INVERSÃO LINEAR:

ESTIMADORES VIA NORMA L1

Neste tópico iremos ver a definição da norma L1 e seu significado físico. Em seguida iremos falar sobre dois estimadores. O primeiro é o estimador de mínimo dos resíduos via norma L1 e segundo é o estimador da variação total (total variation –TV).

Definição da norma L_n

Por definição, a norma ℓ_n de um vetor genérico M-dimensional $\overline{\mathbf{v}}$ é dada por:

$$\|\overline{\mathbf{v}}\|_n = \left[\sum_{i=1}^{M} |v_i|^n \right]^{1/n} \tag{1}$$

Portanto a norma ℓ_1 de um vetor genérico M-dimensional $\overline{\mathbf{v}}$ é dada por:

$$\|\overline{\mathbf{v}}\|_1 = \sum_{i=1}^{M} |v_i| \tag{2}$$

Estimador via a minimização do vetor dos resíduos segundo a norma 1

Seja um vetor N-dimensional $\boldsymbol{\varepsilon}$ que representa a diferença entre os dados geofísicos observados $\overline{\mathbf{y}}^o$ e estimados (ou preditos) $\overline{\mathbf{y}}^C$ tal que

$$\boldsymbol{\varepsilon} = \overline{\mathbf{y}}^o - \overline{\mathbf{y}}^c$$

No tópico 6 deste curso, deduzimos o estimador de mínimos quadrados (MQ) que consiste em minimizar a soma dos quadrados dos resíduos, i.e.,

$$\min_{\bar{\mathbf{p}}\in F} \{Q\} \equiv \min_{\bar{\mathbf{p}}\in F} \{\|\boldsymbol{\varepsilon}\|_2^2\} \equiv \min_{\bar{\mathbf{p}}\in F} \left\|\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}\,\bar{\mathbf{p}}\right\|_2^2 \equiv \min_{\bar{\mathbf{p}}\in F} \left(\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}\,\bar{\mathbf{p}}\right)^{\mathrm{T}} \left(\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}\,\bar{\mathbf{p}}\right)$$

No problema acima minimizamos o vetor dos resíduos $\boldsymbol{\varepsilon}$ usando o quadrado da norma L2. Isto equivale a minimizarmos a norma L2. Vimos no tópico 6 que a minimização desta função Q resulta no estimador dos mínimos quadrados (MQ sobredeterminado). Aqui chamaremos este estimador de

$$\hat{\bar{\mathbf{p}}}_{L2} = \left(\bar{\bar{\mathbf{A}}}^{\mathrm{T}}\bar{\bar{\mathbf{A}}}\right)^{-1} \bar{\bar{\mathbf{A}}}^{\mathbf{T}}\,\bar{\mathbf{y}}^{\mathbf{o}} \quad \text{(estimador MQ via norma L2) (3)}$$

Neste tópico definimos a função Q_1 como a norma ℓ_1 dos resíduos:

$$Q_1 = \|\boldsymbol{\varepsilon}\|_1 = \sum_{i=1}^{N} |\varepsilon_i|. \tag{4}$$

Então iremos minimizar a função Q_1, ou seja, minimizaremos o vetor dos resíduos $\boldsymbol{\varepsilon}$ (N x1) usando a norma ℓ_1:

$$\min_{\bar{\mathbf{p}}} \{Q_1\} = \min_{\bar{\mathbf{p}}} \|\boldsymbol{\varepsilon}\|_1 = \min_{\bar{\mathbf{p}}} \sum_{i=1}^{N} |\varepsilon_i|. \tag{5}$$

Como $\bar{\boldsymbol{\varepsilon}} = \bar{\mathbf{y}}^o - \bar{\bar{\mathbf{A}}}\bar{\mathbf{p}}$ em um problema linear, então a minimização do vetor dos resíduos $\boldsymbol{\varepsilon}$ (N x1) usando a norma ℓ_1, pode ser escrita como

$$\min_{\bar{\mathbf{p}}} \{Q_1\} = \min_{\bar{\mathbf{p}}} \|\boldsymbol{\varepsilon}\|_1 = \min_{\bar{\mathbf{p}}} \sum_{i=1}^{N} |\varepsilon_i| = \min_{\bar{\mathbf{p}}} \left|y^o{}_i - y^c{}_i\right| \tag{6}$$

em que $y^c{}_i$ é a i-ésimo elemento do vetor de dados ajustados (ou calculados ou preditos), i.e, $\bar{\mathbf{y}}^C = \bar{\bar{\mathbf{A}}}\bar{\mathbf{p}}$.

