{Z} = \begin{bmatrix} z^{(1)}[1] & \dots & z^{(n)}[1] \\ \vdots & \ddots & \vdots \\ z^{(1)}[\mu] & \dots & z^{(n)}[\mu] \end{bmatrix} \in \mathbb{R}^{\mu \times n}$$

o modelo pode ser reescrito em forma matricial como:

$$\boldsymbol{U} = \boldsymbol{P}\boldsymbol{Z} + \boldsymbol{E}$$
$$\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{Z} + \boldsymbol{F}$$

Algoritmo da PLSR

Inicialização:

$$
\begin{aligned}
\boldsymbol{U} &\leftarrow \begin{bmatrix} \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(n)} \end{bmatrix} \\
\boldsymbol{S} &\leftarrow \begin{bmatrix} \mathbf{s}^{(1)} & \dots & \mathbf{s}^{(n)} \end{bmatrix} \\
\boldsymbol{K} &\leftarrow \frac{1}{n} \boldsymbol{U}\boldsymbol{U}^{*} \\
\boldsymbol{C} &\leftarrow \frac{1}{n} \boldsymbol{U}\boldsymbol{S}^{*} \\
R_0 &\leftarrow \operatorname{tr}(\boldsymbol{K})
\end{aligned}
$$

Inicializar com zeros matrizes $\boldsymbol{P} \in \mathbb{R}^{m \times \mu}$, $\boldsymbol{Q} \in \mathbb{R}^{l \times \mu}$, $\boldsymbol{V}_1 \in \mathbb{R}^{m \times \mu}$ e um vetor $\boldsymbol{\kappa} \in \mathbb{R}^{\mu}$.

Algoritmo da PLSR

Repetir (*loop*) para $k = 1, \ldots, \mu$:

$$
\begin{aligned}
\boldsymbol{V}_1[,k] &\leftarrow \text{svd}(\boldsymbol{C})\$\mathbf{u}[,1] \\
g &\leftarrow 1/\left((\boldsymbol{V}_1[,k])^* \boldsymbol{K} \mathbf{V}_1[,k]\right) \\
\boldsymbol{P}[,k] &\leftarrow g\boldsymbol{K}\boldsymbol{V}_1[,k] \\
\boldsymbol{Q}[,k] &\leftarrow g\boldsymbol{C}^*\boldsymbol{V}_1[,k] \\
\boldsymbol{K} &\leftarrow \left(\boldsymbol{I} - \boldsymbol{P}[,k](\boldsymbol{V}_1[,k])^*\right)\boldsymbol{K} \\
\boldsymbol{C} &\leftarrow \left(\boldsymbol{I} - \boldsymbol{P}[,k](\boldsymbol{V}_1[,k])^*\right)\boldsymbol{C} \\
\boldsymbol{\kappa}[k] &\leftarrow \text{tr}(\boldsymbol{K})/R_0
\end{aligned}
$$

e, caso desejemos fazer a redução dos dados, devemos inicializar com zeros uma matriz $\boldsymbol{Z} \in \mathbb{R}^{\mu \times n}$, e em seguida, repetir (*loop*) para $k = 1, \ldots, \mu$:

$$
\begin{aligned}
\boldsymbol{Z}[k,] &\leftarrow (\boldsymbol{V}_1[,k])^* \boldsymbol{U} \\
\boldsymbol{U} &\leftarrow \boldsymbol{U} - \boldsymbol{P}[,k]\boldsymbol{Z}[k,] \\
\boldsymbol{S} &\leftarrow \boldsymbol{S} - \boldsymbol{Q}[,k]\boldsymbol{Z}[k,]
\end{aligned}
$$