Introdução à Ciência de Dados para Engenheiros

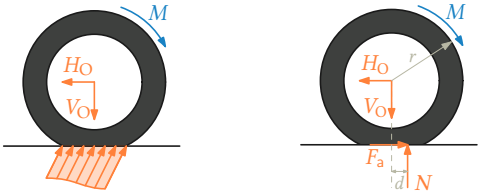
Ajuste de modelos matemáticos
Método dos mínimos quadrados

Renato Maia Matarazzo Orsino
reorsino@usp.br

Conteúdo

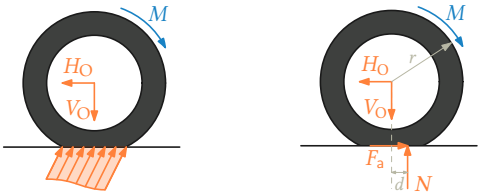
Ensaio de coastdown

Resistência ao rolamento



- Na Mecânica Geral, os modelos mais simplificados de rolamento, consideram que o contato de um cilindro rígido sobre uma pista plana indeformável se dá em uma única linha, paralela ao eixo do cilindro e de dimensões deprezíveis.
- Um modelo mais próximo da realidade deve considerar a área de contato decorrente da deformação tanto do cilíndro quanto da pista. Neste caso, o contato passa a ser modelado por meio de um carregamento distribuído por esta área, ou seja, por um campo de pressões.

Resistência ao rolamento

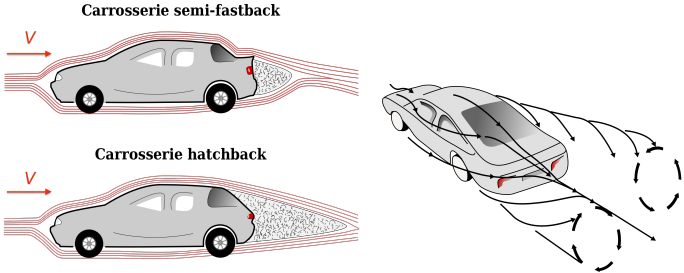


Idealmente, este campo de pressões poderia ser reduzido a uma única força com duas componentes:

- N (normal): ortogonal à superfície de contato.
- A (atrito): paralela à superfície de contato.

No entanto, devido à assimetria deste campo (ver figura), a linha de ação da componente normal não passa pelo "ponto" de contato, mas à direita dele, a uma distância k. Assim, V_O e N produzem um binário de intensidade Nk contrário ao momento M imposto pelo sistema de tração.

Força de arrasto aerodinâmico (*drag*)

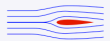


Fonte: Aerodinâmica automotiva (Wikipedia)

Componente da resultante do sistema de forças aerodinâmicas sobre o veículo, na direção oposta à velocidade relativa v do escoamento incidente.

$$F_{\mathrm{D}} \propto v^2$$

Força de arrasto aerodinâmico (*drag*)

Shape and flow	Form Drag	Skin friction
	0%	100%
	~10%	~90%
	~90%	~10%
	100%	0%

Fonte: Aerodinâmica automotiva

A força de arrasto média pode ser estimada pela expressão:

$$F_{\mathrm{D}} = \frac{1}{2}\,\rho\, v^2\, C_D\, A$$

com:

- ρ denotando a densidade do fluido;
- v sendo a velocidade relativa do escoamento incidente sobre o sólido;
- A sendo a área projetada (frontal);
- C_{D} ou C_{X} denotando um adimensional denominado coeficiente de arrasto que é função da forma do sólido e das propriedades de sua superfície.

Ensaio de *coastdown*

Objetivo – encontrar os coeficientes f_0 e f_2 da equação da força de resistência ao movimento do veículo:

$$R = f_0 + f_2 v^2$$

Procedimento (norma ABNT NBR10312)

- Acelerar o veículo a uma velocidade inicial igual ou superior a 105 $\mathrm{km/h}$ e iniciar a desaceleração livre do veículo. Quando o veículo atingir a velocidade igual ou superior a 100 $\mathrm{km/h}$, iniciar as medições de tempo em intervalos iguais de velocidade de 10 $\mathrm{km/h}$ no máximo, até que o veículo atinja uma velocidade igual ou inferior a 30 $\mathrm{km/h}$.
- Deve-se efetuar pelo menos cinco ensaios em cada sentido da pista de rolamento.
- Deve-se registrar os tempos de desaceleração a cada intervalo de velocidade fixado, e a temperatura ambiente e pressão barométrica em cada desaceleração.

Ensaio de *coastdown*

Aplicando a segunda lei de Newton, conclui-se que, neste ensaio, a aceleração pode ser modelada por meio da expressão:

$$a(v) = -c_0 - c_2 v^2, \quad \text{com } c_0 = \frac{f_0}{m} \text{ e } c_2 = \frac{f_2}{m}$$

