2
Regressão por mínimos quadrados parciais

O algoritmo para regressão por mínimos quadrados parciais, ou PLS[1], proposto por S. Wold [56, 57], compõe um método para regressão em fatores cujo objetivo é a predição de um conjunto de variáveis de saída Y baseado na observação de um conjunto de variáveis de entrada X. Este método é consolidado na área de quimiometria para a análise de cromatografias e espectrometrias [22, 23]. Ele tem sido aplicado em outras áreas como monitoramento de processos, marketing ou processamento de imagens [43, 32, 33] devido à robustez do modelo gerado num cenário de muitas variáveis e poucas amostras [52]. Neste capítulo, é explicada a regressão por mínimos quadrados parciais, mostrando a particularidade do modelo, os algoritmos de base usados nesta tese seguido de um exemplo de uso.

2.1
Modelo

A construção de um modelo PLS requer um conjunto de observações (padrões ou amostras) juntamente com o valor das variáveis dependentes. Seja X a matriz contendo as amostras em suas linhas, e Y a matriz contendo os valores para predição em suas linhas.

A regressão PLS modela simultaneamente os fatores (ou variáveis latentes) inerentes tanto em X quanto em Y. Estes fatores são então usados para definir um sub-espaço em X que melhor se adapte à modelagem de Y. Com a regressão por componentes principais (PCR) [1, 30, 26], a rotação definida pelos autovetores é usada para encontrar um sub-espaço em X que subseqüentemente é usado para modelar Y. A abordagem tomada com PLS é muito parecida com a da análise em componentes principais [28] exceto que os fatores são escolhidos de forma a descrever tanto as variáveis em Y quanto em X. Isto é conseguido usando as colunas da matriz Y para estimar

[1] Partial Least Squares.

os fatores em X. Ao mesmo tempo, as colunas de X são usadas para estimar os fatores em Y. Os modelos resultantes são

$$\mathsf{X} = \mathsf{TP} + \mathsf{E}$$

$$\mathsf{Y} = \mathsf{UQ} + \mathsf{F}$$

onde as colunas de T e U são chamadas de escores de X e Y respectivamente, e as de P e Q cargas. As matrizes E e F representam o erro associado à modelagem PLS de X e de Y.

Os fatores em T não são ótimos para estimação das colunas de X como é o caso com PCR, mas sofrem uma rotação de forma a que descrevam simultaneamente a matriz Y. Na situação ideal, as fontes de variação em X são exatamente as mesmas do que as fontes de variação em Y, resultando em fatores idênticos para X e Y. Na prática, X varia de forma não correlacionada com a variação em Y, e desta forma existem t e u tais que $\mathsf{t} \neq \mathsf{u}$, onde t e u são colunas de T e U respectivamente. Entretanto, quando ambas as matrizes são usadas para estimar os fatores de X e Y, estes seguem a relação:

$$\mathsf{u} = b\mathsf{t} + \epsilon$$

onde b representa a relação interna entre u e t. A principal vantagem da regressão PLS está na incorporação de maior informação na fase de modelagem, resultando num modelo mais compacto para predição quando comparado com outros métodos [31, 52].

Os algoritmos apresentados a seguir não seguem exatamente a representação encontrada na literatura por dois motivos. As versões desta tese se distiguem em primeiro, por fornecerem equações mais simples porém menos didáticas. Em segundo por não incorporarem a capacidade de trabalhar com dados ausentes. Historicamente a regressão PLS nasceu do algoritmo NIPALS[2] de H. Wold [55] para cálculo de componentes principais levando em conta dados ausentes. Como o foco desta tese está no tratamento de grandes conjuntos de dados e na melhora do poder de predição, os algoritmos são apresentados de forma mais simples. Por outro lado, seguindo a convenção encontrada na literatura, são apresentadas duas formas do algoritmo: a primeira, PLS1 para o caso de apenas uma variável dependente e a segunda, PLS2 para os demais casos. Esta distinção permite algumas simplificações para o PLS1 como é mostrado a seguir.

[2]Nonlinear estimation by Iterative Partial Least Squares.

