ALGORITMO PARA SUAVIZAÇÃO DE DADOS VIA AJUSTE DE CURVAS POR MÍNIMOS QUADRADOS COM DETECÇÃO DE DADOS INVÁLIDOS

Valcir Orlando
Roberto Vieira da Fonseca Lopes
Departamento de Mecânica Espacial e Controle - Instituto de Pesquisas Espaciais - INPE
São José dos Campos - Brasil

RESUMO

Apresenta-se um algoritmo para suavização ou validação de dados via ajuste de curvas por mínimos quadrados, com procedimento de seleção automática do grau da curva baseado na distribuição Chi-quadrada. O algoritmo compreende um procedimento de detecção de dados inválidos baseado na distribuição F, o qual atinge seu objetivo através do monitoramento da contribuição de cada dado na soma dos quadrados dos resíduos. A aplicação deste procedimento permite que se trabalhe com taxas de dados inválidos sobre dados válidos maiores que o procedimento usual de rejeição de dados baseado em uma faixa de tolerância em torno da curva ajustada. Isto pode ser claramente observado comparando os resultados obtidos nos testes.

ABSTRACT

An algorithm for data smoothing or validation by least square curve fitting, with a Chi-Square procedure of curve degree automatic selection is presented. The algorithm comprises an invalid data detection procedure based on the F distribution, which reaches its objective monitoring the contribution of each data point in the quadratic residuum sum. The procedure application allows handling with a ratio of invalid to valid data greater than the usual procedure of data editing based on a tolerance zone around of adjusted curve. This can be clearly observed by comparing the tests results.

INTRODUÇÃO

O objetivo principal do algoritmo de suavização via ajuste de curvas por mínimos quadrados, a ser apresentado neste trabalho, é permitir um bom desempenho no processo de ajuste, mesmo em presença de dados inválidos, através de eficiente detecção destes. Dados inválidos são automaticamente detectados e rejeitados em duas etapas. A primeira é efetuada através de técnica desenvolvida para essa finalidade com base na distribuição F, aplicável durante processo de seleção automática da curva ajustada, no caso de ajuste polinomial. A outra, utiliza uma técnica usual de rejeição de dados baseada em uma faixa de tolerância em torno da curva ajustada, aplicável somente após a conclusão do processo de ajuste.

O algoritmo pode ser também utilizado apenas como processo de validação de dados, aplicação esta na qual o interesse principal não recai sobre a curva suavizadora em si e sim na detecção e sinalização de dados inválidos com o objetivo de facilitar processamentos futuros.

AJUSTE DE CURVAS POR MÍNIMOS QUADRADOS E TÉCNICAS PARA SELEÇÃO AUTOMÁTICA DO GRAU E PARA REJEIÇÃO DE DADOS INVÁLIDOS

Serão brevemente apresentados nesta seção a técnica de ajuste de curvas por mínimos quadrados, uma técnica de seleção automática do grau da curva ajustada, no caso em que esta é um polinômio, e uma técnica usual de rejeição automática de dados inválidos. Todas estas técnicas se encontram difundidas na literatura técnica e suas descrições, aqui, visam apenas colocar resultados que interessam ao desenvolvimento do algoritmo de suavização a ser apresentado, de modo a facilitar a compreensão do trabalho.

AJUSTE DE CURVAS POR MÍNIMOS QUADRADOS

A técnica de ajuste de curvas por mínimos quadrados é aplicável a conjuntos de dados que sigam qualquer distribuição estatística. Neste trabalho, porém, ficar-se-á restrito a dados que sigam distribuição normal, para os quais essa técnica é equivalente à chamada técnica Chi-quadrado [1]. Considere-se que se dispõe de um conjunto de dados y_i, $i = 1,2,\ldots,m$, afetados de erros aleatórios gaussianos com média nula e não correlacionados entre si. Considerem-se esses dados modelados matematicamente por:

$$y_i = f(x_i, C_0, C_1, \ldots, C_r) + \varepsilon_i \quad , \quad i = 1,2,\ldots,m \quad , \qquad (1)$$

onde ε_i, $i = 1,2,\ldots,m$ é uma sequência branca gaussiana e:

$$f(x, C_0, C_1, \ldots, C_r) = C_0 g_0(x) + C_1 g_1(x) + \ldots + C_r g_r(x) \quad , \qquad (2)$$

sendo $g_k(x)$, $k = 0,1,\ldots,r$ funções ortogonais (conhecidas) da variável independente x e C_k, $K = 0,1,\ldots,r$ são constantes reais.

