7
Extensão Intervalar

A extensão intervalar do método RCRI, que será apresentada nesse capítulo, tem como objetivo utilizar o método dentro do conceito de aritmética intervalar, que é a aritmética definida dentro do conjuntos dos intervalos da reta, $\mathbb{IR}$. A idéia é refazer toda a estrutura da aritmética usual, que está definida nos números reais, para o conjunto dos intervalos na reta. Um pequeno resumo sobre o assunto pode ser encontrado no apêndice C; para um estudo mais detalhado recomenda-se o livro de Stolfi e de Figueiredo (45).

A utilização da aritmética intervalar torna-se cada vez mais comum devido à necessidade de se trabalhar com erros, que podem vir de medidas imprecisas ou erros numéricos, por exemplo. Como os intervalos têm uma estrutura que inclui intuitivamente essa idéia de incerteza, já que não se trata mais de um único valor e sim um conjunto de possíveis valores, a aritmética intervalar passa a ser uma opção muito interessante, como mostra Kreinovich em (24).

Em muitas situações práticas torna-se necessária a interpolação de dados. Tal problema pode ser estendido, com o auxílio da aritmética intervalar, para a interpolação de dados intervalares, que ao longo dos anos vem sendo assunto de muitos trabalhos, como por exemplo, (1), (23), (37) e (41). A idéia da interpolação intervalar é estimar um intervalo de saída a partir de dados de entrada, depois de que o modelo tenha sido treinado com dados de treinamento que descrevem a relação entre esses dados de entrada e o intervalo de saída.

A proposta desse capítulo é utilizar o método RCRI para realizar aproximações intervalares, ou melhor, realizar regressões intervalares. Dessa forma, para um certo dado de entrada, o método fornece como saída não mais um número real e sim um intervalo. O modelo intervalar é praticamente o mesmo do caso real, algumas alterações foram feitas a fim de prepará-lo para receber dados intervalares, como pode ser visto na seção 7.1. Mais adiante será apresentado o algoritmo completo e por último uma aplicação.

7.1
O Que Muda na Estrutura?

O método RCRI será modificado de forma que ele possa trabalhar com dados intervalares. Então, a primeira grande mudança está nos dados de entrada. Os elementos do conjunto Y, que antes eram números reais, agora serão intervalos compactos na reta e o conjunto X continuará como antes, com entradas reais somente. Dessa forma os dados de entrada serão:

$$X = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^d \quad \text{e} \quad Y = \{[y_1], [y_2], \ldots, [y_N]\} \subset \mathbb{IR}$$

Polinômio

Uma vez que Y passa a ser um conjunto de intervalos, a construção do polinômio $\mathbf{p}$ também será alterada. O que será feito é substituir as operações em $\mathbb{R}$ para as operações no $\mathbb{IR}$. Assim, com base na equação (2-2), os coeficientes do polinômio $\mathbf{p}$ serão definidos pelo resultado do seguinte sistema intervalar:

$$\mathbf{M^t M}[\mathbf{a}] = \mathbf{M^t}[\delta] \tag{7-1}$$

onde a matriz real $\mathbf{M}$ e os vetores de intervalos $[\mathbf{a}]$ e $[\delta]$ estão definidos a seguir, considerando $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$.

$$[\delta] = \begin{pmatrix} [y_1] - [\bar{y}] \\ [y_2] - [\bar{y}] \\ \vdots \\ [y_n] - [\bar{y}] \end{pmatrix}_{n} \qquad [\mathbf{a}]^{\mathbf{t}} = \begin{pmatrix} [a] & [a_1] & \ldots & [a_{dd\ldots d}] \end{pmatrix}_{\mathbf{n}(g,\ d)}$$

$$M = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1d} & x_{11}^2 & x_{11}x_{12} & \ldots & x_{1d}^2 & x_{11}^3 & \ldots & x_{1d}^g \\ 1 & x_{21} & \ldots & x_{2d} & x_{21}^2 & x_{21}x_{22} & \ldots & x_{2d}^2 & x_{21}^3 & \ldots & x_{2d}^g \\ & & \vdots & & & & \vdots & & & \vdots & \\ 1 & x_{n1} & \ldots & x_{nd} & x_{n1}^2 & x_{n1}x_{n2} & \ldots & x_{nd}^2 & x_{n1}^3 & \ldots & x_{nd}^g \end{pmatrix}_{n \times \mathbf{n}(g,\ d)}$$