2.2 Algoritmo

A regressão PLS pode ser decomposta nas seguintes etapas:

1. Determinação da estrutura latente
   Dado um conjunto de dados para treinamento, o modelo de regressão é construído. Esta etapa é chamada de calibração ou treinamento;

2. Seleção de fatores
   Dado um segundo conjunto independente, chamado de conjunto de teste, predições são realizadas variando o número de fatores. O número de fatores que fornecer a melhor predição é usado para o modelo. Esta etapa corresponde à validação do modelo.

Seguem os algoritmos PLS2 e PLS1 para treinamento juntamente com o algoritmo para predição.

2.2.1 Regressão PLS2

Sejam X e Y matrizes de dimensão $(n \times m)$ e $(n \times l)$ respectivamente, com as amostras e valores a serem preditos. Para o caso de a matriz Y possuir mais de uma variável, o algoritmo da figura 2.1 é empregado. Para o caso de a matriz X ser de posto cheio, o número k de fatores calculados poder ser igual a m sem problemas de degeneração.

O trecho entre as linhas 4 e 11 corresponde originalmente ao algoritmo NIPALS para cálculo dos autovetores w_i de $\mathsf{X}_i\mathsf{X}_i^\top\mathsf{Y}_i\mathsf{Y}_i^\top\mathsf{X}_i$ via método de potência [17, 21]. Na convergência, o uso da carga[3] w_i para o cálculo do fator t_i comprova o compromisso na explicação entre as variáveis de X e Y.

A regressão de Y em t_i é realizada na linha 12, onde cada elemento de b_i corresponde ao coeficiente de regressão linear por mínimos quadrados da variável correspondente de Y em t_i. Da mesma forma, cada elemento do vetor p_i corresponde ao coeficiente de regressão linear por mínimos quadrados das variáveis de X em t_i. Estas características garantem (seção 2.3.3) que os resíduos calculados nas linhas 14 e 15 são ortogonais a t_i e subseqüentemente que o conjunto de fatores t_i gerado é ortogonal. O cálculo residual das linhas 14 e 15 é feito descontando o fator t_i das matrizes X e Y.

[3] ponderação das variáveis originais para estimação de t_i

	Algoritmo para regressão PLS2
1	$\mathsf{X}_1 \leftarrow \mathsf{X}; \mathsf{Y}_1 \leftarrow \mathsf{Y}$
2	for $\mathtt{i} = 1$ to $\mathtt{k}$
3	$\mathsf{u} \leftarrow$ primeira coluna de Y_i
4	repetir
5	$\mathsf{w}_i \leftarrow \mathsf{X}_i^\top \mathsf{u}$
6	normalizar w_i
7	$\mathsf{t}_i \leftarrow \mathsf{X}_i \mathsf{w}_i$
8	$\mathsf{q} \leftarrow \mathsf{Y}_i^\top \mathsf{t}_i$
9	normalizar q
10	$\mathsf{u} \leftarrow \mathsf{Y}\mathsf{q}$
11	até convergência de w_i
	// Cálculo dos coeficientes
12	$\mathsf{b}_i \leftarrow \mathsf{Y}_i^\top \mathsf{t}_i / \mathsf{t}_i^\top \mathsf{t}_i$
13	$\mathsf{p}_i \leftarrow \mathsf{X}_i^\top \mathsf{t}_i / \mathsf{t}_i^\top \mathsf{t}_i$
	// Cálculo residual
14	$\mathsf{X}_{i+1} \leftarrow \mathsf{X}_i - \mathsf{t}_i \mathsf{p}_i^\top$
15	$\mathsf{Y}_{i+1} \leftarrow \mathsf{Y}_i - \mathsf{t}_i \mathsf{b}_i^\top$
16	end

Figura 2.1: Algoritmo PLS2.

2.2.2 Regressão PLS1

Quando é desejada a predição de apenas uma variável, o algoritmo para regressão por mínimos quadrados parciais apresentado em 2.1 pode ser simplificado. A maior simplificação, que é apresentada na figura 2.2, consiste no cálculo do autovetor w_i. Dado que Y possui apenas uma coluna, a matriz $\mathsf{X}^\top\mathsf{Y}\mathsf{Y}^\top\mathsf{X}$ possui posto igual a 1 porque pode ser expressa na forma $\mathsf{v}\mathsf{v}^\top$, onde $\mathsf{v} = \mathsf{X}^\top\mathsf{Y}$. Logo, seu autovetor não normalizado é trivialmente dado por

$$\mathsf{w}_i = \mathsf{X}^\top\mathsf{Y} \tag{2-1}$$

com autovalor

$$\lambda_i = \mathsf{Y}^\top\mathsf{X}\mathsf{X}^\top\mathsf{Y}$$

Esta simplificação se reflete nas linhas 4 a 11 do algoritmo da figura 2.1. O restante do algoritmo permanece igual com o único detalhe de que o coeficiente de regressão linear b_i passa a ser um escalar.