Considere-se o problema de determinar os valores dos coeficientes C_k que minimizam o funcional:

$$S = \sum_{i=1}^{m} w_i [y_i - \sum_{k=1}^{r} C_K g_k(x_i)]^2 \quad , \tag{3}$$

onde w_i são parâmetros conhecidos.

A solução desse problema, amplamente difundida na literatura [1], [2] é:

$$\hat{C} = [A^T W A]^{-1} A^T W y \quad , \tag{4}$$

$$\text{cov}\,[\hat{C}] = [A^T W A]^{-1} \quad , \tag{5}$$

onde $\hat{C} = [\hat{C}_0, \hat{C}_1, \ldots, \hat{C}_r]^T$ é o vetor cujas componentes são os valores estimados dos coeficientes da curva ajustada, $y = [y_1, y_2, \ldots, y_m]^T$,

$$A = \begin{bmatrix} g_0(x_1) & g_1(x_1) & \cdots & g_r(x_1) \\ g_0(x_2) & g_1(x_2) & \cdots & g_r(x_2) \\ \vdots & \vdots & & \vdots \\ g_0(x_m) & g_1(x_m) & \cdots & g_r(x_m) \end{bmatrix} \quad , \tag{6}$$

$$W = \begin{bmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_m \end{bmatrix} \quad . \tag{7}$$

Neste trabalho só será considerado o caso em que as funções $g_k(x)$ são polinômios de grau k. Um problema importante relacionado a este caso é a escolha do grau do polinômio ajustado. Se o grau for muito alto podem ser incluídos ruídos de freqüências mais altas, componentes dos dados, nos valores suavizados. Por outro lado, se o grau for muito baixo, podem ser suprimidas componentes de freqüências mais altas que ainda contêm informações de interesse, juntamente com o ruído.

No caso presente, respeitadas as hipóteses efetuadas quanto às propriedades estatísticas do erro nos dados e tomando-se $w_i = \frac{1}{\tau_i^2}$, $i = 1,2,\ldots,m$, onde τ_i é o desvio-padrão do erro aleatório no dado y_i, pode-se selecionar automaticamente o grau do polinômio suavizador, através do emprego de uma técnica difundida na literatura, para monitoramento da qualidade do ajuste [1], [2]. Esta técnica será resumidamente descrita a seguir.

TÉCNICA DE SELEÇÃO AUTOMÁTICA DO GRAU DO POLINÔMIO

Considerando válidas todas as hipóteses efetuadas anteriormente, ajustem-se gradativamente aos dados polinômios de grau $r = 1,2,\ldots$. Após cada ajuste, calcule-se o valor de S correspondente pela equação 3 e verifique-se se:

$$S < \chi^2_\alpha \ (m - r - 1) \ , \tag{8}$$

onde $\chi^2_\alpha \ (m - r - 1)$ é um número real positivo tal que:

$$\text{probalidade} \ [\chi^2(m - r - 1) > \chi^2_\alpha(m - r - 1)] = \alpha \ , \tag{9}$$

sendo $\alpha \in [0,1]$ um fator de tolerância prédeterminado e $\chi^2(m-r-1)$ uma variável aleatória Chi-quadrada com $m - r - 1$ graus de liberdade.

Para $r \geq 2$ deve-se verificar também se:

$$S_R = \left| 1 - \frac{S(r)}{S(r-1)} \right| < \xi \ , \tag{10}$$

onde $S(r)$ e $S(r-1)$ representam os valores calculados para S a partir dos ajustes de grau r e $r-1$ respectivamente, e $\xi \in R$ é um fator de tolerância prédeterminado.