Notando que $a = \dfrac{\mathrm{d}v}{\mathrm{d}t}$, $\dfrac{\mathrm{d}s}{\mathrm{d}t} = v \quad \Rightarrow \quad a\dfrac{\mathrm{d}s}{\mathrm{d}t} = v\dfrac{\mathrm{d}v}{\mathrm{d}t} \quad \Rightarrow \quad \mathrm{d}s = \dfrac{v}{a}\mathrm{d}v$, temos:

$$s - 0 = \int_{v_0}^{v} \frac{v}{a(v)}\,\mathrm{d}v = -\int_{v_0}^{v} \frac{v\,\mathrm{d}v}{c_0 + c_2 v^2} = -\frac{1}{2c_2}\ln|c_0 + c_2 v^2|\Big|_{v_0}^{v}$$

$$s(v) = \frac{1}{2c_2}\ln|c_0 + c_2 v_0^2| - \frac{1}{2c_2}\ln|c_0 + c_2 v^2| \;\Rightarrow\; \boxed{s(v) = \frac{1}{2c_2}\ln\left|\frac{c_0 + c_2 v_0^2}{c_0 + c_2 v^2}\right|}$$

Os valores de c_0 e c_2 podem então ser ajustados para que o gráfico da função $s(v)$ se aproxime da melhor forma possível a curva obtida pelos dados experimentais.

Método dos mínimos quadrados

Definição do problema

Dado uma série de dados $(\mathbf{x}^{(i)}, y^{(i)}), i = 1, 2, \ldots, n$, com $y^{(i)} \in \mathbb{R}$, ou seja, com $y^{(i)}$ escalar e $\mathbf{x}^{(i)} \in \mathbb{R}^d$, ou seja, $\mathbf{x}^{(i)}$ sendo um vetor ou escalar, obter o vetor de parâmetros $\boldsymbol{\theta} \in \mathbb{R}^m$ de um modelo matemático $h_{\boldsymbol{\theta}}$ tal que o erro ao estimar as saídas $y^{(i)}$ como $\hat{y}^{(i)} = h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$ seja mínimo sob uma métrica bem definida.

$$e^{(i)} = y^{(i)} - \hat{y}^{(i)} = y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$$

Hipótese: os erros $e^{(i)}$ são realizações de distribuições normais *independentes* de média nula e variância σ_i^2, i.e. $e^{(i)} \sim \mathcal{N}(0, \sigma_i)$:

$$p(e^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[e^{(i)}]^2}{2\sigma_i^2}\right) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})]^2}{2\sigma_i^2}\right) = p(y^{(i)}|\mathbf{x}^{(i)}; \boldsymbol{\theta})$$

Maximização da verossimilhança

Deseja-se maximizar a verossimilhança $L(\boldsymbol{\theta})$ dos parâmetros, ou seja, maximizar a probabilidade conjunta de que os valores de saída observados $y^{(i)}$ sejam os valores estimados pelo modelo $\hat{y}^{(i)} = h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$ acrescidos de erros distribuídos conforme hipótese acima:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y^{(i)}|\mathbf{x}^{(i)}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})]^2}{2\sigma_i^2} \right)$$

Teorema: Se g é uma função monotonicamente crescente, o valor de $\boldsymbol{\theta}$ que maximiza uma função $f(\boldsymbol{\theta})$ é o mesmo que maximiza $g(f(\boldsymbol{\theta}))$.

Dado o caráter crescente da função log (logaritmo natural), o valor de $\boldsymbol{\theta}$ que maximiza $\log L(\boldsymbol{\theta})$ é o mesmo que maximiza $L(\boldsymbol{\theta})$.

Maximização da verossimilhança

$$\begin{aligned}
\log L(\boldsymbol{\theta}) &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})]^2}{2\sigma_i^2}\right) \\
&= \sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})]^2}{2\sigma_i^2}\right)\right] \\
&= \sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi\sigma_i}}\right] - \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})]^2
\end{aligned}$$

Dado que o primeiro termo da última expressão acima é positivo e independe de $\boldsymbol{\theta}$, pode-se concluir que o máximo valor de $\boldsymbol{\theta}$ é obtido quando a forma quadrática $J(\boldsymbol{\theta})$ do segundo termo, também positiva, é mínima:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})]^2$$

Identificação da função objetivo

Supondo que todos os $e^{(i)}$ são realizações de distribuições normais de mesma variância, i.e. todos os $\sigma_i^2 = \sigma^2$, pode-se fatorar a expressão de $J(\boldsymbol{\theta}) = S(\boldsymbol{\theta})/\sigma^2$, e o valor de $\boldsymbol{\theta}$ que maximiza $L(\boldsymbol{\theta})$ pode ser obtido minimizando:

$$S(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})]^2$$

Esta é a versão original do problema de mínimos quadrados (Gauss, Legendre). Também se pode considerar que os erros de estimativa associados a diferentes pontos de dados venham de distribuições com variâncias distintas. Neste caso, temos uma matriz diagonal $R \in \mathbb{R}^{n\times n}$ de covariância de erros de medidas dada por:

$$\boldsymbol{R} = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}$$

Identificação da função objetivo

Neste caso, podemos reescrever a função objetivo $J(\boldsymbol{\theta})$ na forma:

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{y} - \hat{\mathbf{y}})^{*}\boldsymbol{R}^{-1}(\mathbf{y} - \hat{\mathbf{y}})$$

com $(\cdot)^{*}$ denotando a transposta de $(\cdot)$ e:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \quad \text{e} \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(n)} \end{bmatrix} = \begin{bmatrix} h_{\boldsymbol{\theta}}(\mathbf{x}^{(1)}) \\ \vdots \\ h_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}) \end{bmatrix}$$

Note que tal forma de expressar a função objetivo $J(\boldsymbol{\theta})$ também permite estender a formulação do problema para casos em que seja razoável supor que as distribuições dos erros não sejam independentes entre si. Neste último caso, no entanto, a matriz de covariância $\boldsymbol{R}$ deixaria de ser diagonal.