	Algoritmo para regressão PLS1
1	$\mathsf{X}_1 \leftarrow \mathsf{X}; \mathsf{Y}_1 \leftarrow \mathsf{Y}$
2	for $\mathtt{i} = 1$ to $\mathtt{k}$
3	$\mathsf{w}_i \leftarrow \mathsf{X}_i^\top \mathsf{Y}$
4	$\mathsf{w}_i \leftarrow \mathsf{w}_i/(\mathsf{Y}^\top \mathsf{X}\mathsf{X}^\top \mathsf{Y})^{1/2}$ // normalização de w_i
5	$\mathsf{t}_i \leftarrow \mathsf{X}_i \mathsf{w}_i$
	// Cálculo dos coeficientes
6	$b_i \leftarrow \mathsf{Y}_i^\top \mathsf{t}_i / \mathsf{t}_i^\top \mathsf{t}_i$
7	$\mathsf{p}_i \leftarrow \mathsf{X}_i^\top \mathsf{t}_i / \mathsf{t}_i^\top \mathsf{t}_i$
	// Cálculo residual
8	$\mathsf{X}_{i+1} \leftarrow \mathsf{X}_i - \mathsf{t}_i \mathsf{p}_i^\top$
9	$\mathsf{Y}_{i+1} \leftarrow \mathsf{Y}_i - b_i \mathsf{t}_i$
10	end

Figura 2.2: Algoritmo PLS1.

2.2.3 Predição

Uma vez calculado o modelo PLS para regressão com um dos algoritmos apresentados para PLS1 ou PLS2, podemos aplicá-lo a um novo conjunto de amostras X' para predição das variáveis dependentes Y'. Esta tarefa é realizada com w, p e b obtidos na fase de treinamento, decompondo a matriz X' nos fatores estimados t' e, em seguida, realizando a predição com estes. O algoritmo é mostrado na figura 2.3.

Repare que o número de amostras em X' não é necessariamente igual ao de X. Observe também que determinar o número de fatores h usados para predição, requer um procedimento de validação mostrado com um exemplo na seção a seguir.

2.3 Propriedades

A regressão por mínimos quadrados parciais possui um grande conjunto de propriedades [21, 52]. Dentre estas, algumas importantes para o desenvolvimento dos algoritmos desta tese, são mostradas a seguir.

Algoritmo para predição PLS	
1	$\mathsf{X}'_1 \leftarrow \mathsf{X}'$
2	$\mathsf{Y}' \leftarrow 0$
3	for $\mathtt{i} = 1$ to $\mathtt{h}$
4	$\mathsf{t}'_i \leftarrow \mathsf{X}'_i\mathsf{w}_i$
5	$\mathsf{Y}' \leftarrow \mathsf{Y}' + \mathsf{t}'_i\mathsf{b}_i^\top$
	// Cálculo residual
6	$\mathsf{X}'_{i+1} \leftarrow \mathsf{X}'_i - \mathsf{t}'_i\mathsf{p}_i^\top$
7	end

Figura 2.3: Predição PLS.

2.3.1 Estrutura Latente

Na convergência ao final da linha 11 para o algoritmo PLS2 ou linha 4 para o algoritmo PLS1, w_i corresponde ao autovetor normalizado de $\mathsf{X}_i^\top\mathsf{Y}_i\mathsf{Y}_i^\top\mathsf{X}_i$ associado ao maior autovalor. Isto é, $\mathsf{X}_i^\top\mathsf{Y}_i\mathsf{Y}_i^\top\mathsf{X}_i\mathsf{w}_i = \lambda_i\mathsf{w}_i$ com $\|\mathsf{w}_i\| = 1$. Desta forma, os fatores PLS revelam a estrutura latente da matriz $\mathsf{X}_i^\top\mathsf{Y}_i\mathsf{Y}_i^\top\mathsf{X}_i$.