Quando pela primeira vez uma das condições, 8 ou 10, for afirmativa, aceite-se o ajuste atual como satisfatório. A satisfação da condição 8 para um determinado valor de r indica que, para esse valor, S passa a se aproximar de uma variável aleatória Chi-quadrada com m-r-1 graus de liberdade [1], [2]. Isto significa que para o primeiro valor de r, que satisfaz à condição 8, a curva ajustada se aproxima, nos pontos relativos aos instantes de amostragem dos dados, dos valores esperados dos dados correspondentes. O modelo matemático suposto para os dados na equação 1 passa então a se tornar adequado. Os valores da curva ajustada, calculados nos instantes de amostragem dos dados, podem então ser encarados como estimativas do valor real da grandeza medida, nesses instantes, já que tomando o valor esperado de y_i, $i = 1,2,\ldots,m$ e considerando a hipótese, inicialmente efetuada, de que os dados só estão afetados de erros aleatórios correspondentes a ruídos brancos gaussianos, tem-se: $E[y_i] = y_{R_i}$.

O motivo de testar também a condição 10, indicativa de convergência relativa, é o de que a existência de possíveis dados inválidos, com erros muito elevados, no conjunto pode fazer com que a condição 8 só venha a ser satisfeita pela primeira vez para valor irrealisticamente elevado do grau. Depois de grandes saltos iniciais decrescentes, o valor de S tende a se estabilizar, variando muito pouco. Neste caso, a convergência relativa, expressa pela inequação 10, seria alcançada antes da convergência absoluta, expressa pela inequação 8. Um valor comumente utilizado para o fator de tolerância ξ, citado na literatura, é 0,1 [2].

A seguir será apresentada uma técnica de rejeição de dados inválidos, também difundida na literatura, que será incorporada ao algoritmo de suavização desenvolvido como um refinamento complementar à téc

nica desenvolvida para a mesma finalidade.

DESCRIÇÃO RESUMIDA DE TÉCNICA USUAL DE REJEIÇÃO AUTOMÁTICA DE DADOS.

Considerem-se também aqui como válidas todas as hipóteses anteriormente efetuadas quanto ao erro nos dados e sua distribuição e suponha-se ter efetuado o ajuste de um polinômio que melhor represente o valor esperado dos dados no intervalo considerado. Calculem-se então os valores suavizados destes por:

$$\hat{y}_i = \sum_{k=0}^{r} \hat{C}_k \, g_k(x_i) \ , \quad i = 1,2,\ldots,m \ . \tag{11}$$

Calcule-se a seguir o resíduo, R_i, $i = 1,2,\ldots,m$, entre o valor de cada dado original, y_i, e o correspondente valor suavizado, verificando-se a seguir, para cada resíduo calculado, se:

$$|R_i| = |\, y_i - \bar{y}_i| > \eta\tau_i \ , \tag{12}$$

onde η é um fator de tolerância preestabelecido e τ_i é o desvio-padrão do dado y_i, $i = 1,2,\ldots,m$.

Cada vez que a desigualdade acima não for satisfeita aceita-se então o dado y_i como válido. Caso contrário, rejeita-se esse dado zerando o valor de w_i correspondente ao peso a ele associado, conforme a equação 3. Sempre que for constatada a existência de pelo menos um dado inválido, repete-se a técnica de ajuste polinomial utilizando os novos pesos para os dados rejeitados. Uma vez de posse dos novos coeficientes, repete-se a verificação da desigualdade 12. O procedimento deve ser repetido até que não se detecte, nessa verificação, nenhum ponto inválido.

Se a taxa de dados inválidos sobre dados válidos for muito alta, o procedimento descrito acima pode rejeitar também dados válidos e, com isso, não apresentar um bom desempenho. Pode, ainda, ocorrer a rejeição de dados válidos, mesmo para baixo valores dessa taxa, se existirem no conjunto utilizado no ajuste dados com erros muito elevados. Esses dados poderiam influir no ajuste de modo a, em suas proximidades, desviar a curva ajustada dos valores esperados dos dados válidos. Com isto, pontos válidos localizados nessas proximidades poderiam cair fora da região de tolerância em torno da curva ajustada, o que causaria as suas rejeições.

A técnica de rejeição automática a ser apresentada na seqüência, foi desenvolvida com o objetivo de elaborar um algoritmo no qual a ocorrência de problemas como os acima citados seja reduzida.