Além disso, $[\bar{y}]$ representa a média de um conjunto de intervalos, no caso Y. Na aritmética intervalar a média do conjunto $Y = \{[y_1], [y_2], \ldots, [y_N]\}$ é definida por:

$$[\bar{y}] = \left[\frac{1}{N} \sum_{i=1}^{N} \inf[y_i], \frac{1}{N} \sum_{i=1}^{N} \sup[y_i] \right].$$

Classificação

A classificação realizada em cada nó interno, que decide se um ponto de entrada está no filho à esquerda ou à direita, era feita através de $\mathbf{p}(\mathbf{x})$: se para o dado de entrada $\mathbf{x}$ tinha-se $\mathbf{p}(\mathbf{x}) \leq 0$, então ele era classificado para o filho à esquerda; caso contrário, sua classificação era para o filho à direita.

Agora que $\mathbf{p}(\mathbf{x})$ não é mais um número real e sim um intervalo, essa classificação sofrerá algumas modificações: se para o dado de entrada $\mathbf{x}$ tem-se $\mathbf{mid}[\mathbf{p}(\mathbf{x})] \leq 0$, então ele é classificado para o filho à esquerda; caso contrário, sua classificação é para o filho à direita; onde $\mathbf{mid}[\mathbf{p}(\mathbf{x})]$ indica o ponto médio do intervalo $[\mathbf{p}(\mathbf{x})]$. Dessa forma os conjuntos X_1, X_2, Y_1 e Y_2 serão definidos por:

$$\begin{array}{lllll} X_1 & = & \{\mathbf{x} \in X \mid \mathbf{mid}[\mathbf{p}(\mathbf{x})] \leq 0\} & X_2 & = & \{\mathbf{x} \in X \mid \mathbf{mid}[\mathbf{p}(\mathbf{x})] > 0\} \\ Y_1 & = & \{[y_i] \in Y \mid \mathbf{x}_i \in X_1\} & Y_2 & = & \{[y_i] \in Y \mid \mathbf{x}_i \in X_2\} \end{array}$$

Critérios de Parada

O único critério de parada que muda é o 3º. A mudança será simplesmente que em vez de Y_1 e Y_2 serão usados os vetores W_1 e W_2, formados pelos pontos médios dos intervalos:

$$W_1 = \{\mathbf{mid}[y_i] \mid [y_i] \in Y_1\} \qquad W_2 = \{\mathbf{mid}[y_i] \mid [y_i] \in Y_2\}.$$

Sejam $\overline{W}_1$, $\overline{W}_2$, S_1^2 e S_2^2 as médias e variâncias amostrais de W_1 e W_2. Escolhido o nível de significância α e calculado o valor de t_0 de acordo com a expressão (7-2) abaixo, esse critério de parada passará a ser feito da seguinte maneira: se $-t_{\frac{\alpha}{2},n_1+n_2-2} < t_0 < t_{\frac{\alpha}{2},n_1+n_2-2}$, esse nó não será dividido e passa a ser uma folha; caso contrário, a divisão será mantida.

$$t_0 = \frac{\overline{W}_1 - \overline{W}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \qquad S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \tag{7-2}$$

Estimativa Local

Como o vetor Y é composto por intervalos, os elementos nas folhas também serão intervalos. As estimativas locais e globais foram modificadas para se enquadrarem aos dados intervalares. A equação (7-3) mostra como fica a nova estimativa local:

$$\widehat{[c_k]} = \frac{\sum_{i=1}^{N} \mathbf{G}_k(\mathbf{x}_i)[y_i]}{\sum_{i=1}^{N} \mathbf{G}_k(\mathbf{x}_i)}. \tag{7-3}$$

Estimativa Global

Com base na equação (3-8) a estimação global dos intervalos nas folhas serão as coordenadas do vetor $[\mathbf{c}]$, que é a solução do sistema (7-4) definido a seguir:

$$\mathbf{G^t G}[\mathbf{c}] = \mathbf{G^t}[\mathbf{y}] \tag{7-4}$$

onde $\mathbf{G}$, $[\mathbf{c}]$ e $[\mathbf{y}]$ são dados por:

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1(\mathbf{x}_1) & \mathbf{G}_2(\mathbf{x}_1) & \dots & \mathbf{G}_l(\mathbf{x}_1) \\ \mathbf{G}_1(\mathbf{x}_2) & \mathbf{G}_2(\mathbf{x}_2) & \dots & \mathbf{G}_l(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \mathbf{G}_1(\mathbf{x}_N) & \mathbf{G}_2(\mathbf{x}_N) & \dots & \mathbf{G}_l(\mathbf{x}_N) \end{pmatrix} \quad [\mathbf{c}] = \begin{pmatrix} [c_1] \\ \vdots \\ [c_l] \end{pmatrix} \quad [\mathbf{y}] = \begin{pmatrix} [y_1] \\ [y_2] \\ \vdots \\ [y_N] \end{pmatrix}$$