Justamente, o uso da carga w_i para o cálculo dos escores t_i, comprova o compromisso do modelo PLS entre a explicação das variáveis de X_i e Y_i. De fato, na convergência temos também

$$\mathsf{X}_i\mathsf{X}_i^\top\mathsf{Y}_i\mathsf{Y}_i^\top\mathsf{X}_i\mathsf{w}_i = \lambda\mathsf{X}_i\mathsf{w}_i$$

ou seja

$$\mathsf{X}_i\mathsf{X}_i^\top\mathsf{Y}_i\mathsf{Y}_i^\top\mathsf{t}_i = \lambda\mathsf{t}_i$$

e de forma semelhante

$$\mathsf{Y}_i^\top\mathsf{X}_i\mathsf{X}_i^\top\mathsf{Y}_i\mathsf{b}_i = \lambda\mathsf{b}_i$$

Outra abordagem [21], reforça este ponto mostrando que a cada iteração do PLS, as variáveis latentes ou componentes escolhidos em X e Y possuem máxima covariância, ou seja a seguinte expressão é maximizada

$$\mathrm{Cov}(\mathsf{f}, \mathsf{g}) = \mathsf{f}^\top\mathsf{g}/n \tag{2-2}$$

com

$$\begin{aligned} \mathsf{f} &= \mathsf{X}\mathsf{d} & |\mathsf{d}| &= 1 \\ \mathsf{g} &= \mathsf{Y}\mathsf{e} & |\mathsf{e}| &= 1 \end{aligned}$$

PLS maximiza 2-2 escolhendo $\mathsf{f} = \mathsf{t}$ e $\mathsf{g} = \mathsf{Y}\mathsf{b}/\|\mathsf{b}\|$.

2.3.2 Ortogonalidade

As variáveis do resíduo X_{i+1} da matriz X_i são ortogonais ao fator t_i. De fato, usando a equação da linha 14 do algoritmo 2.1 temos

$$\begin{aligned} \mathsf{X}_{i+1}^\top \mathsf{t}_i &= (\mathsf{X}_i - \mathsf{t}_i \mathsf{p}_i^\top)^\top \mathsf{t}_i \\ &= \mathsf{X}_i^\top \mathsf{t}_i - \mathsf{p}_i \mathsf{t}_i^\top \mathsf{t}_i \\ &= \mathsf{X}_i^\top \mathsf{t}_i - \frac{\mathsf{X}_i^\top \mathsf{t}_i}{\mathsf{t}_i^\top \mathsf{t}_i} \mathsf{t}_i^\top \mathsf{t}_i \\ &= \mathsf{X}_i^\top \mathsf{t}_i - \mathsf{X}_i^\top \mathsf{t}_i \\ &= 0 \end{aligned} \tag{2-3}$$

Com isto, temos que o conjunto de fatores $\{\mathsf{t}_i\}$ $1 \leq i \leq m$, compõe uma base ortogonal em $\mathbb{R}^m$. Podemos provar que para todo $d > 0$, $i + d = j$ implica $\mathsf{t}_i^\top \mathsf{t}_j = 0$. Para isso, provamos por indução sobre d que $d > 0$ implica $\mathsf{t}_i^\top \mathsf{t}_{i+d} = 0$ e $\mathsf{X}_{i+d}^\top \mathsf{t}_i = 0$.
Usando como base $d = 1$, temos

$$\begin{aligned} \mathsf{t}_i^\top \mathsf{t}_{i+1} &= \mathsf{t}_i^\top \mathsf{X}_{i+1} \mathsf{w}_{i+1} \\ &= (\mathsf{X}_{i+1}^\top \mathsf{t}_i)^\top \mathsf{w}_{i+1} \\ &= 0 \end{aligned} \tag{2-4}$$

já que $\mathsf{X}_{i+1}^\top \mathsf{t}_i = 0$ por (2-3).