TÉCNICA DESENVOLVIDA PARA REJEIÇÃO AUTOMÁTICA DE DADOS INVÁLIDOS

A idéia básica da técnica consiste na monitoração da contribuição de cada dado no valor de S, expresso na equação 3. Para isto, após a efetuação de um determinado ajuste, propõe-se inicialmente com $m' = m$ o cálculo de:

$$\frac{F_j(1,\ m'-r-2)}{m'-r-2} = \frac{\frac{1}{\tau_j^2}\left[y_j - \sum_{k=0}^{r} C_k g_k(x_j)\right]^2}{\sum_{\substack{i=1 \\ i\neq j}}^{m} \frac{1}{\tau_i^2}\left[y_i - \sum_{k=0}^{r} C_k g_k(x_i)\right]^2}\ ,\quad j = 1,2,\ldots,m \quad (13)$$

O lado direito da expressão acima é a relação entre a parcela relativa à contribuição do dado y_j para o valor de S sobre o valor que teria S sem essa parcela. Para valor realístico do grau, r, do polinômio ajustado este deve se aproximar do valor esperado dos dados nos instantes de amostragem. Com isso, o numerador da relação acima se aproxima de uma variável aleatória Chi-quadrada com um grau de liberdade e o denominador de uma variável Chi-quadrada com $m-r-2$ graus de liberdade. As variáveis $F_j(1,\ m-r-2)$, $j = 1,2,\ldots,m$ se aproximariam então de uma variável aleatória com distribuição F, já que uma variável seguindo essa distribuição é definida por [1], [3]:

$$F\ (\nu_1,\ \nu_2) = \frac{\chi^2(\nu_1)/\nu_1}{\chi^2(\nu_2)/\nu_2}\ ,$$

onde $\chi^2(\nu_1)$ e $\chi^2(\nu_2)$ são duas variáveis aleatórias Chi-quadradas com ν_1 e ν_2 graus de liberdade respectivamente.

No caso de $F_j(1, m'-r-2)$ tem-se, portanto, $\nu_1 = 1$ e $\nu_2 = m'-r-2$.

Uma vez calculados os valores de $F_j(1, m-r-2)$, $j = 1,2,\ldots,m$, pela equação 13, verifica-se se é satisfeita para cada valor calculado a desigualdade:

$$F_j(1, m'-r-2) > F_\beta(1, m'-r-2)\ , \quad (15)$$

onde F_β é um número real positivo e β um fator de tolerância predeterminado tal que:

$$\text{probabilidade}\ [F(1, m'-r-2) > F_\beta(1, m'-r-2)] = \beta\ , \quad (16)$$

sendo $F(1, m'-r-2)$ uma variável aleatória com distribuição F para a qual $\nu_1 = 1$ e $\nu_2 = m'-r-2$.

Se o valor calculado para $F_j(1, m'-r-2)$ ultrapassar, para um determinado j, o valor $F_\beta(1, m'-r-2)$, isto significará que a contribuição do ponto y_j para o valor total de S é maior do que seria esperado considerando-se as propriedades estatísticas que, por hipótese, possuem os erros nos dados (seqüência gaussiana branca), desde que a escolha do fator de tolerância, β, tenha recaído sobre um valor suficientemente pequeno. O dado y_j deve então, por esse motivo, ser rejeitado zerando o peso, w_j, a ele associado.

Cada vez que pelo menos um dado inválido é detectado repete-se o procedimento de ajuste com conseqüente repetição do procedimento de rejeição de dados, agora tomando-se $m' = m-p$, sendo p o número total

de pontos rejeitados.

Note-se que para cada dado rejeitado o número de graus de liberdade do denominador de $F_j(1, m' - r - 2)$ diminui de uma unidade. Isto é facilmente verificado, já que a rejeição de um ponto qualquer, y_i, através da anulação do peso a ele associado, $w_i = 1/\tau_i^2$, equivale a considerar que $\tau_i^2 = \infty$, o que anula a parcela do denominador correspondente a esse dado. Este é o motivo de tomar $m' = m - p$.

Para que os testes de verificação da desigualdade 15 possam ser efetuados, deve-se dispor de uma tabela de $F_\beta(\nu_1, \nu_2)$ para o valor escolhido de β, $\nu_1 = 1$ e diversos valores de ν_2.