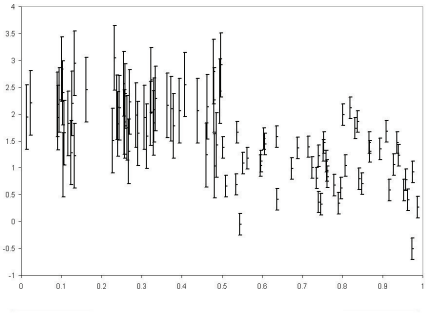
Estimativa Minmax

Além das estimativas locais e globais pode ser interessante, para a extensão intervalar, definir uma terceira forma de estimar os elementos nas folhas, que será chamada por estimativa minmax. Nessa nova forma serão considerados o mínimo entre os inf e o máximo entre os sup, de acordo com a equação (7-5) a seguir.

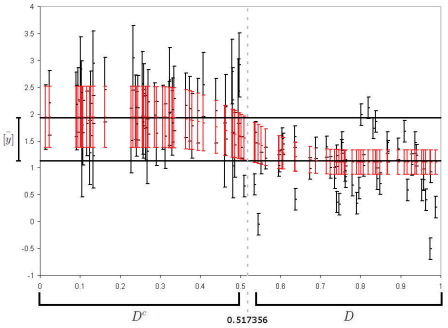
$$\widehat{[c_k]} = \left[\min_{1 \le i \le N} \mathbf{G}_k(\mathbf{x}_i) \inf[y_i] \ , \ \max_{1 \le i \le N} \mathbf{G}_k(\mathbf{x}_i) \sup[y_i] \right] \tag{7-5}$$

Exemplo 7.1.1 Esse exemplo é a extensão dos exemplos 2.2.1 e 3.3.2, os quais apresentam um problema onde os elementos do conjunto X são reais e o grau do polinômio que divide o domínio é um. A extensão intervalar substitui y_i por $[y_i]$, ou seja, substitui o que antes era um número real por um intervalo na reta. A figura 7.1(a) mostra os novos dados de entrada. As estimativas local, global e minmax estão apresentadas nas figuras 7.1(b), 7.1(c) e 7.1(d), respectivamente.

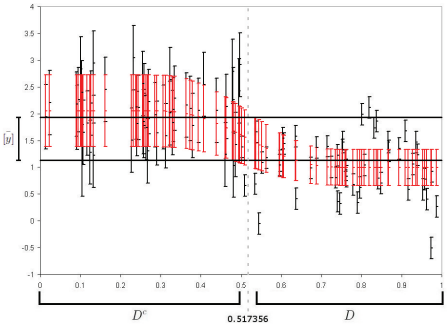
Observe que os tamanhos dos intervalos da esquerda são maiores que os da direita e que isso foi capturado pelas três estimativas. Veja também que o método RCRI intervalar fornece uma estimativa melhor do que simplesmente $[\bar{y}]$, no caso da estimativa local e global, e $[\min_{1 \le i \le N}(\inf[y_i]), \max_{1 \le i \le N}(\sup[y_i])]$, no caso da estimativa minmax. Nesse caso é considerada uma melhor estimativa aquela cujo ponto médio e tamanho do intervalo estimado estão mais próximos dos valores do ponto médio e do tamanho do intervalo desejados.



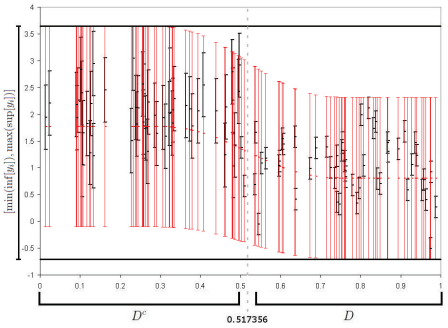
7.1(a): Dados de entrada



7.1(b): Estimativa Local



7.1(c): Estimativa Global



7.1(d): Estimativa Minmax

Figura 7.1: Exemplo intervalar em uma variável.

7.2 Algoritmo Intervalar

A seguir está o algoritmo para a versão intervalar do RCRI, que mostra cada passo desde a criação dos dados de entrada até as diferentes estimativas já apresentadas.