Regressão linear e não-linear

Em problemas de regressão não-linear, ou seja, em que a expressão de $h_{\boldsymbol{\theta}}$ dependa de forma não-linear de $\boldsymbol{\theta}$, pode não haver solução única para o problema de mínimo.

Nestes casos, é possível que seja conhecida uma estimativa inicial $\boldsymbol{\theta}_0$ para os parâmetros buscados (que tenha, por exemplo, ordem de grandeza próxima aos valores que se deseja encontrar). Neste caso, podemos considerar uma versão aumentada da função objetivo que penaliza soluções que se afastem demasiadamente de $\boldsymbol{\theta}_0$, adotando para tal, uma matriz de pesos definida positiva $\boldsymbol{P}_0 \in \mathbb{R}^{m \times m}$:

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^* \boldsymbol{P}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\mathbf{y} - \hat{\mathbf{y}})^* \boldsymbol{R}^{-1}(\mathbf{y} - \hat{\mathbf{y}})$$

Método do gradiente descendente

Método do gradiente descendente

1. Faça uma estimativa inicial para $\boldsymbol{\theta}$, obtenha o valor de $\nabla_{\boldsymbol{\theta}} J = \left[\dfrac{\partial J}{\partial \theta_j}\right]$ para cada um dos pontos de dados:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} \quad \text{e} \quad \nabla_{\boldsymbol{\theta}} J = \begin{bmatrix} \dfrac{\partial J}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial J}{\partial \theta_m} \end{bmatrix}$$

2. Em seguida, aplique correções ao valor inicial dando pequenos passos no sentido oposto aos dos valores de $\nabla_{\boldsymbol{\theta}} J$, adotando uma taxa de aprendizado α pequena:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \, \nabla_{\boldsymbol{\theta}} J$$

Método do gradiente descendente

Em particular, se tomarmos:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})]^2 \quad \Rightarrow \quad \nabla_{\boldsymbol{\theta}} J = -\sum_{i=1}^{n} \frac{1}{\sigma_i^2}[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})](\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}})\Big|_{\mathbf{x}=\mathbf{x}^{(i)}}$$

Em suma, seria possível repetir, até a convergência, o seguinte procedimento:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \sum_{i=1}^{n} \frac{1}{\sigma_i^2}[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})](\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}})\Big|_{(i)}$$

Método do gradiente descendente estocástico

Uma variação deste algoritmo que, em geral, tende a uma convergência mais rápida (porém menos precisa) é o método incremental (ou estocástico) de gradiente descendente que consiste em atualizar o valor de $\boldsymbol{\theta}$ para cada novo dado que chega. Neste caso, repete-se (até que as oscilações de valor dos elementos de $\boldsymbol{\theta}$ fique abaixo de uma dada margem):

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \frac{1}{\sigma_i^2}[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})](\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}})\Big|_{\mathbf{x}=\mathbf{x}^{(i)}}$$

Ou ainda, se usarmos $S(\boldsymbol{\theta})$ em vez de $J(\boldsymbol{\theta})$ como função objetivo:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \left[y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})\right](\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}})\Big|_{\mathbf{x}=\mathbf{x}^{(i)}}$$

Regressão Linear Multivariada (RLM)

Definição do problema

Um problema de regressão linear pode ser classificado como uma RLM se existir uma função $\boldsymbol{\phi}$ que converte cada vetor de atributos $\mathbf{x}^{(i)} \in \mathbb{R}^d$ em um vetor de características $\boldsymbol{\phi}(\mathbf{x}^{(i)}) \in \mathbb{R}^m$, tal que o modelo matemático em questão possa ser escrito como:

$$h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(i)})$$

Nestes casos, o algoritmo de gradiente descendente, baseado na função objetivo $S(\boldsymbol{\theta})$, se torna simplesmente:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \sum_{i=1}^{n} \left[y^{(i)} - \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(i)}) \right] \boldsymbol{\phi}(\mathbf{x}^{(i)})$$

Já o algoritmo incremental de gradiente descendente se torna:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \left[y^{(i)} - \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(i)}) \right] \boldsymbol{\phi}(\mathbf{x}^{(i)})$$

Exemplo

Se $\mathbf{x} \in \mathbb{R}^2$ e o modelo for um polinômio de segundo grau da forma:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_1^2 + \theta_5 x_1 x_2 + \theta_6 x_2^2$$

temos $\boldsymbol{\theta} \in \mathbb{R}^6$ e:

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix}$$

Hipótese

O vetor de parâmetros pode ser descrito como uma combinação linear dos $\boldsymbol{\phi}(\mathbf{x}^{(i)})$:

$$\boldsymbol{\theta} = \sum_{i=1}^{n} \beta_i \boldsymbol{\phi}(\mathbf{x}^{(i)})$$