Quando o procedimento de seleção automática do grau for utilizado no ajuste, propõe-se a utilização desta técnica de rejeição antes de testar o grau de cada curva ajustada. Cada vez que pelo menos um ponto inválido for detectado, o procedimento de ajuste deve ser reiniciado a partir do grau unitário. Quando nenhum ponto inválido é detectado, então efetua-se o teste de verificação do grau da curva, aceitando ou não o ajuste, dependendo do resultado deste teste. Se o ajuste não é aceito, então, de acordo com o procedimento de seleção automática do grau descrito anteriormente, deve-se repetir o ajuste para uma curva de grau uma unidade maior. Efetuado o novo ajuste repete-se, antes do teste do grau, a aplicação da técnica de rejeição. O processo termina quando para um determinado grau o ajuste é aceito, pois isto implica a não-detecção de qualquer ponto inválido em toda a interação presente.

Para refinar o procedimento, no algoritmo para suavização de dados desenvolvido, após a convergência do processo acima descrito, aplica-se ainda a técnica usual de rejeição automática de dados inválidos com base na faixa de $\eta\tau_i$ ao redor da curva ajustada. Eventuais dados inválidos que possam não ter sido dectados pela primeira técnica o serão provavelmente por esta última, já que a aplicação a priori da primeira deverá pelo menos reduzir a taxa de dados inválidos sobre os válidos, melhorando o desempenho da última.

ALGORÍTMO DESENVOLVIDO PARA SUAVIZAÇÃO DE DADOS, VIA AJUSTE POLINOMIAL, COM SELEÇÃO AUTOMÁTICA DO GRAU E REJEIÇÃO DE DADOS INVÁLIDOS

Na figura 1 apresenta-se um diagrama esquemático do algoritmo. Ele consiste nos seguintes passos:

1. Tomar $m' = m$ (número de pontos) e seguir ao passo 2.

2. Ajustar aos dados uma curva de grau r pelo procedimento de mínimos quadrados descrito anteriormente (iniciar com $r = 1$). Feito o ajuste, calcular S e S_R através das equações 3 e 10 respectivamente. Se $r = 1$, calcular apenas o valor de S. Tomar $j = 1$ e seguir para o passo 3.

3. Calcular $F_j(1, m' - r - 2)$ com auxílio da equação 13 e comparar o valor encontrado com o valor tabelado $F_\beta(1, m' - r - 2)$ e:

   a) Se $F_j(1, m' - r - 2) < F_\beta(1, m' - r - 2)$ e $j < m$, então incrementar j de uma unidade e repetir o passo presente. Se $j = m$ ir para o passo 4.

b) Se $F_j(1, m' - r - 2) > F_\beta(1, m' - r - 2)$, então tomar $w_i = 0$ e retornar ao passo 2 fazendo $r = 1$ e decrementando m' de uma unidade.

4. Comparar o valor de S, calculado no passo 2, com o valor tabelado $\chi^2_\alpha\ (m' - r - 1)$ e:

a) Se $S > \chi^2_\alpha\ (m' - r - 1)$ e $r = 1$, retornar ao passo 2 incrementando r de uma unidade. Se $r > 1$, então testar a convergência relativa, isto é, verificar se $S_R > \xi$. Em caso afirmativo, retornar ao passo 2 incrementando r de uma unidade. Caso contrário considerar a curva de grau ajustada como satisfatória e ir para o passo 5.

b) Se $S < \chi^2_\alpha\ (m' - r - 1)$, então foi obtida a convergência absoluta. Verificar se $S(r-1)/S(r) > \theta$, sendo θ um número inteiro positivo predeterminado, e, se essa desigualdade for afirmativa não aceitar a curva de grau r ajustada como satisfatória. Para valor adequadamente escolhido de θ, se a desigualdade acima é constatada, isto geralmente significa que, embora a curva ajustada tenha satisfeito à condição de convergência absoluta, a condição de convergência relativa está ainda longe de ser atingida. Está será provavelmente atingida para o grau $r + 1$. Ajustar então nova curva de grau $r + 1$ e considerá-la satisfatória, saltando a seguir ao próximo passo. No caso em que $S(r-1)/S(r) < \theta$, então aceitar a curva de grau r como satisfatória e dirigir-se ao próximo passo.

