O índice de detecção $D(k)$

Formulação, propriedades e uso na detecção de falhas
a partir da saída medida $y(k)$ e da saída predita $\hat{y}(k)$

Heitor Gonçalves Santana Helena Bigliani Ferreira
Gabriel Rios de Souza

Nota técnica complementar ao artigo do projeto final de PCS5787 Ciência dos Dados e Big Data
Programa de Pós-Graduação em Engenharia Elétrica, Escola Politécnica da USP, 2026

Esta nota detalha o índice de detecção $D(k)$ empregado no artigo *Detecção de falhas no processo Tennessee Eastman por resíduos de modelos NARX/NARMAX*. Parte-se do par formado pela saída medida $y(k)$ e pela saída predita $\hat{y}(k)$ de cada subsistema, define-se o resíduo nas suas duas formas, um passo à frente e de simulação livre, e mostra-se como os resíduos dos cinco subsistemas são reunidos numa única estatística, a média móvel do quadrado da distância de Mahalanobis. As propriedades que justificam o uso do índice são enunciadas como lemas e proposições e demonstradas passo a passo, a saber: a relação exata entre os dois tipos de resíduo, a absorção de desvios persistentes pelo resíduo um passo à frente, a invariância a mudanças de escala, o valor esperado em operação normal, a decomposição do valor esperado em falha num termo de média e num termo de dispersão, o efeito da janela da média móvel e a ordem de grandeza do atraso de detecção. Todas as afirmações são confrontadas com os dados de Rieth et al. e com os modelos ARX do artigo, e a nota termina com o procedimento de uso, o código correspondente e as limitações do método.

Sumário

1 Contexto histórico **2**

2 Do par $y(k)$, $\hat{y}(k)$ ao resíduo **2**
2.1 O modelo e as duas predições 2
2.2 Relação exata entre os dois resíduos 3
2.3 Absorção de desvios persistentes 4
2.4 O vetor de resíduos 5

3 Formulação do índice $D(k)$ **6**
3.1 Estimação da referência em operação normal 6
3.2 Distância de Mahalanobis e média móvel 6
3.3 Limiar e regra de alarme 7

4 Propriedades do índice **7**
4.1 Normalização e correlação 7
4.2 Comportamento em operação normal 9
4.3 Como uma falha aparece no índice: média e dispersão 10
4.4 O papel da janela da média móvel 11
4.5 Ordem de grandeza do atraso 12

5 Uso do índice na detecção de falhas **13**
5.1 Procedimento 13
5.2 Implementação 15

5.3 Do alarme ao subsistema 15

6 Limitações e cuidados **16**

7 Observações finais **16**

1 Contexto histórico

A ideia de detectar falhas comparando o que um processo faz com o que um modelo diz que ele deveria fazer é conhecida como redundância analítica e remonta ao início da década de 1970. Beard [2] e Jones [12] propuseram filtros de detecção de falhas que geram sinais sensíveis a falhas específicas, e Mehra e Peschon [15] formularam a abordagem pelas inovações, segundo a qual, num sistema bem modelado, a sequência de erros de predição um passo à frente de um filtro de Kalman é um ruído branco de média nula e covariância conhecida, de modo que qualquer afastamento dessas propriedades denuncia uma mudança no sistema. O levantamento de Willsky [18] organizou esses trabalhos e consolidou a separação, usada até hoje, entre a *geração de resíduos*, que produz sinais próximos de zero em operação normal, e a *avaliação de resíduos*, que decide, a partir deles, se há falha. As décadas seguintes trouxeram as relações de paridade de Chow e Willsky [5], os observadores de Frank [7] e as sínteses de Basseville e Nikiforov [1], Gertler [8] e Isermann [11].

A avaliação de um vetor de resíduos tem raízes ainda mais antigas, na estatística multivariada. Mahalanobis [14] introduziu, em 1936, uma distância que leva em conta a covariância das variáveis, e Hotelling [9] empregou o seu quadrado, a estatística T^2, no controle de qualidade multivariado, aplicação em que ela se tornou o padrão do monitoramento estatístico de processos [4]. A suavização temporal da estatística, para tornar visíveis pequenas mudanças persistentes, aparece nos gráficos de controle de médias móveis e, na forma exponencial, no gráfico MEWMA de Lowry et al. [13].

O índice $D(k)$ do artigo combina essas duas tradições: os resíduos são gerados por modelos de entrada e saída identificados pelo algoritmo FROLS [3], e são avaliados pela média móvel da estatística T^2 de Hotelling, com limiar calibrado em dados de operação normal que não participaram da identificação.

2 Do par $y(k)$, $\hat{y}(k)$ ao resíduo

2.1 O modelo e as duas predições

Considera-se um subsistema com saída $y(k)$ e entradas $u_1(k), \ldots, u_p(k)$, todas padronizadas pela média e pelo desvio padrão das simulações normais de treino. O modelo ARX identificado tem a forma

$$y(k) = \sum_{i=1}^{n_y} a_i \, y(k-i) + \sum_{j=1}^{p} \sum_{l=0}^{n_u} b_{jl} \, u_j(k-l) + e(k), \tag{1}$$

em que $e(k)$ é a parcela que o modelo não explica, com $n_y = 3$ e $n_u = 2$ no artigo; os coeficientes não selecionados pelo FROLS são nulos. Com o operador de atraso q^{-1}, definido por $q^{-1}x(k) = x(k-1)$, e os polinômios

$$A(q) = 1 - \sum_{i=1}^{n_y} a_i \, q^{-i}, \qquad B_j(q) = \sum_{l=0}^{n_u} b_{jl} \, q^{-l}, \tag{2}$$

a equação (1) escreve-se de forma compacta como

$$A(q) \, y(k) = \sum_{j=1}^{p} B_j(q) \, u_j(k) + e(k). \tag{3}$$

O mesmo modelo produz duas predições distintas da saída, conforme se usem, nos termos autorregressivos, as saídas medidas ou as saídas que o próprio modelo calculou.