Dessa forma, o algoritmo de gradiente descendente pode ser reescrito como:

$$\begin{aligned}
\boldsymbol{\theta} &\leftarrow \sum_{i=1}^{n} \beta_i \boldsymbol{\phi}(\mathbf{x}^{(i)}) + \alpha \sum_{i=1}^{n} \left[y^{(i)} - \sum_{j=1}^{n} \beta_j \boldsymbol{\phi}(\mathbf{x}^{(j)})^* \boldsymbol{\phi}(\mathbf{x}^{(i)}) \right] \boldsymbol{\phi}(\mathbf{x}^{(i)}) \\
&= \sum_{i=1}^{n} \left[\beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \boldsymbol{\phi}(\mathbf{x}^{(i)})^* \boldsymbol{\phi}(\mathbf{x}^{(j)}) \beta_j \right) \right] \boldsymbol{\phi}(\mathbf{x}^{(i)})
\end{aligned}$$

Matriz núcleo e função núcleo

Em outras palavras, o algoritmo de gradiente descendente equivale a:

$$\beta_i \leftarrow \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \boldsymbol{\phi}(\mathbf{x}^{(i)})^* \boldsymbol{\phi}(\mathbf{x}^{(j)}) \beta_j \right)$$

ou, ainda, em formato matricial:

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \alpha(\mathbf{y} - \boldsymbol{K}\boldsymbol{\beta})$$

com $\boldsymbol{K} = [K_{ij}]$ sendo a matriz núcleo cujas entradas são definidas como:

$$K_{ij} = \boldsymbol{\phi}(\mathbf{x}^{(i)})^* \boldsymbol{\phi}(\mathbf{x}^{(j)}) = \mathrm{K}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$$

Pode-se definir também a função núcleo $\mathrm{K}(\mathbf{u}, \mathbf{v})$ cujo valor corresponde ao produto interno de $\boldsymbol{\phi}(\mathbf{u})$ por $\boldsymbol{\phi}(\mathbf{v})$.

$$\mathrm{K}(\mathbf{u}, \mathbf{v}) = \boldsymbol{\phi}(\mathbf{u})^* \boldsymbol{\phi}(\mathbf{v})$$

Matriz núcleo e função núcleo

Deseja-se, no entanto, que seja possível calcular $\mathrm{K}(\mathbf{u}, \mathbf{v})$ sem a necessidade de se calcular explicitamente $\boldsymbol{\phi}(\mathbf{u})$ ou $\boldsymbol{\phi}(\mathbf{v})$, uma vez que, tipicamente, $m \gg d$. Para diversos casos isto é possível:

- Núcleo polinomial de grau k: suponha que $\boldsymbol{\phi}(\mathbf{u})$ seja constituído por todos os monômios da forma $u_1^{i_1} u_2^{i_2} \ldots u_d^{i_d}$ de grau igual ou inferior a k, ou seja, com $0 \leq i_1 + i_2 + \ldots + i_d \leq k$. Neste caso, $m = \binom{d+k}{k} = \frac{(d+k)!}{d!k!}$. No entanto, $\mathrm{K}(\mathbf{u}, \mathbf{v}) = \boldsymbol{\phi}(\mathbf{u})^* \boldsymbol{\phi}(\mathbf{v})$ pode ser calculado simplesmente como:

$$\mathrm{K}(\mathbf{u}, \mathbf{v}) = (\mathbf{u}^* \mathbf{v} + c)^k$$

- Núcleo Gaussiano: considera um espaço de características de dimensão infinita, medindo a similaridade entre u e v por meio da expressão:

$$\mathrm{K}(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{u} - \mathbf{v}\|^2}{2\sigma^2}\right)$$

Teorema de Mercer

Para que uma dada função $\mathrm{K}(\mathbf{u}, \mathbf{v})$ defina de forma consistente um núcleo (ou seja, represente um produto interno entre dois vetores de características), é necessário e suficiente que para qualquer conjunto de dados $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}, (n < \infty)$, a matriz núcleo correspondente $\boldsymbol{K} = [K_{ij}]$ seja simétrica e definida positiva.

Uma vez ajustados os valores dos β_i, pode-se expressar o modelo $h_{\boldsymbol{\theta}}(x) = \boldsymbol{\theta}^* \boldsymbol{\phi}(x)$ em termos da função núcleo escolhida:

$$h_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{n} \beta_i \mathrm{K}(\mathbf{x}^{(i)}, x)$$

Pseudo-inversa de Moore-Penrose

Definição: dado $\boldsymbol{A} \in \mathbb{R}^{r \times s}$ a pseudo-inversa $\boldsymbol{A}^{\#} \in \mathbb{R}^{s \times r}$ é a única matriz que satisfaz às quatro seguintes propriedades:

1. $\boldsymbol{A}\boldsymbol{A}^{\#}\boldsymbol{A} = \boldsymbol{A}$
2. $\boldsymbol{A}^{\#}\boldsymbol{A}\boldsymbol{A}^{\#} = \boldsymbol{A}^{\#}$
3. $(\boldsymbol{A}\boldsymbol{A}^{\#})^{*} = \boldsymbol{A}\boldsymbol{A}^{\#}$
4. $(\boldsymbol{A}^{\#}\boldsymbol{A})^{*} = \boldsymbol{A}^{\#}\boldsymbol{A}$

Uso da pseudo-inversa de Moore-Penrose

Caso I: se não houver solução $\boldsymbol{\xi}$ para o sistema linear $\boldsymbol{A}\boldsymbol{\xi} = \mathbf{y}$, $\boldsymbol{\xi} \leftarrow \boldsymbol{A}^{\#}\mathbf{y}$ é a solução que minimiza o erro quadrático $\|\boldsymbol{A}\boldsymbol{\xi} - \mathbf{y}\|^2$. Neste caso, $\boldsymbol{A}^{\#} = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*$.