A condição de mínimo é que a derivada da função Q_1 (equação 4) em relação ao vetor de parâmetros $\bar{\mathbf{p}}$ seja zero. Note que esta função não é diferenciável se um dos elementos de $\varepsilon_1, \ldots, \varepsilon_N$ é zero. Vamos ignorar este fato e computar a derivada em relação ao k-ésimo parâmetro p_k , nos pontos em que os elementos de $\boldsymbol{\varepsilon}$ não são zero:

$$\frac{\partial Q_1}{\partial p_k} = \sum_{i=1}^{N} \frac{\partial|\varepsilon_i|}{\partial p_k} = \sum_{i=1}^{N} a_{ik}\,\mathrm{sgn}(\varepsilon_i) \tag{7}$$

em que a_{ik} é o ik-ésimo elemento da matriz de sensibilidade $\overline{\overline{\mathbf{A}}}$ e $\mathrm{sgn}(\varepsilon_i)$ é o sinal do i-ésimo resíduo. Note que a derivada da função Q_1 (equação 4) em relação ao vetor de parâmetros $\overline{\mathbf{p}}$ depende do resíduo $\boldsymbol{\varepsilon}$, como $\boldsymbol{\varepsilon} = \overline{\mathbf{y}}^o - \overline{\mathbf{y}}^c = \overline{\mathbf{y}}^o - \overline{\overline{\mathbf{A}}}\overline{\mathbf{p}}$ então estamos diante de um problema rigorosamente não-linear. Isto porque a derivada da função depende dos parâmetros desconhecidos $\overline{\mathbf{p}}$ que queremos estimar.

Apesar do problema de minimização dos resíduos segundo a norma 1 ser um rigorosamente não-linear, uma das alternativas é resolver através de um método iterativo dos mínimos quadrados reponderados (IRLS – iteratively reweighted least sqaures) .

Para tanto vamos reescrever o sinal do i-ésimo resíduo como

$$\mathrm{sgn}(\varepsilon_i) = \frac{\varepsilon_i}{|\varepsilon_i|},$$

Então podemos reescrever a equação 7 como

$$\frac{\partial Q_1}{\partial p_k} = \sum_{i=1}^{N} \frac{\partial|\varepsilon_i|}{\partial p_k} = \sum_{i=1}^{N} a_{ik}\,\mathrm{sgn}(\varepsilon_i) = \sum_{i=1}^{N} a_{ik}\,\frac{1}{|\varepsilon_i|}\varepsilon_i \tag{8}$$

Em notação matricial o gradiente da função Q_1 (equação 4) em relação ao vetor de parâmetros $\overline{\mathbf{p}}$ é:

$$\overline{\nabla}_{\overline{\mathbf{p}}}\{Q_1\} = \overline{\overline{\mathbf{A}}}^T\overline{\overline{\mathbf{W}}}\boldsymbol{\varepsilon} = \overline{\overline{\mathbf{A}}}^T\overline{\overline{\mathbf{W}}}\left(\overline{\mathbf{y}}^{\mathbf{o}} - \overline{\overline{\mathbf{A}}}\overline{\mathbf{p}}\right), \tag{9}$$

em que $\overline{\overline{\mathbf{W}}}$ (N x N) é uma matriz diagonal de pesos cujo i-ésimo elemento da diagonal é o inverso do valor absoluto do resíduo estimado na iteração anterior.

$$\overline{\overline{\mathbf{W}}} = \begin{bmatrix} 1/|\varepsilon_1| & & & \\ & 1/|\varepsilon_2| & & \\ & & \ddots & \\ & & & 1/|\varepsilon_N| \end{bmatrix}_{N\times N} \tag{9a}$$