Definição 1 (Predição um passo à frente)**.** A predição um passo à frente (*one-step-ahead*, OSA) é

$$\hat{y}(k \,|\, k-1) = \sum_{i=1}^{n_y} a_i \, y(k-i) + \sum_{j=1}^{p} \sum_{l=0}^{n_u} b_{jl} \, u_j(k-l), \tag{4}$$

isto é, a melhor estimativa de $y(k)$ dadas as saídas *medidas* até $k-1$ e as entradas até k.

Definição 2 (Simulação livre)**.** A predição de simulação livre $\hat{y}_s(k)$ é obtida pela recursão

$$\hat{y}_s(k) = \sum_{i=1}^{n_y} a_i \, \hat{y}_s(k-i) + \sum_{j=1}^{p} \sum_{l=0}^{n_u} b_{jl} \, u_j(k-l), \qquad k > L, \tag{5}$$

iniciada com as saídas medidas, $\hat{y}_s(k) = y(k)$ para $k \leq L$, em que $L = 3$ é o maior atraso do modelo. A saída medida nunca mais entra na recursão: o modelo é alimentado só pelas entradas.

Definição 3 (Resíduos)**.** O resíduo um passo à frente e o resíduo de simulação livre são as diferenças entre a saída medida e as duas predições:

$$r_{\text{osa}}(k) = y(k) - \hat{y}(k \,|\, k-1), \qquad r_{\text{sim}}(k) = y(k) - \hat{y}_s(k). \tag{6}$$

No artigo, o resíduo um passo à frente é denotado por $e(k)$; aqui se usa $r_{\text{osa}}(k)$ para distingui-lo da parcela não explicada $e(k)$ da equação (1), com a qual coincide quando o modelo é avaliado nos próprios dados de identificação.

Substituindo-se (4) em (6) e usando-se (2), obtém-se uma forma que será usada adiante:

$$r_{\text{osa}}(k) = y(k) - \sum_{i=1}^{n_y} a_i \, y(k-i) - \sum_{j=1}^{p} B_j(q) \, u_j(k) = A(q) \, y(k) - \sum_{j=1}^{p} B_j(q) \, u_j(k). \tag{7}$$

Exemplo 1 (Subsistema S2, *stripper*)**.** Em S2, a saída é a temperatura do *stripper* (xmeas_18) e as entradas são a válvula de vapor do *stripper* (xmv_9, u_1) e a vazão de fundo do separador, que alimenta o *stripper* (xmeas_14, u_2). O FROLS selecionou cinco termos, e as duas predições são

$$\begin{aligned}
\hat{y}(k \,|\, k-1) &= 1{,}4200 \, y(k-1) - 0{,}8080 \, y(k-2) + 0{,}2798 \, y(k-3) \\
&\quad + 0{,}2915 \, u_1(k) - 0{,}1848 \, u_1(k-2), \\
\hat{y}_s(k) &= 1{,}4200 \, \hat{y}_s(k-1) - 0{,}8080 \, \hat{y}_s(k-2) + 0{,}2798 \, \hat{y}_s(k-3) \\
&\quad + 0{,}2915 \, u_1(k) - 0{,}1848 \, u_1(k-2).
\end{aligned}$$

As duas expressões têm os mesmos coeficientes e diferem apenas no que alimenta os termos autorregressivos. Nas simulações normais de validação, o desvio padrão de r_{osa} é 0,087 e o de r_{sim} é 0,233, ambos em unidades do desvio padrão da saída.

2.2 Relação exata entre os dois resíduos

Os dois resíduos parecem medir coisas diferentes, mas, para o modelo ARX, um é uma versão filtrada do outro.

Proposição 1 (Resíduo de simulação livre como resíduo um passo à frente filtrado)**.** *Para o modelo ARX* (1)*, com a simulação livre iniciada pelas saídas medidas, o resíduo de simulação*

livre é a saída do filtro $1/A(q)$ excitado pelo resíduo um passo à frente, com condições iniciais nulas:

$$r_{\mathrm{sim}}(k) = \sum_{i=1}^{n_y} a_i\, r_{\mathrm{sim}}(k-i) + r_{\mathrm{osa}}(k), \quad k > L, \qquad r_{\mathrm{sim}}(k) = 0, \quad k \le L, \tag{8}$$

ou, na notação de operadores, $A(q)\, r_{\mathrm{sim}}(k) = r_{\mathrm{osa}}(k)$, isto é, $r_{\mathrm{sim}}(k) = A(q)^{-1} r_{\mathrm{osa}}(k)$.