Este é o caso de um problema de mínimos quadrados linear em que se adota $\sigma_i^2 = \sigma^2, \forall i$:

$$J_n(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \boldsymbol{\theta}^*\boldsymbol{\phi}(\mathbf{x}^{(i)})\right)^2$$

O valor de $\boldsymbol{\theta} = \boldsymbol{\theta}^{(n)}$ que minimiza o valor de J_n é dado por:

$$\left.\frac{\partial J_n}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(n)}} = 0 \quad \Rightarrow \quad -\frac{1}{\sigma^2}\sum_{i=1}^{n}\boldsymbol{\phi}(\mathbf{x}^{(i)})\left(y^{(i)} - \boldsymbol{\phi}(\mathbf{x}^{(i)})^*\boldsymbol{\theta}^{(n)}\right) = 0$$

$$\Rightarrow \quad -\frac{1}{\sigma^2}\boldsymbol{\Phi}\left(\mathbf{y} - \boldsymbol{\Phi}^*\boldsymbol{\theta}^{(n)}\right) = 0 \quad \text{com} \quad \boldsymbol{\Phi} = \begin{bmatrix}\boldsymbol{\phi}(\mathbf{x}^{(1)}) & \dots & \boldsymbol{\phi}(\mathbf{x}^{(n)})\end{bmatrix} \in \mathbb{R}^{m\times n}$$

Uso da pseudo-inversa de Moore-Penrose

De fato, identificando: $\boldsymbol{A} = \boldsymbol{\Phi}^*$ e $\boldsymbol{\xi} = \boldsymbol{\theta}^{(n)}$:

$$\boldsymbol{\theta}^{(n)} = \left[\sum_{i=1}^{n} \boldsymbol{\phi}(\mathbf{x}^{(i)})\boldsymbol{\phi}(\mathbf{x}^{(i)})^*\right]^{-1} \left[\sum_{i=1}^{n} \boldsymbol{\phi}(\mathbf{x}^{(i)})y^{(i)}\right] = (\boldsymbol{\Phi}\boldsymbol{\Phi}^*)^{-1}\boldsymbol{\Phi}\mathbf{y} = \boldsymbol{\Phi}^{*\#}\mathbf{y}$$

Caso II: se houver infinitas soluções $\boldsymbol{\xi}$ para o sistema linear $\boldsymbol{A}\boldsymbol{\xi} = \mathbf{y}$, então $\boldsymbol{\xi} \leftarrow \boldsymbol{A}^{\#}\mathbf{y}$ fornece a solução que minimiza $\|\boldsymbol{\xi}\|^2$.

Este é o caso, por exemplo, de um problema de mínimos quadrados linear em que se adota $\sigma_i^2 = \sigma^2, \forall i$, $\boldsymbol{\theta}^{(0)} = 0$ e é necessário considerar $\eta > 0$ para que não haja singularidade na inversão de matrizes, ou seja:

$$J_n(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \boldsymbol{\theta}^*\boldsymbol{\phi}(\mathbf{x}^{(i)})\right)^2 + \frac{1}{2}\eta\|\boldsymbol{\theta}\|^2 = \frac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\Phi}^*\boldsymbol{\theta}\|^2 + \frac{1}{2}\eta\|\boldsymbol{\theta}\|^2$$

Uso da pseudo-inversa de Moore-Penrose

Neste caso, $\boldsymbol{\theta}^{(n)} = \boldsymbol{\Phi}^{*\#}\mathbf{y}$ corresponde ao valor de $\boldsymbol{\theta}$ que não apenas torna $\mathbf{y} - \boldsymbol{\Phi}^*\boldsymbol{\theta} = 0$ como também torna $\|\boldsymbol{\theta}\|$ mínimo, de forma a garantir que $J_n(\boldsymbol{\theta})$ também seja mínimo.

Também recai no caso II formulações de regressão linear multivariada via uso de matrizes núcleo. Nestes casos, $\boldsymbol{\beta} \leftarrow \boldsymbol{K}^{\#}\mathbf{y}$ é a solução procurada.

Caso III: havendo solução única $\boldsymbol{\xi}$ para o sistema linear $\boldsymbol{A}\boldsymbol{\xi} = \mathbf{y}$, $\boldsymbol{\xi} \leftarrow \boldsymbol{A}^{\#}\mathbf{y} = \boldsymbol{A}^{-1}\mathbf{y}$ é o valor desta solução. Neste caso, $\boldsymbol{A}^{\#} = \boldsymbol{A}^{-1}$.

Quando o usuário baseia seu algoritmo de solução numérica no uso de matrizes pseudo-inversas não é necessário que ele faça qualquer distinção *a priori* entre os problemas que recaiam nos casos I, II ou III. Uma solução de mínimos quadrados de um problema de regressão linear generalizada em que se adota $\sigma_i^2 = \sigma^2, \forall i$ e $\boldsymbol{\theta}^{(0)} = 0$ será sempre dada por:

$$\boldsymbol{\theta}^{(n)} = \boldsymbol{\Phi}^{*\#}\mathbf{y}$$

Método Recursivo de Mínimos Quadrados (RLS)