Assumindo que $\mathsf{X}_{i+d-1}^{\top}\mathsf{t}_i = 0$ e $\mathsf{t}_i^{\top}\mathsf{t}_{i+d-1} = 0$, temos que

$$\begin{aligned}\mathsf{t}_i^{\top}\mathsf{t}_{i+d} &= \mathsf{t}_i^{\top}\mathsf{X}_{i+d}\mathsf{w}_{i+d}\\ &= (\mathsf{X}_{i+d}^{\top}\mathsf{t}_i)^{\top}\mathsf{w}_{i+d}\\ &= (\mathsf{X}_{i+d-1}^{\top}\mathsf{t}_i - \mathsf{p}_{i+d-1}\mathsf{t}_{i+d-1}^{\top}\mathsf{t}_i)^{\top}\mathsf{w}_{i+d}\\ &= 0\end{aligned}$$

e

$$\begin{aligned}\mathsf{X}_{i+d}^{\top}\mathsf{t}_i &= \mathsf{X}_{i+d-1}^{\top}\mathsf{t}_i - \mathsf{p}_{i+d-1}\mathsf{t}_{i+d-1}^{\top}\mathsf{t}_i\\ &= 0\end{aligned}$$

Logo os fatores do conjunto $\{\mathsf{t}_1, \cdots, \mathsf{t}_m\}$ são mutuamente ortogonais. Existem versões do algoritmo [52] nas quais os fatores não são ortogonais apesar de o modelo gerado ser o mesmo. A ortogonalidade permite para o usuário da regressão uma melhor interpretação dos resultados, sendo mantida nas versões apresentadas nesta tese.

2.3.3
Propriedades específicas do algoritmo PLS1

A ortogonalidade dos fatores t_i permite que sejam feitas algumas simplificações no algoritmo PLS1. O cálculo dos resíduos Y_i pode ser dispensado:

$$\begin{aligned}\mathsf{Y}_i &= \mathsf{Y}_{i-1} - b_{i-1}\mathsf{t}_{i-1}\\ &= \mathsf{Y}_{i-2} - b_{i-2}\mathsf{t}_{i-2} - b_{i-1}\mathsf{t}_{i-1}\\ &= \mathsf{Y}_1 - b_1\mathsf{t}_1 - \cdots - b_i\mathsf{t}_i\end{aligned}$$

fazendo com que

$$\begin{aligned}\mathsf{w}_i &= \mathsf{X}_i^{\top}\mathsf{Y}_i\\ &= \mathsf{X}_i^{\top}(\mathsf{Y}_1 - b_1\mathsf{t}_1 - \cdots - b_i\mathsf{t}_i)\\ &= \mathsf{X}_i^{\top}\mathsf{Y}_1\\ &= \mathsf{X}_i^{\top}\mathsf{Y}\end{aligned}$$

já que os resíduos X_{i+1} são ortogonais aos fatores t_i.

Da mesma forma, podemos substituir a expressão do cálculo de p_i por

$$\begin{aligned}\mathsf{p}_i &= \mathsf{X}_i^\top \mathsf{t}_i / \mathsf{t}_i^\top \mathsf{t}_i \\ &= \mathsf{X}_1^\top \mathsf{t}_i / \mathsf{t}_i^\top \mathsf{t}_i \\ &= \mathsf{X}^\top \mathsf{t}_i / \mathsf{t}_i^\top \mathsf{t}_i\end{aligned}$$

2.4 Seleção de fatores

Um procedimento comum para determinar o número ótimo de fatores, ou componentes, a serem usados na predição, consiste numa estimativa da falta de precisão do modelo conseguida com o cálculo do PRESS [4]. Este tipo de procedimento, chamado de validação cruzada [1, 16, 31], usa um conjunto independente de r amostras X juntamente com variáveis conhecidas em Y. Para cada valor de h no algoritmo de predição PLS, obtemos a predição $\mathsf{Y}'(h)$ e calculamos

$$\text{PRESS} = \sum_{j}^{l} \sum_{i}^{r} (\mathsf{Y}_{ij} - \mathsf{Y}'_{ij}(h))^2 \tag{2-5}$$

onde l é o número de colunas, ou variáveis dependentes, de Y, e Y_{ij} e Y'_{ij} indicam o elemento da linha i coluna j da respectiva matriz. A figura 2.4 mostra um gráfico típico do PRESS obtido em função do número de componentes h. Apesar de o número de fatores usados poder ser tão grande quanto se deseje, limitado teoricamente pelo posto da matriz X, na prática o valor do PRESS começa a divergir ou atinge um mínimo antes que todos os fatores tenham sido incluídos. A inclusão de fatores adicionais a partir deste ponto resulta no aumento do valor do PRESS, o que significa uma pior predição. Isto se deve ao fato de o ruído presente nas variáveis de X estar sendo usado para modelar Y, ou seja, está ocorrendo *overfitting*. O fato de não incluir os fatores que descrevem ruído, permite aumentar o poder preditivo do modelo PLS.