Demonstração. Para $k \le L$, a Definição 2 dá $\hat{y}_s(k) = y(k)$, logo $r_{\mathrm{sim}}(k) = 0$. Para $k > L$, a recursão (5) pode ser reescrita, passando os termos autorregressivos para o lado esquerdo, como

$$\hat{y}_s(k) - \sum_{i=1}^{n_y} a_i\, \hat{y}_s(k-i) = \sum_{j=1}^{p} B_j(q)\, u_j(k), \quad \text{isto é,} \quad A(q)\, \hat{y}_s(k) = \sum_{j=1}^{p} B_j(q)\, u_j(k).$$

Aplicando-se $A(q)$, que é linear, à definição $r_{\mathrm{sim}}(k) = y(k) - \hat{y}_s(k)$, vem

$$A(q)\, r_{\mathrm{sim}}(k) = A(q)\, y(k) - A(q)\, \hat{y}_s(k) = A(q)\, y(k) - \sum_{j=1}^{p} B_j(q)\, u_j(k).$$

O lado direito é exatamente a expressão (7) de $r_{\mathrm{osa}}(k)$, donde $A(q)\, r_{\mathrm{sim}}(k) = r_{\mathrm{osa}}(k)$. Expandindo-se $A(q)$ pela definição (2), tem-se $r_{\mathrm{sim}}(k) - \sum_i a_i\, r_{\mathrm{sim}}(k-i) = r_{\mathrm{osa}}(k)$, que é (8). A recursão começa em $k = L+1$ e usa os valores $r_{\mathrm{sim}}(k-i)$ com $k - i \ge 1$, que são nulos para $k - i \le L$; as condições iniciais são, portanto, nulas. ■

Observação 1. A Proposição 1 foi verificada nos dados: filtrando-se os resíduos um passo à frente das 75 simulações normais de validação e das simulações com as falhas 1 e 19 por $1/A(q)$ de cada subsistema, recuperam-se os resíduos de simulação livre com diferença máxima de 10^{-6}, atribuível ao armazenamento em precisão simples. A relação vale exatamente para o ARX, que é o modelo usado na detecção; para o NARX e o NARMAX, em que a saída entra em produtos com outros regressores, ela não vale em geral.

Corolário 1 (Ganho estático)**.** *Se o modelo é estável, isto é, se as raízes de $z^{n_y} - a_1 z^{n_y-1} - \cdots - a_{n_y}$ têm módulo menor que 1, e se $r_{\mathrm{osa}}(k)$ tende a uma constante c, então $r_{\mathrm{sim}}(k)$ tende a $c/A(1)$, em que $A(1) = 1 - \sum_i a_i$.*

Demonstração. Seja $r_{\mathrm{osa}}(k) = c$ para k grande. Procura-se uma solução constante $r_{\mathrm{sim}}(k) = \bar{r}$ de (8): substituindo-se, vem $\bar{r} = \sum_i a_i\, \bar{r} + c$, isto é, $\bar{r}\,(1 - \sum_i a_i) = c$, ou $\bar{r} = c/A(1)$. Como o filtro é estável, a diferença entre $r_{\mathrm{sim}}(k)$ e essa solução satisfaz a recursão homogênea $x(k) = \sum_i a_i\, x(k-i)$, cujas soluções decaem a zero, e $r_{\mathrm{sim}}(k) \to c/A(1)$. ■

2.3 Absorção de desvios persistentes

O Corolário 1 tem uma leitura inversa, que explica por que o artigo concatena os dois resíduos.

Proposição 2 (O resíduo um passo à frente absorve desvios persistentes)**.** *Suponha-se que, a partir do instante k_0, a saída medida passe a ter um desvio constante b em relação ao que o modelo explica, isto é, $y(k) = y_0(k) + b$ para $k \ge k_0$, em que $y_0(k)$ é a saída que seria medida sem a falha, e que as entradas não mudem. Então, para $k \ge k_0 + n_y$,*

$$r_{\mathrm{osa}}(k) = r^0_{\mathrm{osa}}(k) + A(1)\, b, \qquad r_{\mathrm{sim}}(k) = r^0_{\mathrm{sim}}(k) + b, \tag{9}$$

em que r^0 denota os resíduos sem a falha. O resíduo um passo à frente acusa apenas a fração $A(1)$ do desvio; o de simulação livre, o desvio inteiro.

Demonstração. Pela expressão (7), $r_{\text{osa}}(k) = y(k) - \sum_i a_i\, y(k-i) - \sum_j B_j(q)u_j(k)$. Para $k \geq k_0 + n_y$, todos os instantes k,$k-1$,...,$k-n_y$ são posteriores a k_0, de modo que $y(k) = y_0(k) + b$ e $y(k-i) = y_0(k-i) + b$. Substituindo-se,

$$r_{\text{osa}}(k) = \Big[y_0(k) - \sum_i a_i\, y_0(k-i) - \sum_j B_j(q)u_j(k)\Big] + b - \sum_i a_i\, b = r^0_{\text{osa}}(k) + b\Big(1 - \sum_i a_i\Big),$$

que é a primeira igualdade, pois o colchete é $r^0_{\text{osa}}(k)$. Quanto à simulação livre, a recursão (5) não usa as saídas medidas depois de $k = L$, e, como as entradas não mudam, $\hat{y}_s(k)$ é o mesmo com ou sem falha; logo $r_{\text{sim}}(k) = y_0(k) + b - \hat{y}_s(k) = r^0_{\text{sim}}(k) + b$. ■

A Tabela 1 mostra que, em quatro dos cinco subsistemas, $A(1)$ é pequeno, pois os modelos têm polos próximos de 1, isto é, dinâmicas lentas em relação ao período de amostragem de 3 min. Em S5, por exemplo, o resíduo um passo à frente acusa só 6,7% de um desvio persistente, ao passo que o de simulação livre o acusa por inteiro. Em contrapartida, pelo Corolário 1, o filtro $1/A(q)$ amplifica as baixas frequências do ruído normal, o que explica por que r_{sim} tem desvio padrão de duas a quatro vezes maior que r_{osa} em operação normal e é fortemente autocorrelacionado. Os dois resíduos são, portanto, complementares: o um passo à frente reage depressa a mudanças e tem pouco ruído, mas esquece os desvios que persistem; o de simulação livre acumula os desvios persistentes, ao custo de mais ruído.