Definição do problema

Dado uma série de dados $(\mathbf{x}^{(i)}, y^{(i)}), i = 1, 2, \ldots, n$, com $y^{(i)} \in \mathbb{R}$ (escalar) e uma função $\boldsymbol{\phi}$ que converte cada vetor de atributos $\mathbf{x}^{(i)} \in \mathbb{R}^d$ em um vetor de características $\boldsymbol{\phi}(\mathbf{x}^{(i)}) \in \mathbb{R}^m$, obter o vetor de parâmetros $\boldsymbol{\theta} \in \mathbb{R}^m$ do modelo matemático $h_{\boldsymbol{\theta}}(x) = \boldsymbol{\theta}^* \boldsymbol{\phi}(x)$ tal que ao estimar as saídas $y^{(i)}$ como $\hat{y}^{(i)} = \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(i)})$ seja mínimo o valor da função objetivo:

$$J_n(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^* \boldsymbol{P}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\mathbf{y} - \hat{\mathbf{y}})^* \boldsymbol{R}^{-1}(\mathbf{y} - \hat{\mathbf{y}})$$

com:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \quad \text{e} \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(n)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(1)}) \\ \vdots \\ \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(n)}) \end{bmatrix}$$

A função objetivo J_n penaliza não somente os erros de estimativa $(\mathbf{y} - \hat{\mathbf{y}})$ como também vetores de parâmetros que se afastem demasiadamente de $\boldsymbol{\theta}_0$, adotando para tal, uma matriz de pesos definida positiva $\boldsymbol{P}_0 \in \mathbb{R}^{m \times m}$.

Definição do problema

Admitamos que as variâncias dos erros de estimativa dos diferentes pontos de dados possam ser admitidas independentes, de tal forma que a matriz de covariânca de erros $\boldsymbol{R} \in \mathbb{R}^{n \times n}$ seja diagonal:

$$\boldsymbol{R} = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}$$

Dessa forma, a expressão para a função objetivo $J_n(\theta)$ pode ser reescrita na forma:

$$J_n(\theta) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^* \boldsymbol{P}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}\left(y^{(i)} - \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(i)})\right)^2$$

Dedução do algoritmo

Note que, para todo $k > 0$:

$$\frac{\partial J_k}{\partial \boldsymbol{\theta}} = \boldsymbol{P}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \sum_{i=1}^{k} \frac{1}{\sigma_i^2}\left(y^{(i)} - \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(i)})\right)\boldsymbol{\phi}(\mathbf{x}^{(i)}) \tag{1}$$

$$\frac{\partial J_{k+1}}{\partial \boldsymbol{\theta}} = \frac{\partial J_k}{\partial \boldsymbol{\theta}} - \frac{1}{\sigma_{k+1}^2}\left(y^{(k+1)} - \boldsymbol{\theta}^* \boldsymbol{\phi}(\mathbf{x}^{(k+1)})\right)\boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \tag{2}$$

O valor de $\boldsymbol{\theta} = \boldsymbol{\theta}_k$ que minimiza o valor de J_k é dado por:

$$\left.\frac{\partial J_k}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k} = 0 \quad \Rightarrow \quad \boldsymbol{P}_0^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_0) - \sum_{i=1}^{k} \boldsymbol{\phi}(\mathbf{x}^{(i)}) \frac{1}{\sigma_i^2}\left(y^{(i)} - \boldsymbol{\phi}(\mathbf{x}^{(i)})^* \boldsymbol{\theta}_k\right) = 0 \tag{3}$$

$$\boldsymbol{\theta}_k = \boldsymbol{P}_k \left[\boldsymbol{P}_0^{-1}\boldsymbol{\theta}_0 + \sum_{i=1}^{k} \frac{1}{\sigma_i^2} y^{(i)} \boldsymbol{\phi}(\mathbf{x}^{(i)})\right] \tag{4}$$

Dedução do algoritmo

$\boldsymbol{P}_k$ é definida como:

$$\boldsymbol{P}_k^{-1} = \boldsymbol{P}_0^{-1} + \sum_{i=1}^{k} \frac{1}{\sigma_i^2} \boldsymbol{\phi}(\mathbf{x}^{(i)}) \boldsymbol{\phi}(\mathbf{x}^{(i)})^* \tag{5}$$

Definam-se também:

$$S_{k+1} = \boldsymbol{\phi}(\mathbf{x}^{(k+1)})^* \boldsymbol{P}_k \boldsymbol{\phi}(\mathbf{x}^{(k+1)}) + \sigma_{k+1}^2 \tag{6}$$

$$\boldsymbol{K}_{k+1} = \frac{1}{S_{k+1}} \boldsymbol{P}_k \boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \tag{7}$$

Notando que:

$$\boldsymbol{P}_{k+1}^{-1} = \boldsymbol{P}_k^{-1} + \frac{1}{\sigma_{k+1}^2} \boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \boldsymbol{\phi}(\mathbf{x}^{(k+1)})^* \tag{8}$$

Dedução do algoritmo

Utilizando o lema de inversão de matrizes:

$$\boldsymbol{P}_{k+1} = \left[\boldsymbol{I} - \underbrace{\frac{1}{S_{k+1}}\boldsymbol{P}_k\boldsymbol{\phi}(\mathbf{x}^{(k+1)})}_{\boldsymbol{K}_{k+1}}\boldsymbol{\phi}(\mathbf{x}^{(k+1)})^*\right]\boldsymbol{P}_k = \left(\boldsymbol{I} - \boldsymbol{K}_{k+1}\boldsymbol{\phi}(\mathbf{x}^{(k+1)})^*\right)\boldsymbol{P}_k \tag{9}$$