Algoritmo Intervalar: Regressão Construtiva por Regiões Implícitas

Passo 0) Sejam $Y = \{[y_1], [y_2], \ldots, [y_N]\} \subset \mathbb{IR}$, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^d$, g = grau do polinômio $\mathbf{p}$, d_{lim} = profundidade limite, $n_{lim} = qN > n(g,d)$, $0 < \alpha < 1$ parâmetro do teste de hipótese do 4º critério de parada e λ = parâmetro da função $\mathbf{G}$, todos determinados a priori.

Passo 1) Para X e Y definidos, encontre os coeficientes de $\mathbf{p}$ de acordo com a equação (7-1).

Passo 2) Defina,

$$\begin{aligned} X_1 &= \{\mathbf{x} \in X \mid \mathbf{mid}[\mathbf{p}(\mathbf{x})] \leq 0\} \\ X_2 &= \{\mathbf{x} \in X \mid \mathbf{mid}[\mathbf{p}(\mathbf{x})] > 0\} \\ Y_1 &= \{[y_i] \in Y \mid \mathbf{x}_i \in X_1\} \\ Y_2 &= \{[y_i] \in Y \mid \mathbf{x}_i \in X_2\} \\ n_1 &= \text{ número de elementos em } X_1 \\ n_2 &= \text{ número de elementos em } X_2 \end{aligned}$$

Passo 3) Se $n_1 > n_{lim}$ e $n_2 > n_{lim}$, então siga para o passo 4.
Se $n_1 \leq n_{lim}$ ou $n_2 \leq n_{lim}$, então esse nó será uma folha. Pule para o passo 9 para definir os intervalos nas folhas.

Passo 4) Seja d a profundidade desse nó da árvore.
Se $d < d_{lim}$, então siga para o passo 5.
Se $d \geq d_{lim}$, então esse nó será uma folha. Pule para o passo 9 para definir os intervalos nas folhas.

Passo 5) Seja t_0 é dado pela equação (7-2).
Se $-t_{\frac{\alpha}{2},n_1+n_2-2} < t_o < t_{\frac{\alpha}{2},n_1+n_2-2}$, siga para o passo 6.
Se $t_o < -t_{\frac{\alpha}{2},n_1+n_2-2}$ ou $t_o > t_{\frac{\alpha}{2},n_1+n_2-2}$, então esse nó será uma folha. Pule para o passo 9 para definir os intervalos nas folhas.

Passo 6) É criado um nó interno e os coeficientes de $\mathbf{p}$ são armazenados por ele.

Passo 7) Para determinar o filho à esquerda desse nó interno, volte ao passo 1 com $X = X_1$ e $Y = Y_1$.

Passo 8) Para determinar o filho à direita desse nó interno, volte ao passo 1 com $X = X_2$ e $Y = Y_2$.

Passo 9) Para estimar os intervalos nas folhas escolha entre os métodos local, global e minmax, apresentados em (7-3), (7-4) e (7-5), respectivamente.

7.3 Aplicação

A aplicação intervalar apresentada nessa seção será em geologia do petróleo, semelhante à apresentada na seção 6.2 do capítulo anterior. A motivação de implementar esse problema de forma intervalar vem do fato de que muitas vezes não importa tanto o valor de y, no caso rhob, e sim o provável intervalo que ele se encontra.

A amostra inicial em cada um dos poços foi dividia por dez, para gerar os dados intervalares da seguinte maneira: para cada poço, em profundidade, os dados foram agrupados de 10 em 10. Os atributos associados a cada um desses grupos foram as médias dos atributos, e os limites do intervalo de rhob foram o menor e o maior valor de rhob dentro do grupo.

Dessa forma o tamanho de cada amostra foi reduzido a sua décima parte e 90% dela foi usada como dados de aprendizagem e os outros 10% como dados fora da amostra. A redução no tamanho da amostra implica em uma redução também no valor do parâmetro d_{lim}, que passa a assumir $d_{lim} = 4$. Já os demais parâmetros foram os mesmos do caso não intervalar. Novamente os dados de aprendizagem foram escolhidos de forma aleatória em 10 diferentes rodadas, a fim de garantir uma boa amostragem e descartar a possibilidade de que o sorteio favoreça ou prejudique o modelo.