Tabela 1: Ganho estático $A(1) = 1 - \sum_i a_i$ dos modelos ARX, maior módulo dos polos e desvios padrão dos dois resíduos nas simulações normais de validação

Subsistema	$A(1)$	$1/A(1)$	Maior polo	$\sigma(r_{\text{osa}})$	$\sigma(r_{\text{sim}})$	$\sigma(r_{\text{sim}})/\sigma(r_{\text{osa}})$
S1 refrigeração do reator	0,627	1,60	0,691	0,694	0,763	1,1
S2 *stripper*	0,108	9,24	0,858	0,087	0,233	2,7
S3 pressão do reator	0,089	11,26	0,896	0,209	0,559	2,7
S4 separador	0,161	6,23	0,827	0,365	0,758	2,1
S5 compressor	0,067	14,84	0,941	0,196	0,625	3,2

A Figura 1 ilustra o fenômeno na falha 1, um degrau na razão de alimentação entre A e C, para o subsistema S3. Nas primeiras horas após a falha, a pressão medida sobe muito; a predição um passo à frente acompanha a medição, porque é realimentada por ela, e seu resíduo pouco se afasta de zero, enquanto a simulação livre, que só conhece as entradas, não reproduz a subida, e seu resíduo chega a 16 desvios padrão. É por esse mecanismo que, no artigo, o resíduo um passo à frente isolado detecta 62,1% das amostras da falha 1, ao passo que o de simulação livre detecta 98,1%.

2.4 O vetor de resíduos

Os resíduos dos cinco subsistemas são empilhados num único vetor. No detector principal do artigo, com os resíduos um passo à frente e de simulação livre dos modelos ARX, o vetor tem $m = 10$ componentes,

$$\mathbf{r}(k) = \big[\, r_{\text{osa},1}(k) \;\cdots\; r_{\text{osa},5}(k) \;\; r_{\text{sim},1}(k) \;\cdots\; r_{\text{sim},5}(k) \,\big]^{\mathsf{T}} \in \mathbb{R}^m, \tag{10}$$

em que o índice $i = 1,\ldots,5$ identifica o subsistema; os detectores que usam apenas um tipo de resíduo têm $m = 5$. Cada componente é calculada a partir da saída medida $y_i(k)$ e da saída predita correspondente, de modo que $\mathbf{r}(k)$ resume, num instante, o quanto cada subsistema se afasta do comportamento previsto pelo seu modelo.

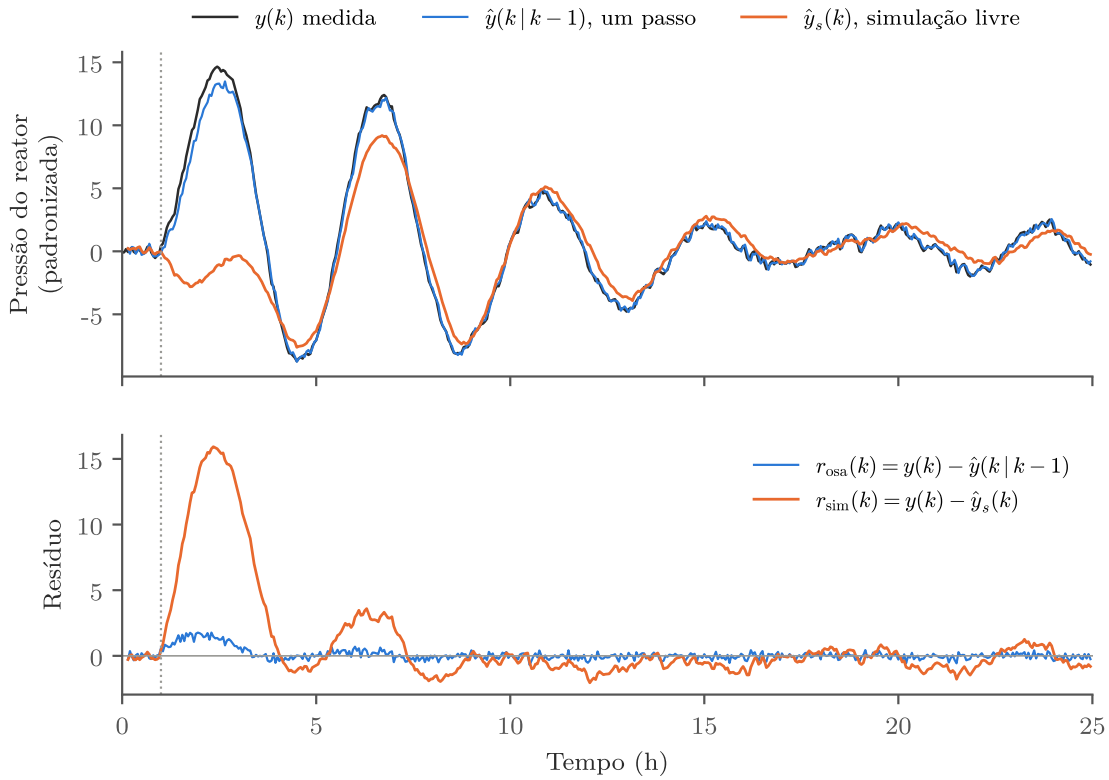


Figura 1: Subsistema S3 na simulação de teste 426 com a falha 1, iniciada em 1 h (linha pontilhada). Acima, a saída medida e as duas predições do mesmo modelo ARX; abaixo, os resíduos correspondentes.