Notando ainda que:

$$\boldsymbol{K}_{k+1}S_{k+1} = \boldsymbol{P}_k\boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \tag{10}$$

$$\boldsymbol{K}_{k+1}\sigma^2_{k+1} = \boldsymbol{P}_k\boldsymbol{\phi}(\mathbf{x}^{(k+1)}) - \boldsymbol{K}_{k+1}\boldsymbol{\phi}(\mathbf{x}^{(k+1)})^*\boldsymbol{P}_k\boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \tag{11}$$

$$\boldsymbol{K}_{k+1}\sigma^2_{k+1} = \left(\boldsymbol{I} - \boldsymbol{K}_{k+1}\boldsymbol{\phi}(\mathbf{x}^{(k+1)})^*\right)\boldsymbol{P}_k\boldsymbol{\phi}(\mathbf{x}^{(k+1)}) = \boldsymbol{P}_{k+1}\boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \tag{12}$$

$$\boldsymbol{K}_{k+1} = \frac{1}{\sigma^2_{k+1}}\boldsymbol{P}_{k+1}\boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \tag{13}$$

Dedução do algoritmo

Assim, o valor de $\boldsymbol{\theta} = \boldsymbol{\theta}_{k+1}$ que minimiza J_{k+1} é dado por:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{P}_{k+1}\left[\boldsymbol{P}_0^{-1}\boldsymbol{\theta}_0 + \sum_{i=1}^{k+1}\frac{1}{\sigma_i^2}y^{(i)}\boldsymbol{\phi}(\mathbf{x}^{(i)})\right] \tag{14}$$

$$= (\boldsymbol{I} - \boldsymbol{K}_{k+1}\boldsymbol{\phi}(\mathbf{x}^{(k+1)})^*)\boldsymbol{\theta}_k + \underbrace{\frac{1}{\sigma_{k+1}^2}\boldsymbol{P}_{k+1}\boldsymbol{\phi}(\mathbf{x}^{(k+1)})}_{\boldsymbol{K}_{k+1}}y^{(k+1)} \tag{15}$$

Finalmente:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \boldsymbol{K}_{k+1}\left(y^{(k+1)} - \boldsymbol{\phi}(\mathbf{x}^{(k+1)})^*\boldsymbol{\theta}_k\right)$$

Método recursivo de mínimos quadrados (RLS)

Em outras palavras, o algoritmo recursivo para a solução de um problema de regressão linear multivariada consiste em iterar, para $k = 0, 1, \ldots, n-1$, iniciando com um valor arbitrário para $\boldsymbol{\theta}_0 \in \mathbb{R}^m$ e com uma matriz definida positiva $\boldsymbol{P}_0 \in \mathbb{R}^{m \times m}$:

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \left(\frac{y^{(k+1)} - \boldsymbol{\phi}(\mathbf{x}^{(k+1)})^* \boldsymbol{\theta}_k}{\boldsymbol{\phi}(\mathbf{x}^{(k+1)})^* \boldsymbol{P}_k \boldsymbol{\phi}(\mathbf{x}^{(k+1)}) + \sigma_{k+1}^2} \right) \boldsymbol{P}_k \boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \tag{16}$$

$$\boldsymbol{P}_{k+1} \leftarrow \boldsymbol{P}_k - \frac{1}{\boldsymbol{\phi}(\mathbf{x}^{(k+1)})^* \boldsymbol{P}_k \boldsymbol{\phi}(\mathbf{x}^{(k+1)}) + \sigma_{k+1}^2} \boldsymbol{P}_k \boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \boldsymbol{\phi}(\mathbf{x}^{(k+1)})^* \boldsymbol{P}_k \tag{17}$$

Método recursivo de mínimos quadrados (RLS)

Alternativamente, a fim de reduzir o número de operações, pode-se reescrever o algoritmo anterior na forma ($k = 0, 1, \ldots, n-1$):

$$\mathbf{z}_{k+1} \leftarrow \boldsymbol{P}_k \boldsymbol{\phi}(\mathbf{x}^{(k+1)}) \tag{18}$$

$$\gamma_{k+1} \leftarrow \frac{1}{\boldsymbol{\phi}(\mathbf{x}^{(k+1)})^* \mathbf{z}_{k+1} + \sigma^2_{k+1}} \tag{19}$$

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \gamma_{k+1}\left(y^{(k+1)} - \boldsymbol{\phi}(\mathbf{x}^{(k+1)})^* \boldsymbol{\theta}_k\right)\mathbf{z}_{k+1} \tag{20}$$

$$\boldsymbol{P}_{k+1} \leftarrow \boldsymbol{P}_k - \gamma_{k+1}\mathbf{z}_{k+1}\mathbf{z}^*_{k+1} \tag{21}$$