A condição de mínimo é que a derivada da função Q_1 (equação 4) em relação ao vetor de parâmetros $\bar{\mathbf{p}}$ seja zero. Então fazendo

$$\bar{\nabla}_{\bar{\mathbf{p}}}\{Q_1\} = \bar{\bar{\mathbf{A}}}^T \bar{\bar{\mathbf{W}}}\left(\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}\bar{\mathbf{p}}\right) = \bar{\mathbf{0}}$$

temos

$$\bar{\bar{\mathbf{A}}}^T \bar{\bar{\mathbf{W}}}\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}^T \overline{\overline{\mathbf{W}\mathbf{A}}}\hat{\bar{\mathbf{p}}}_{L1} = \bar{\mathbf{0}}.$$

Logo chegamos ao sistema de equações:

$$\bar{\bar{\mathbf{A}}}^T \overline{\overline{\mathbf{W}\mathbf{A}}}\hat{\bar{\mathbf{p}}}_{L1} = \bar{\bar{\mathbf{A}}}^T \bar{\bar{\mathbf{W}}}\bar{\mathbf{y}}^{\mathbf{o}} \qquad (10)$$

em que $\hat{\bar{\mathbf{p}}}_{L1}$ é o estimador dos parâmetros via a minimização do vetor dos resíduos através da norma 1.

Como a matriz de peso $\bar{\bar{\mathbf{W}}}$ no sistema de equação 10 depende dos resíduos (i.e., depende dos parâmetros), então estamos diante de um sistema de equações não lineares. Uma das alternativas é usar o método iterativo dos mínimos quadrados reponderados (IRLS). Então, na k-ésima iteração estimamos um vetor de parâmetros $\hat{\bar{\mathbf{p}}}_{L1}^{(k)}$. O algoritmo começa com a solução de mínimos quadrados sobre determinado (estimador chamado neste tópico de $\hat{\bar{\mathbf{p}}}_{L2}$, veja a equação 3). Então, na iteração k = 1 estimamos:

$$\hat{\bar{\mathbf{p}}}_{L1}^{(1)} = \hat{\bar{\mathbf{p}}}_{L2},$$

em seguida, computamos o vetor de resíduos $\boldsymbol{\varepsilon}^{(1)} = \left(\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}\hat{\bar{\mathbf{p}}}_{L1}^{(1)}\right)$ e a matriz de pesos $\bar{\bar{\mathbf{W}}}^{(1)}$ (equação 9a) na iteração k=1. Então, ainda na iteração k=1, resolvemos o sistema de equações (equação 10):

$$\bar{\bar{\mathbf{A}}}^T \bar{\bar{\mathbf{W}}}^{(k)} \bar{\bar{\mathbf{A}}}\hat{\bar{\mathbf{p}}}_{L1}^{(k+1)} = \bar{\bar{\mathbf{A}}}^T \bar{\bar{\mathbf{W}}}^{(k)}\bar{\mathbf{y}}^{\mathbf{o}}$$

para estimarmos $\hat{\bar{\mathbf{p}}}_{L1}^{(2)}$ e o vetor de resíduos $\boldsymbol{\varepsilon}^{(2)} = \left(\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}\hat{\bar{\mathbf{p}}}_{L1}^{(2)}\right)$ e a matriz de pesos $\bar{\bar{\mathbf{W}}}^{(2)}$ (equação 9a) na iteração k=1+1. O processo é repetido até que a desigualdade

$$\frac{\left\| \hat{\overline{\mathbf{p}}}_{L1}^{(k+1)} - \hat{\overline{\mathbf{p}}}_{L1}^{(k)} \right\|_2}{1 + \left\| \hat{\overline{\mathbf{p}}}_{L1}^{(k+1)} \right\|_2} \leq \tau$$

em que τ é um número pequeno chamado de tolerância.

Este procedimento falha quando se algum resíduo for igual a zero por causa da matriz de pesos (equação 9a). A solução é definir um valor pequeno r de modo que se em alguma iteração k o i-ésimo resíduo $\varepsilon_i^{(k)}$ for menor que r então faz-se $\left|\varepsilon_i^{(k)}\right| = r$, e consequentemente teremos que o i-ésimo elemento da diagonal da matriz de pesos $\overline{\overline{\mathbf{W}}}^{(k)}$ é igual a $w_{ii}^{(k)} = 1/r$.