3 Formulação do índice $D(k)$

3.1 Estimação da referência em operação normal

O índice compara $\mathbf{r}(k)$ com o comportamento do vetor de resíduos em operação normal, descrito pelo vetor de médias $\boldsymbol{\mu}$ e pela matriz de covariância $\mathbf{S}$. Ambos são estimados nas $R_v = 75$ simulações normais de validação (simulações 351 a 425), que não participam da identificação dos modelos nem do teste. Sendo $\mathbf{r}_j(k)$ o vetor de resíduos da simulação j no instante k, e $\mathcal{K}$ o conjunto dos instantes em que todos os resíduos estão definidos ($k > L$), com $N = R_v\,|\mathcal{K}|$ amostras ao todo, os estimadores são

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{j=1}^{R_v}\sum_{k\in\mathcal{K}} \mathbf{r}_j(k), \qquad \hat{\mathbf{S}} = \frac{1}{N-1}\sum_{j=1}^{R_v}\sum_{k\in\mathcal{K}} \big[\mathbf{r}_j(k) - \hat{\boldsymbol{\mu}}\big]\big[\mathbf{r}_j(k) - \hat{\boldsymbol{\mu}}\big]^{\mathsf{T}}. \tag{11}$$

Nos dados do artigo, $N = 75 \times 497 = 37\,275$. As médias estimadas são praticamente nulas (a maior, em módulo, é 0,009), como se espera de modelos identificados por mínimos quadrados com termos de entrada que absorvem o nível; a diagonal de $\hat{\mathbf{S}}$ contém os quadrados dos desvios padrão da Tabela 1; e os elementos fora da diagonal são relevantes, sobretudo entre os dois resíduos de um mesmo subsistema, com correlação de 0,91 em S1 e de 0,38 a 0,54 nos demais. O número de condição de $\hat{\mathbf{S}}$ é 224. No restante do texto, $\boldsymbol{\mu}$ e $\mathbf{S}$ denotam essas estimativas.

3.2 Distância de Mahalanobis e média móvel

Definição 4 (Índice de detecção)**.** O quadrado da distância de Mahalanobis do vetor de resíduos à referência normal é

$$d^2(k) = \big[\mathbf{r}(k) - \boldsymbol{\mu}\big]^{\mathsf{T}}\mathbf{S}^{-1}\big[\mathbf{r}(k) - \boldsymbol{\mu}\big], \tag{12}$$

e o índice de detecção é a sua média móvel causal numa janela de W amostras,

$$D(k) = \frac{1}{W} \sum_{j=0}^{W-1} d^2(k-j) = \frac{1}{W} \sum_{j=0}^{W-1} \big[\mathbf{r}(k-j) - \boldsymbol{\mu}\big]^{\mathsf{T}} \mathbf{S}^{-1} \big[\mathbf{r}(k-j) - \boldsymbol{\mu}\big]. \tag{13}$$

No artigo, $W = 10$ amostras (30 min). Quando $\mathbf{S}$ é singular ou mal condicionada, $\mathbf{S}^{-1}$ é substituída pela pseudoinversa de Moore e Penrose [16], $\mathbf{S}^{+}$.

A expressão (13) é exatamente a do artigo. Cada parcela tem um papel: a subtração de $\boldsymbol{\mu}$ centra os resíduos; a matriz $\mathbf{S}^{-1}$ normaliza as escalas e desconta as correlações (Lemas 1 e 2); a forma quadrática reduz o vetor a um número não negativo, que cresce com qualquer afastamento, em qualquer direção (Lema 4); e a média móvel reduz a variância do índice, o que torna visíveis mudanças pequenas mas persistentes (Proposição 3). O índice depende apenas de $y_i(k)$ e das predições $\hat{y}_i$: para calculá-lo em tempo real basta, a cada amostra, medir as saídas, atualizar as duas predições de cada subsistema, formar $\mathbf{r}(k)$ e atualizar a média móvel.

3.3 Limiar e regra de alarme

O limiar τ é o quantil de ordem $q = 0{,}99$ dos valores de $D(k)$ nas próprias simulações normais de validação. Sendo $D_{(1)} \le \cdots \le D_{(n)}$ esses valores ordenados, usa-se o quantil com interpolação linear (tipo 7 de Hyndman e Fan [10], o padrão do NumPy e do MATLAB):

$$h = (n-1)\,q + 1, \qquad \tau = D_{(\lfloor h \rfloor)} + \big(h - \lfloor h \rfloor\big)\big[D_{(\lceil h \rceil)} - D_{(\lfloor h \rfloor)}\big]. \tag{14}$$

A regra de alarme é

$$\text{alarme em } k \iff D(k) > \tau, \tag{15}$$

e, por construção, 1% das amostras normais de validação excedem τ, o que fixa a taxa nominal de falsos alarmes (FAR) em 1%. Nos dados do artigo, para o vetor de dez resíduos ARX, $\tau = 23{,}58$ com $W = 1$ e $\tau = 18{,}76$ com $W = 10$. A taxa de detecção (FDR) é a fração das amostras posteriores ao início da falha em que há alarme, e o atraso de detecção, no artigo, é o tempo até o primeiro de três alarmes consecutivos.