5. Calcular os resíduos $R_i = y_i - f(x_i, C_0, C_1, \ldots, C_m)$ para todo i inteiro pertencente ao intervalo [1,m] e não correspondentes a dados que foram rejeitados. Verificar então para todos os resíduos calculados se é válida a desigualdade:

$$| R_i | > \eta \tau_i \ . \qquad (17)$$

Sempre que para um dado valor de R_i esta desigualdade for verdadeira fazer $w_i = 0$ e decrementar m' de uma unidade. Se, uma vez que todos os resíduos tenham sido verificados, ocorrer que, para pelo menos um deles, a desigualdade foi verdadeira retornar ao passo 1 fazendo $r = 1$. Em caso de, nesta verificação, não se constatar nenhum ponto inválido, isto é, para nenhum resíduo a desigualdade se verificou, então chegou-se ao fim do processo.

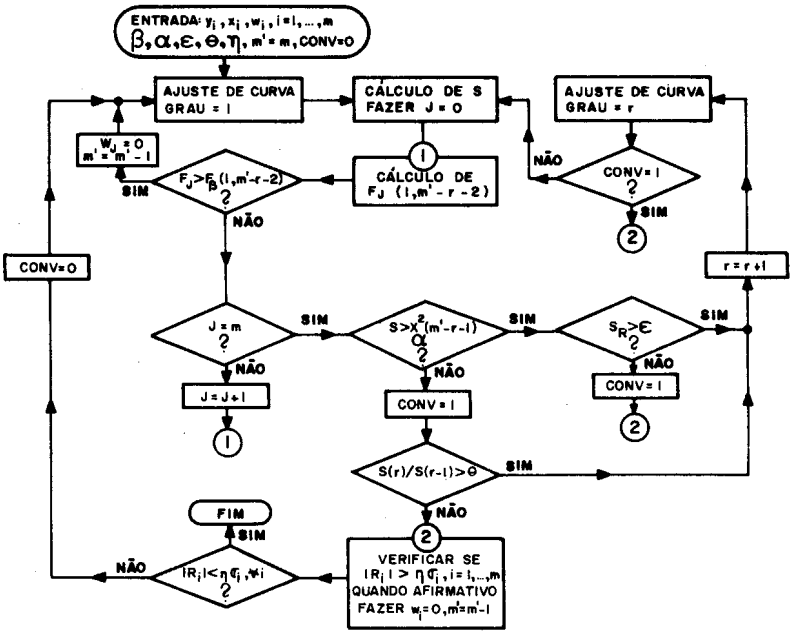


Fig. 1 - Diagrama de blocos do algoritmo de suavização de dados.

TESTE NUMÉRICO

Descrição do Teste e Apresentação de Resultados

O algoritmo foi testado através de sua aplicação e dados simulados de rastreamento de satélite artificial. Esses dados foram obtidos a partir da simulação da órbita do satélite europeu TD-1A, calculando valores da distância entre uma estação de rastreamento e o satélite, supondo uma taxa de amostragem de um dado por segundo. A cada valor calculado, adicionou-se um erro aleatório de média nula e desvio-padrão de 50 m, calculado a partir de uma rotina geradora de números aleatórios com distribuição normal. A simulação de dados inválidos foi feita através da adição de erros irrealisticamente elevados a alguns dos dados simulados.

Na rotina implementada para ajuste polinomial por mínimos quadrados, utilizaram-se para as funções $g_i(x)$, na equação 2, polinômios de Chebyshev que tornam a matriz A^TWA da equação 4 melhor condicionada quanto à facilidade de inversão. Isto se deve ao fato de os polinômios de Chebyshev variarem entre -1 e 1. Os polinômios de Chebyshev satisfazem a seguinte relação de recursão [2], [4]:

$$g_{k+1}(x) = 2xg_k(x) - g_{k-1}(x), \quad k = 1,2,\ldots \qquad (18)$$

com $g_0(x) = 1$ e $g_1(x) = x$.