Qual o significado físico da minimização da Norma 1 do vetor de resíduos

Ao minimizarmos o vetor dos resíduos segundo a norma 1 em relação ao vetor de parâmetros $\overline{\mathbf{p}}$ estamos minimizando a função Q_1 (equação 4). Então minimizamos o sinal dos resíduos. Minimizar o sinal dos resíduos significa que podemos ter valores elevados dos resíduos, porém o somatório dos sinais destes resíduos deve ser mínimo. Isto faz com que a minimização da norma L1 dos resíduos despreze dados observados que sejam espúrios (outiliers). Dizemos que a minimização da norma L1 dos resíduos é mais robusta no sentido de permitir grandes resíduos (diferença entre os dados observados e os dados ajustados ou preditos). A Figura 1 mostra em pontos pretos um conjunto de observações geofísicas. Note que há um único dado espúrio (outilier). As retas tracejadas coloridas representam retas de ajustes (dados ajustados ou preditos) minimizando-se as normas 2 (reta L2) e 1 (reta L1) do vetor de resíduos $\boldsymbol{\varepsilon}$. O resíduo na i-ésima coordenada é a diferença entre o dado observado (pontos pretos) e o ajustado (retas ajustadas via normas L1 e L2). Note que a reta ajustada via norma L2 é atraída pelo ponto espúrio (outlier). Ao contrário, a minimização da norma L1 dos resíduos permite grandes resíduos pois minimiza-se o sinal dos resíduos. Logo a reta ajustada via norma L1 NÂO é atraída pelo ponto espúrio (outlier).

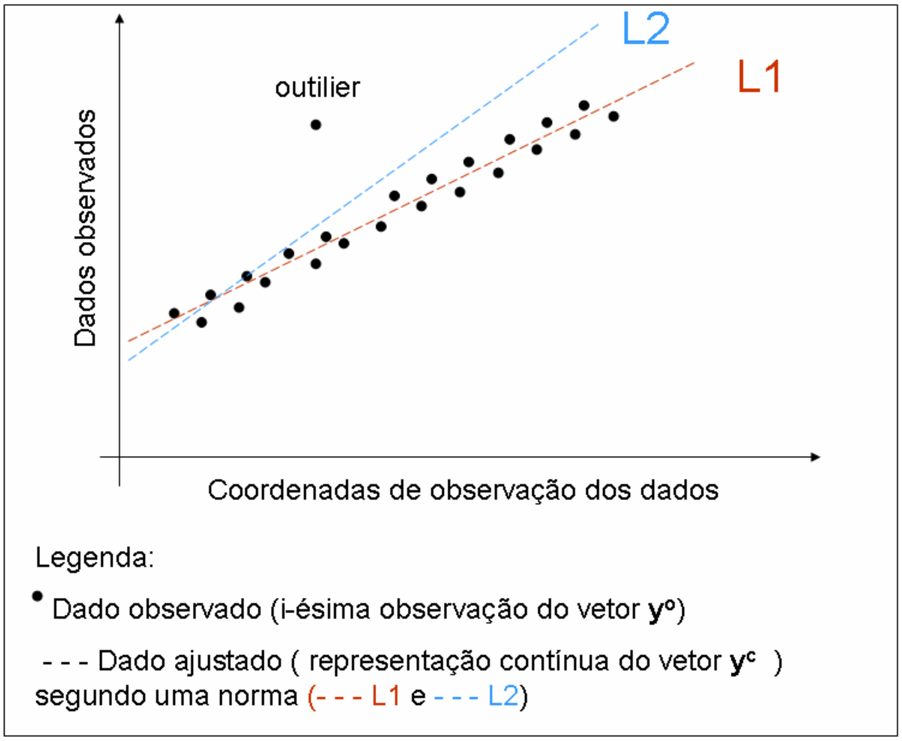

Figura 1

Estimador da Variação Total (TV)