Observação 2 (Por que o limiar é empírico)*.* Se os resíduos fossem gaussianos e independentes no tempo, os limiares poderiam ser obtidos da distribuição qui-quadrado (Lemas 3 e Proposição 3): $\chi^2_{10;\,0{,}99} = 23{,}21$ para $W = 1$ e $\chi^2_{100;\,0{,}99}/10 = 13{,}58$ para $W = 10$. O primeiro valor é próximo do empírico, 23,58, mas o segundo é bem menor que 18,76. A razão é a autocorrelação dos resíduos, sobretudo dos de simulação livre (Proposição 1): as W parcelas da média móvel não são independentes, a variância de $D(k)$ diminui menos do que o previsto, e um limiar teórico produziria muito mais falsos alarmes que o nominal. O quantil empírico (14) incorpora automaticamente a distribuição real, inclusive a autocorrelação, as caudas não gaussianas e as imperfeições dos modelos.

4 Propriedades do índice

Nesta seção, $\mathbf{r}$ denota um vetor aleatório de resíduos com média $\boldsymbol{\mu}$ e covariância $\mathbf{S}$ em operação normal, e $\mathbf{S}$ é suposta definida positiva, exceto onde se indica o uso da pseudoinversa.

4.1 Normalização e correlação

Lema 1 (Caso de resíduos não correlacionados)**.** *Se* $\mathbf{S} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2)$*, então*

$$d^2(k) = \sum_{l=1}^{m} \left(\frac{r_l(k) - \mu_l}{\sigma_l} \right)^2, \tag{16}$$

a soma dos quadrados dos escores padronizados dos resíduos.

Demonstração. A inversa de uma matriz diagonal é a diagonal dos inversos, $\mathbf{S}^{-1} = \operatorname{diag}(1/\sigma_1^2, \ldots, 1/\sigma_m^2)$. Escrevendo-se $\mathbf{z} = \mathbf{r}(k) - \boldsymbol{\mu}$, a forma quadrática fica $\mathbf{z}^\mathsf{T}\mathbf{S}^{-1}\mathbf{z} = \sum_l \sum_n z_l \, [\mathbf{S}^{-1}]_{ln} \, z_n$, e, como só os termos com $l = n$ são não nulos, $\mathbf{z}^\mathsf{T}\mathbf{S}^{-1}\mathbf{z} = \sum_l z_l^2/\sigma_l^2$, que é (16). ∎

O Lema 1 mostra a primeira função de $\mathbf{S}^{-1}$: dividir cada resíduo pelo seu desvio padrão, para que um resíduo de $\sigma = 0{,}76$, como o de simulação livre de S1, e um de $\sigma = 0{,}087$, como o um passo à frente de S2, pesem igualmente. Sem essa normalização, o índice seria dominado pelos resíduos de maior variância, e uma falha que só alterasse S2 passaria despercebida. A segunda função aparece quando os resíduos são correlacionados.

Exemplo 2 (Dois resíduos correlacionados)**.** Sejam dois resíduos padronizados ($\mu_l = 0$, $\sigma_l = 1$) com correlação ρ, $|\rho| < 1$. Então

$$\mathbf{S} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \qquad \mathbf{S}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix},$$

como se verifica pelo produto $\mathbf{S}\,\mathbf{S}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1-\rho^2 & -\rho+\rho \\ \rho-\rho & -\rho^2+1 \end{bmatrix} = \mathbf{I}$. Para $\mathbf{r} = [z_1 \; z_2]^\mathsf{T}$,

$$d^2 = \frac{1}{1-\rho^2} \begin{bmatrix} z_1 & z_2 \end{bmatrix} \begin{bmatrix} z_1 - \rho z_2 \\ -\rho z_1 + z_2 \end{bmatrix} = \frac{z_1^2 - 2\rho\, z_1 z_2 + z_2^2}{1-\rho^2}. \tag{17}$$

Dois casos esclarecem o efeito de $\rho > 0$:

- *Desvios concordantes*, $z_1 = z_2 = z$: o numerador vale $z^2 - 2\rho z^2 + z^2 = 2z^2(1-\rho)$, e, como $1-\rho^2 = (1-\rho)(1+\rho)$, tem-se $d^2 = 2z^2/(1+\rho)$. A soma dos quadrados seria $2z^2$; quando $\rho \to 1$, $d^2 \to z^2$. Dois resíduos que andam juntos carregam essencialmente a mesma informação, e a distância de Mahalanobis os conta uma vez só.
- *Desvios discordantes*, $z_1 = -z_2 = z$: o numerador vale $2z^2(1+\rho)$ e $d^2 = 2z^2/(1-\rho)$, que cresce sem limite quando $\rho \to 1$. Se dois resíduos normalmente andam juntos, vê-los em sentidos opostos é muito improvável em operação normal, e o índice acusa isso fortemente.

A Figura 2 mostra a situação real de S1, em que os resíduos um passo à frente e de simulação livre têm correlação 0,91. A região de operação normal é uma elipse alongada na diagonal, e não um círculo. O ponto P está a uma distância euclidiana pequena da origem, com $z_1^2 + z_2^2 = 2{,}65$, dentro do círculo que contém 99% dos dados normais, mas fora da elipse, com $d^2 = 56$, porque contraria a correlação normal entre os dois resíduos. Um índice que somasse os quadrados dos resíduos não o detectaria; a distância de Mahalanobis o detecta.