Resultados

Em cada uma das 10 vezes que o método foi rodado foram determinados a porcentagem dos pontos fora da amostra que estavam dentro do intervalo estimado e o tamanho médio dos intervalos estimados. Então, para cada uma das três formas de estimar os intervalos (local, global ou minmax), foram computados 10 diferentes valores da porcentagem de acertos e 10 diferentes tamanhos médios dos intervalos. As estatísticas para esses valores encontram-se nas tabelas 7.3 e 7.1 a seguir.

Como base de comparação, para cada poço, os tamanhos dos intervalos formados pelos valores mínimo e máximo de rhob dentro da amostra são:

$$\begin{array}{rclrcl} \text{Norte} & = & 1,1261 & \text{Sul} & = & 1,6517 \\ \text{Oeste} & = & 1,1472 & \text{Leste} & = & 1,1496. \end{array}$$

Talvez, mais interessante do que o dados da tabela 7.1 seja a informação de quanto o tamanho do intervalo foi reduzido com a aplicação do método. Esse resultados encontra-se na tabela 7.2.

	RCRI (Local)	RCRI (Global)	RCRI (Minmax)	RCRI (Local)	RCRI (Global)	RCRI (Minmax)
mediana	0,497	0,876	0,953	0,399	0,740	1,257
MAD	0,003	0,026	0,029	0,003	0,030	0,055
mínimo	0,487	0,804	0,836	0,390	0,673	1,174
máximo	0,503	0,920	1,044	0,408	0,831	1,423
		Norte			Sul	

	RCRI (Local)	RCRI (Global)	RCRI (Minmax)	RCRI (Local)	RCRI (Global)	RCRI (Minmax)
mediana	0.505	1.012	0.889	0.445	0.853	0.928
MAD	0.006	0.057	0.034	0.002	0.040	0.027
mínimo	0.496	0.920	0.813	0.433	0.805	0.857
máximo	0.511	1.142	0.933	0.450	1.021	0.980
		Oeste			Leste	

Tabela 7.1: Estatísticas, dos quatro poços, para a média dos tamanhos dos intervalos gerados.

	RCRI (Local)	RCRI (Global)	RCRI (Minmax)
Norte:	$55,87\%$	$22,21\%$	$15,37\%$
Sul:	$75,84\%$	$55,20\%$	$23,90\%$
Oeste:	$55,98\%$	$11,79\%$	$22,51\%$
Leste:	$61,29\%$	$25,80\%$	$19,28\%$

Tabela 7.2: Redução no tamanho do intervalo fornecido, por cada método e em cada um dos quatro poços.

Então, juntando as informações das tabelas 7.2 e 7.3, pode-se concluir, por exemplo, que no caso do poço Norte, em média, o tamanho do intervalo fornecido pelo método RCRI com a estimativa local foi reduzido em $55,87\%$ e a porcentagem de acertos, em média, se manteve em $92,27\%$. Por outro lado, ainda para o poço Norte, em média, o tamanho do intervalo fornecido pelo método RCRI com a estimativa minmax foi reduzido em apenas $15,37\%$, mas a porcentagem de acertos, em média, foi $99,92\%$.

Avaliando a relação diminuição do tamanho do intervalo e porcentagem de acertos, parece que a forma local fornece um boa solução. Se for possível permitir alguns erros, menos de 10% em geral, o método RCRI com a forma local tem como saída intervalos bem reduzidos, passando de 60% para o poço Leste.

Diante dessas informações, deve-se escolher o método que forneça a maior redução no tamanho do intervalo sem deixar de garantir uma porcentagem mínima de acertos exigida.

	RCRI (Local)	RCRI (Global)	RCRI (Minmax)	RCRI (Local)	RCRI (Global)	RCRI (Minmax)
mediana	92,27%	98,38%	99,92%	93,66%	98,14%	99,77%
MAD	0,38%	0,33%	0,08%	0,38%	0,34%	0,08%
mínimo	91,56%	97,78%	99,80%	92,86%	97,49%	99,70%
máximo	92,67%	98,79%	100,00%	94,53%	99,09%	100,00%
		Norte			Sul	

	RCRI (Local)	RCRI (Global)	RCRI (Minmax)	RCRI (Local)	RCRI (Global)	RCRI (Minmax)
mediana	92,27%	99,81%	99,68%	94,93%	98,92%	99,90%
MAD	1,23%	0,19%	0,19%	0,49%	0,25%	0,10%
mínimo	89,09%	99,48%	99,09%	92,22%	97,64%	99,61%
máximo	94,03%	100,00%	100,00%	95,47%	99,21%	100,00%
		Oeste			Leste	

Tabela 7.3: Estatísticas, dos quatro poços, para as porcentagens de acerto.