RLS para problemas não-lineares

Dado uma série de dados $(\mathbf{x}^{(i)}, y^{(i)}), i = 1, 2, \ldots, n$, com $y^{(i)} \in \mathbb{R}$ (escalar) e $\mathbf{x}^{(i)} \in \mathbb{R}^d$ (vetor de atributos), obter o vetor de parâmetros $\boldsymbol{\theta} \in \mathbb{R}^m$ do modelo matemático $h(\mathbf{x}; \boldsymbol{\theta})$ tal que ao estimar as saídas $y^{(i)}$ como $\hat{y}^{(i)} = h(\mathbf{x}^{(i)}; \boldsymbol{\theta})$ seja mínimo o valor da função objetivo:

$$J_n(\boldsymbol{\theta}) = \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\right)^* \boldsymbol{P}_0^{-1}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\right) + \frac{1}{2}(\mathbf{y} - \hat{\mathbf{y}})^* \boldsymbol{R}^{-1}(\mathbf{y} - \hat{\mathbf{y}}) \tag{22}$$

$$= \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\right)^* \boldsymbol{P}_0^{-1}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\right) + \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 \tag{23}$$

com:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \quad \text{e} \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(n)} \end{bmatrix} = \begin{bmatrix} h(\mathbf{x}^{(1)}; \theta) \\ \vdots \\ h(\mathbf{x}^{(n)}; \theta) \end{bmatrix}$$

Algoritmo recursivo para estimativa de parâmetros

$$e_k \leftarrow y^{(k)} - h(\mathbf{x}^{(k)}; \boldsymbol{\theta}^{(k-1)}) \tag{24}$$

$$\mathbf{g}^{(k)} \leftarrow \nabla_{\boldsymbol{\theta}} h(\mathbf{x} = \mathbf{x}^{(k)}; \boldsymbol{\theta} = \boldsymbol{\theta}^{(k-1)}) \tag{25}$$

$$s_k \leftarrow \sigma_k^2 + \mathbf{g}^{(k)^*} \boldsymbol{P}_{k-1} \mathbf{g}^{(k)} \tag{26}$$

$$\boldsymbol{\theta}^{(k)} \leftarrow \boldsymbol{\theta}^{(k-1)} + \frac{e_k}{s_k} \boldsymbol{P}_{k-1} \mathbf{g}^{(k)} \tag{27}$$

$$\boldsymbol{P}_k \leftarrow \boldsymbol{P}_{k-1} - \frac{1}{s_k} \boldsymbol{P}_{k-1} \mathbf{g}^{(k)} \mathbf{g}^{(k)^*} \boldsymbol{P}_{k-1} \tag{28}$$

- $\mathbf{g}^{(k)} \in \mathbb{R}^m$: valor do gradiente, calculado em $\mathbf{x}^{(k)}$, do modelo com respeito aos seus parâmetros, considerando ajuste recursivo até o $(k-1)$-ésimo dado.
- $s_k \in \mathbb{R}$: soma da variância σ_k^2 admitida *a priori* com a incerteza que o novo dado de entrada $\mathbf{x}^{(k)}$ traz ao modelo ajustado até o $(k-1)$-ésimo dado.
- $\boldsymbol{P}_k \in \mathbb{R}^{m \times m}$: matriz de covariância do erro de estimação dos parâmetros, tendo como base os k primeiros dados da série.

Algoritmo recursivo para estimativa de parâmetros

1. Para um modelo linear generalizado, $h(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\theta}^{*}\boldsymbol{\phi}(\mathbf{x})$ e, dessa forma, temos $\mathbf{g}^{(k)} = \boldsymbol{\phi}(\mathbf{x}^{(k)})$.
2. O algoritmo de gradiente descendente estocástico pode ser entendido como uma aproximação do método recursivo em que não se faz a atualização recursiva da matriz $\boldsymbol{P}_k$ ou da incerteza s_k. Em vez disso, adota-se uma taxa de aprendizado fixa e pequena α ($0 < \alpha \ll 1$) e substitui-se $\boldsymbol{P}_k$ por $\alpha\boldsymbol{I}$ e s_k apenas por σ_k^2:

$$e_k \leftarrow y^{(k)} - h(\mathbf{x}^{(k)};\boldsymbol{\theta}^{(k-1)}) \tag{29}$$

$$\mathbf{g}^{(k)} \leftarrow \nabla_{\boldsymbol{\theta}} h(\mathbf{x}=\mathbf{x}^{(k)};\boldsymbol{\theta}=\boldsymbol{\theta}^{(k-1)}) \tag{30}$$

$$\boldsymbol{\theta}^{(k)} \leftarrow \boldsymbol{\theta}^{(k-1)} + \alpha\frac{e_k}{\sigma_k^2}\mathbf{g}^{(k)} \tag{31}$$

Algoritmo recursivo para estimativa de parâmetros

Na ausência de qualquer informação para uma estimativa inicial de $\boldsymbol{\theta}^{(0)}$, é típico arbitrar o valor inicial das componentes do vetor de parâmetros e adotar $\boldsymbol{P}_0 = \dfrac{1}{\eta}\boldsymbol{I}$ com $\eta > 0$ (quanto menor η, menor a confiança na estimativa inicial). Assim:

$$\boldsymbol{P}_k = \left[\eta\boldsymbol{I} + \sum_{i=1}^{k} \frac{1}{\sigma_i^2}\mathbf{g}^{(i)}\mathbf{g}^{(i)*}\right]^{-1} \tag{32}$$

Neste caso, o valor de η atua como um parâmetro de regularização, garantindo a existência da inversa e, portanto, tornando possível a definição da matriz $\boldsymbol{P}_k$ ainda que $\displaystyle\sum_{i=1}^{k} \frac{1}{\sigma_i^2}\mathbf{g}^{(i)}\mathbf{g}^{(i)*}$ não tenha posto completo.