Lema 2 (Invariância a transformações afins)**.** *Seja $\tilde{\mathbf{r}} = \mathbf{M}\mathbf{r} + \mathbf{c}$, com $\mathbf{M}$ invertível e $\mathbf{c}$ constante. Se $\tilde{\boldsymbol{\mu}}$ e $\tilde{\mathbf{S}}$ são a média e a covariância de $\tilde{\mathbf{r}}$, então $\tilde{d}^2(k) = d^2(k)$ para todo k, e o índice $D(k)$, o limiar e os alarmes não se alteram.*

Demonstração. Pela linearidade da média, $\tilde{\boldsymbol{\mu}} = \mathbb{E}[\mathbf{M}\mathbf{r} + \mathbf{c}] = \mathbf{M}\boldsymbol{\mu} + \mathbf{c}$, logo $\tilde{\mathbf{r}} - \tilde{\boldsymbol{\mu}} = \mathbf{M}(\mathbf{r} - \boldsymbol{\mu})$. A covariância é $\tilde{\mathbf{S}} = \mathbb{E}[\mathbf{M}(\mathbf{r}-\boldsymbol{\mu})(\mathbf{r}-\boldsymbol{\mu})^\mathsf{T}\mathbf{M}^\mathsf{T}] = \mathbf{M}\mathbf{S}\mathbf{M}^\mathsf{T}$, e, pela regra da inversa de um produto de matrizes invertíveis, $\tilde{\mathbf{S}}^{-1} = (\mathbf{M}^\mathsf{T})^{-1}\mathbf{S}^{-1}\mathbf{M}^{-1}$. Então

$$\tilde{d}^2 = (\mathbf{r}-\boldsymbol{\mu})^\mathsf{T}\mathbf{M}^\mathsf{T}\,(\mathbf{M}^\mathsf{T})^{-1}\mathbf{S}^{-1}\mathbf{M}^{-1}\,\mathbf{M}(\mathbf{r}-\boldsymbol{\mu}) = (\mathbf{r}-\boldsymbol{\mu})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{r}-\boldsymbol{\mu}) = d^2.$$

Como $D(k)$ é a média de valores de d^2 e o limiar é um quantil dos mesmos valores, ambos ficam inalterados, e também os alarmes. O argumento vale igualmente para as estimativas (11), pois a média e a covariância amostrais se transformam da mesma maneira. ∎

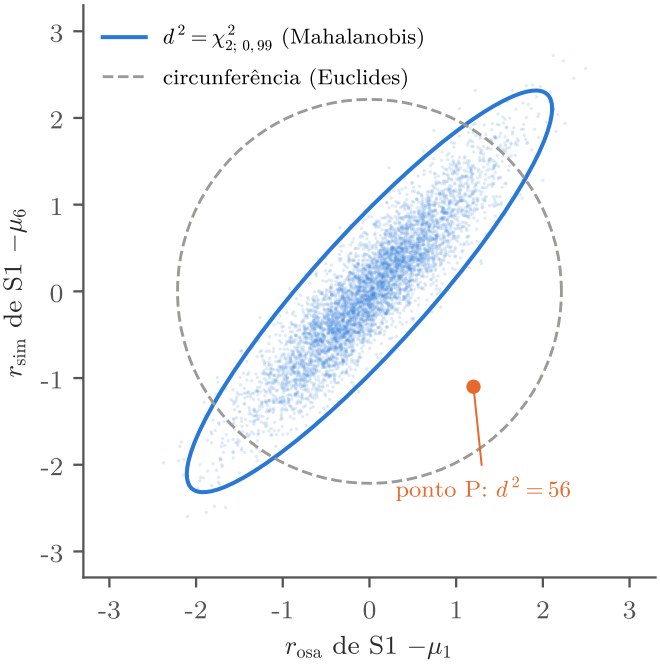


Figura 2: Resíduos um passo à frente e de simulação livre de S1 nas simulações normais de validação (pontos), a elipse $d^2 = \chi^2_{2;\,0,99}$ da distância de Mahalanobis e a circunferência euclidiana de mesma área relativa. O ponto P está dentro da circunferência e fora da elipse.

O Lema 2 garante que o índice não depende de escolhas arbitrárias de unidades ou de escala: tanto faz medir a temperatura em graus Celsius ou em kelvin, padronizar ou não as variáveis, ou combinar os resíduos linearmente. Em particular, como $\mathbf{M} = \text{diag}(1/\sigma_l)$ é invertível, o Lema 1 é um caso particular: a normalização pelos desvios padrão já está embutida em $\mathbf{S}^{-1}$.

Observação 3 (Pseudoinversa). Quando um resíduo é combinação linear exata de outros, $\mathbf{S}$ é singular e não tem inversa. A pseudoinversa $\mathbf{S}^{+}$ [16] restringe a forma quadrática ao subespaço em que os resíduos efetivamente variam, o que equivale a descartar as combinações redundantes, e coincide com $\mathbf{S}^{-1}$ quando esta existe. Com número de condição 224, a matriz do artigo é invertível, e a pseudoinversa é apenas uma salvaguarda numérica.

4.2 Comportamento em operação normal

Lema 3 (Valor esperado em operação normal)**.** *Se* $\mathbf{r}(k)$ *tem média* $\boldsymbol{\mu}$ *e covariância* $\mathbf{S}$*, então* $\mathbb{E}[d^2(k)] = m$ *e* $\mathbb{E}[D(k)] = m$*. Se, além disso,* $\mathbf{r}(k)$ *é gaussiano,* $d^2(k)$ *segue a distribuição qui-quadrado com* m *graus de liberdade, de variância* $2m$*.*

Demonstração. Um escalar é igual ao seu traço, e o traço é invariante a permutações cíclicas, $\text{tr}(\mathbf{XY}) = \text{tr}(\mathbf{YX})$. Com $\mathbf{z} = \mathbf{r}(k) - \boldsymbol{\mu}$,

$$d^2 = \mathbf{z}^{\mathsf{T}}\mathbf{S}^{-1}\mathbf{z} = \text{tr}\left(\mathbf{z}^{\mathsf{T}}\mathbf{S}^{-1}\mathbf{z}\right) = \text{tr}\left(\mathbf{S}^{-1}\mathbf{z}\mathbf{z}^{\mathsf{T}}\right).$$