Além do PRESS, outras estatísticas podem ser calculadas para análise do modelo gerado. Medindo-se o erro quadrático do resíduo X_i pode-se avaliar o quanto da matriz está sendo usado no modelo. Se for medido o erro quadrático das colunas do resíduo X_i, pode-se ter uma estimativa da contribuição de cada variável para o modelo. Por outro lado, se for calculado

[4] Prediction Residual Sum of Squares.

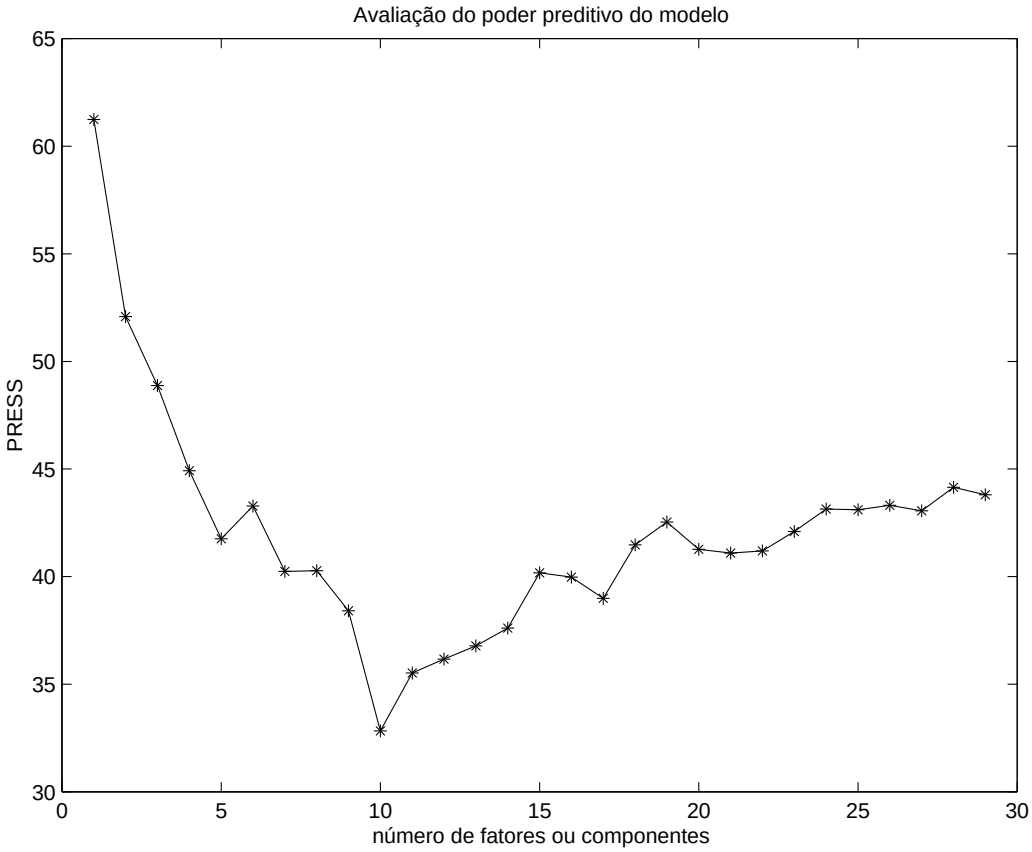

Figura 2.4: Gráfico do PRESS em função do número de fatores.

o erro das amostras no resíduo, podem ser detectados *outliers*, amostras discrepantes com a distribuição geral.

Finalmente, o processo de validação cruzada consiste em repetir o procedimento de calibração e ajuste várias vezes para um conjunto de dados, repartindo-o num subconjunto de treino e outro de teste. Isto fornece robustez aos parâmetros do modelo. Um tipo de validação cruzada muito usada é o *Leave-one-out*, onde o conjunto de teste contém apenas uma amostra e o ajuste é realizado o número de amostras vezes.