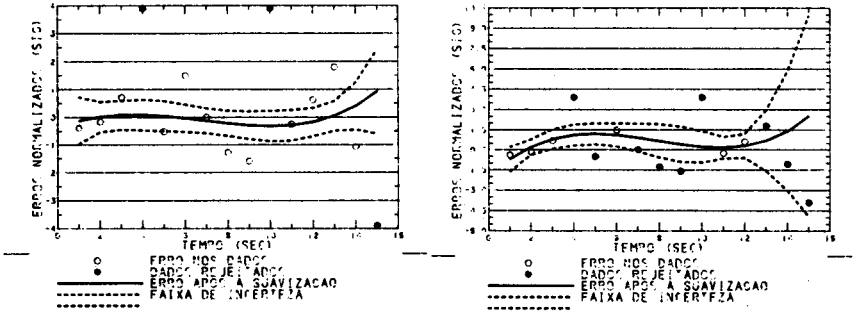
A mudança da variável independente t para a variável x, nos instantes de amostragem de dados t_i, i = 1,2,...,m, foi feita pela equação:

$$x_i = (2t_i - t_m - t_i)/(t_m - t_1). \qquad (19)$$

Para os fatores de tolerância definidos utilizaram-se os seguintes valores: $\alpha = 0{,}05$, $\xi = 0{,}1$, $\beta = 0{,}01$, $\theta = 2$, $\eta = 3$.

Nas figuras 2 e 3 são apresentados graficamente alguns dos resultados obtidos nos testes efetuados. Em cada gráfico são apresentados os valores do erro com o qual se contaminou cada dado simulado, a curva do erro real entre o polinômio suavizador e valores reais da distância estação de rastreamento-satélite simulada e a faixa de incerteza de um desvio-padrão do erro de suavização em torno da curva do erro real. São ainda plotados em destaque, em cada gráfico, os pontos que foram rejeitados durante o processo. Os valores plotados se encontram normalizados em unidades de desvio-padrão do erro aleatório adicionado aos dados simulados: $\tau_i = \tau = 50$ m, i = 1,2,...,m.

A figura 2.a apresenta os resultados da aplicação do algoritmo a quinze dados. Dentre esses quinze pontos existem três, aos quais foram adicionados erros irrealisticamente grandes, de tal modo a torná-los inválidos; y_4, y_{10} e y_{15}. A título de comparação, são apresentados na figura 2.b os resultados obtidos com a aplicação, aos mesmos pontos, de algoritmo de suavização empregando apenas a técnica usual de rejeição de dados com base na faixa $\eta\tau_i$ em torno da curva ajustada.



(a) (b)

Fig. 2 - Resultados do teste de aplicação dos algoritmos desenvolvido e usual a 15 dados: a) algoritmo desenvolvido e b) algoritmo usual.

Observe-se, na figura 2.a, que todos os três dados inválidos existentes no conjunto foram realmente detectados e rejeitados durante a aplicação do algoritmo desenvolvido. O mesmo não acontece com a aplicação do algoritmo que inclui apenas a técnica usual de rejeição de dados, conforme pode ser visto na figura 2.b, na qual está evidenciada a rejeição adicional de seis pontos válidos. Esse mau desempenho da técnica usual se deve a alta taxa de dados inválidos sobre dados válidos utilizada, 1:5. Esta técnica começa a apresentar problemas quando a citada taxa atinge valores da ordem de 1:10 [2]. Comparando a curva do erro real normalizado apresentada na figura 2.a, correspondente à aplicação do algoritmo de suavização desenvolvido, com a apresentada na figura 2, verifica-se que a qualidade do ajuste é bastante superior no primeiro caso, devendo-se isto ao fato de neste caso nenhum ponto válido ser rejeitado. Pode-se ainda ser verificado na figura 2 que a faixa de incerteza em torno da curva ajustada permanece menor que um desvio padrão do erro aleatório dos dados, τ, em praticamente todo o intervalo, alargando-se apenas nas extremidades. Esse alargamento é explicado pelo tratamento assimétrico dispensado aos dados por técnicas de ajuste de curvas. Em ambos os casos apresentados na figura 2 a curva final ajustada foi de grau 3. É importante ainda destacar que, no caso da aplicação do algoritmo desenvolvido, todos os três pontos inválidos foram detectados durante o processo de seleção automática do grau através da técnica de rejeição baseada na distribuição F.