A função TV $\varphi^{TV}(\mathbf{p})$ (Rudin et al., 1992) é definida como

$$\varphi^{TV}(\mathbf{p}) = \left\| \overline{\overline{\mathbf{B}}}\,\mathbf{p} \right\|_1 , \qquad (11)$$

em que $\overline{\overline{\mathbf{B}}}$ é uma matriz representando o operador discreto de primeiras derivadas em relação as direções horizontais de distribuição dos parâmetros e $\|\cdot\|_1$ denota norma ℓ_1. Note que o produto $\overline{\overline{\mathbf{B}}}\ \overline{\mathbf{p}}$ representa a diferença entre parâmetros fisicamente adjacentes. Portanto, a menos de uma constante, o produto $\overline{\overline{\mathbf{B}}}\ \overline{\mathbf{p}}$ quantifica uma aproximação da primeira derivada[1] da função contínua dos parâmetros. Veja o tópico 10 para relembrar os detalhes. A equação 11 representa o regularizador de Tikhonov de ordem 1 mas usando-se a norma L1.

A função TV não penalize as discontinuidades da distribuição espacial do vetor de parâmetros $\mathbf{p}$ de um modelo interpretativo (Vogel and Oman, 1996; 1998). Então, minimizando-se a função TV introduziremos a informação a prior que a distribuição espacial do vetor de parâmetros não será suave, mas descontínua.

Usando a definição da norma ℓ_1, a função TV function dada na equação 11 pode ser reescrita como

$$\varphi^{TV}(\mathbf{p}) = \sum_{l=1}^{L} \left| p_i - p_j \right|, \tag{12}$$

em que l entende-se pelo l-ésimo par, p_i e p_j de parâmetros espacialmente adjacentes em relação as direções de distribuição destes parâmetros e L é o número total de pares de parâmetros espacialmente adjacentes.

Como a função TV $\varphi^{TV}(\mathbf{p})$ não é diferenciável quando $p_i = p_j$, em geral usa-se a aproximação

[1] Derivada de uma função $f(x)$ em relação a x, considerando x_1 como um número particular no domínio de $f(x)$ é dada como:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x) - f(x_1)}{\Delta x}$$

$$\varphi^{TV}(\mathbf{p}) \approx \varphi_{\beta}{}^{TV}(\mathbf{p}) = \sum_{l=1}^{L}\left[(p_i - p_j)^2 + \beta\right]^{1/2} \qquad (13)$$

Proposta por Acar and Vogel (1994), em que β é um valor pequeno positivo. Assim a função $\varphi_{\beta}{}^{TV}(\mathbf{p})$ evita dificuldades associadas com a não diferenciabilidade da função $\varphi^{TV}(\mathbf{p})$ através da aproximação dos valores absolutos da função original $\varphi^{TV}(\mathbf{p})$ por uma função suave que remove a discontinuidade da derivada. A função $\varphi_{\beta}{}^{TV}(\mathbf{p})$ (equação 13) é em geral usada no problema vinculado não linear de minimizar

$$\lambda(\mathbf{p}) = \|\boldsymbol{\varepsilon}\|_2^2 + \mu(\delta)\varphi_{\beta}{}^{TV}(\mathbf{p})$$

Para detalhes veja a tese de doutorado do aluno do Observatório Nacional: Cristiano Mendel Martins (Martins, C.M. 2009)

Referencias

Acar, R., and C. R. Vogel, 1994, Analysis of total variation penalty methods: Inverse Problems, **10**, 1217–1229.

Martins, C.M. 2009, Inversão gravimétrica do relevo 3d de bacias sedimentares e da variação da densidade usando informação a priori sobre o ambiente geológico: Tese de doutorado do Observatório Nacional.

Rudin, L., S. Osher, and E. Fatemi, 1992, Nonlinear total variation based noise removal algorithms: Physica D, **60**, 259–68.

Vogel, C. R., and M. E. Oman, 1996, Iterative methods for total variation denoising: SIAM Journal of Scientific Computing, **17**, 227–238.

______, 1998, Fast, robust total variation-based reconstruction of noisy blurred images: IEEE Transactions on Image Processing, **7**, 813-824.