A Figura 3 apresenta os resultados obtidos em outro teste, no qual foram utilizados trinta dados dentre os quais são inválidos os pontos y_4, y_{10}, y_{15}, y_{20}, y_{25} e y_{30}. Comparando os resultados obtidos com a aplicação do algoritmo desenvolvido (Figura 3.a) com os obtidos na aplicação do algoritmo que utiliza apenas a técnica usual de rejeição (3.b), verifica-se que, também neste teste, a qualidade do ajuste é melhor no primeiro caso, embora aqui a diferença não seja tão sensível quanto no teste anterior. Note-se pela figura 3b que a aplicação apenas da técnica usual de rejeição resultou na rejeição de um ponto válido, y_1, além dos seis pontos realmente inválidos. Com a aplicação do algoritmo desenvolvido, só os seis pontos realmente inválidos foram rejeitados, conforme se verifica na figura 3a. Isto explica a diferença de qualidade entre os ajustes, evidenciada pelas curvas do erro real. Ainda com relação ao presente teste, convém mencionar que, durante a aplicação do algoritmo desenvolvido, apenas três dos seis pontos inválidos existentes no conjunto foram detectados a priori pela técnica de rejeição baseada na distribuição F, ainda durante o processo de seleção automática do grau. Os outros três pontos foram detectados após uma primeira convergência do processo de seleção automática do grau, por caírem fora da região de 3τ em torno da curva ajustada. Verificou-se assim, que a técnica usual de rejeição, também utilizada no algorítmo em complemento da técnica baseada na distribuição F, obteve sucesso na rejeição dos três dados inválidos restantes devido à redução da taxa de dados inválidos sobre dados válidos promovida pela aplicação a priori da técnica de rejeição desenvolvida.

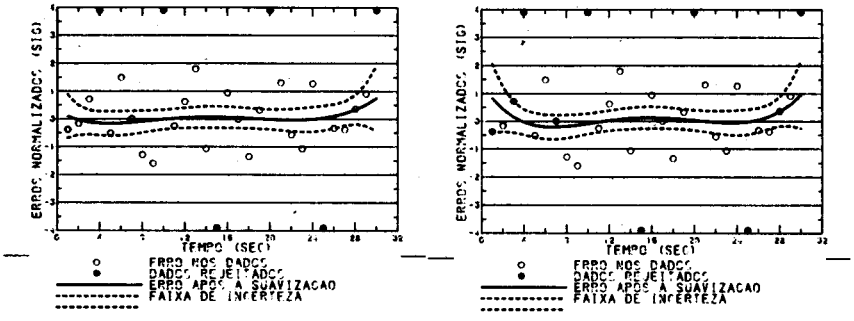


Fig. 3 - Resultados do teste de aplicação dos algoritmos desenvolvido e usual a 30 dados: a)algoritmo desenvolvido e b) algoritmo usual.

CONCLUSÕES

Pode-se concluir, com base nos testes efetuados,que o algoritmo apresentado permite que se trabalhe com taxas de dados inválidos sobre dados válidos maiores que as permitidas pelo emprego apenas da técnica usual de rejeição,com base na faixa de $\eta\tau$ em torno da curva ajustada. Também se verificou nos testes que, em casos de existência de dados com erros muito elevados, mesmo a baixa taxa de dados inválidos sobre válidos, o algoritmo desenvolvido apresenta melhor desempenho. Para esses casos, a técnica de rejeição usual pode rejeitar dados válidos, já que pontos com erros muito elevados podem causar desvios consideráveis da curva ajustada, em suas proximidades, fazendo com que dados válidos localizados nessas regiões possam cair fora da faixa de tolerância.

Em termos de tempo de processamento,o algoritmo desenvolvido apresenta um pequeno acréscimo em relação ao que emprega apenas a técnica usual de rejeição,mas, como ambos os procedimentos possuem a característica de processamento em lote,essa desvantagem não é significativa na maioria das aplicações.

REFERÊNCIAS

[1] Meyer, S.L., "Data analyses for scientistics and engineers", John Wilew & Sons Inc., New York, 1975.

[2] Wertz, J.R., "Spacecraft attitude determination and control, D. Reidel, London, 1978.

[3] Wine, R.L., "Statistics for scientists and engineers", Prentice-Hall Inc., Englewood Cliffs, N.J., 1964.

[4] Hamming, E.W. "Digital filters", Prentice-Hall Inc., Englewood Cliffs, N.J., 1977.