Como o traço e a esperança são lineares, podem ser trocados de ordem:

$$\mathbb{E}[d^2] = \text{tr}\left(\mathbf{S}^{-1}\,\mathbb{E}[\mathbf{z}\mathbf{z}^{\mathsf{T}}]\right) = \text{tr}\left(\mathbf{S}^{-1}\mathbf{S}\right) = \text{tr}(\mathbf{I}_m) = m.$$

Pela linearidade da esperança, $\mathbb{E}[D(k)] = \frac{1}{W}\sum_{j=0}^{W-1}\mathbb{E}[d^2(k-j)] = \frac{1}{W}\,W\,m = m$. No caso gaussiano, seja $\mathbf{S} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$ a decomposição de Cholesky e $\mathbf{w} = \mathbf{L}^{-1}\mathbf{z}$. Então $\mathbf{w}$ é gaussiano de média nula e covariância $\mathbf{L}^{-1}\mathbf{S}(\mathbf{L}^{-1})^{\mathsf{T}} = \mathbf{L}^{-1}\mathbf{L}\mathbf{L}^{\mathsf{T}}(\mathbf{L}^{\mathsf{T}})^{-1} = \mathbf{I}_m$, isto é, suas m componentes são normais padrão independentes, e $d^2 = \mathbf{z}^{\mathsf{T}}(\mathbf{L}\mathbf{L}^{\mathsf{T}})^{-1}\mathbf{z} = \mathbf{w}^{\mathsf{T}}\mathbf{w} = \sum_l w_l^2$, que é, por definição, uma variável qui-quadrado com m graus de liberdade, de variância $2m$. ■

Nos dados de validação, a média de d^2 é 9,9997, praticamente igual a $m = 10$, como o Lema 3 prevê (a pequena diferença vem do fator $N-1$ em (11)); a variância é 21,0, próxima de $2m = 20$, o que indica que a hipótese gaussiana é razoável para d^2 em cada instante. O valor m é, assim, a referência natural do índice: em operação normal, $D(k)$ flutua em torno do número de resíduos.

4.3 Como uma falha aparece no índice: média e dispersão

Lema 4 (Decomposição do valor esperado em falha)**.** *Suponha-se que, durante uma falha, o vetor de resíduos passe a ter média $\boldsymbol{\mu} + \boldsymbol{\delta}$ e covariância $\mathbf{S}_f$, enquanto o índice continua a usar a referência normal $\boldsymbol{\mu}$, $\mathbf{S}$. Então*

$$\mathbb{E}_f\left[d^2(k)\right] = \underbrace{\operatorname{tr}\left(\mathbf{S}^{-1}\mathbf{S}_f\right)}_{\textit{dispersão}} + \underbrace{\boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}^{-1}\boldsymbol{\delta}}_{\textit{média}}. \tag{18}$$

Demonstração. Seja $\mathbf{z} = \mathbf{r}(k) - \boldsymbol{\mu}$. Em falha, $\mathbb{E}_f[\mathbf{z}] = \boldsymbol{\delta}$, e, somando-se e subtraindo-se $\boldsymbol{\delta}$, $\mathbf{z} = (\mathbf{z} - \boldsymbol{\delta}) + \boldsymbol{\delta}$, em que $\mathbf{z} - \boldsymbol{\delta}$ tem média nula e covariância $\mathbf{S}_f$. Então

$$\mathbb{E}_f\left[\mathbf{z}\mathbf{z}^{\mathsf{T}}\right] = \mathbb{E}_f\left[(\mathbf{z}-\boldsymbol{\delta})(\mathbf{z}-\boldsymbol{\delta})^{\mathsf{T}}\right] + \mathbb{E}_f\left[\mathbf{z}-\boldsymbol{\delta}\right]\boldsymbol{\delta}^{\mathsf{T}} + \boldsymbol{\delta}\,\mathbb{E}_f\left[\mathbf{z}-\boldsymbol{\delta}\right]^{\mathsf{T}} + \boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}} = \mathbf{S}_f + \boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}},$$

pois os dois termos cruzados se anulam. Usando-se o traço como na prova do Lema 3,

$$\mathbb{E}_f[d^2] = \operatorname{tr}\left(\mathbf{S}^{-1}\mathbb{E}_f[\mathbf{z}\mathbf{z}^{\mathsf{T}}]\right) = \operatorname{tr}\left(\mathbf{S}^{-1}\mathbf{S}_f\right) + \operatorname{tr}\left(\mathbf{S}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}\right) = \operatorname{tr}\left(\mathbf{S}^{-1}\mathbf{S}_f\right) + \boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}^{-1}\boldsymbol{\delta},$$

em que, no último passo, se usou de novo a permutação cíclica, $\operatorname{tr}(\mathbf{S}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}) = \operatorname{tr}(\boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}^{-1}\boldsymbol{\delta}) = \boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}^{-1}\boldsymbol{\delta}$. ■

Corolário 2. *(a) Sem falha, $\boldsymbol{\delta} = \mathbf{0}$ e $\mathbf{S}_f = \mathbf{S}$, e* (18) *recai em $\operatorname{tr}(\mathbf{I}) = m$, como no Lema 3.*

(b) Uma falha que só desloca a média ($\mathbf{S}_f = \mathbf{S}$) eleva $\mathbb{E}[d^2]$ de m para $m+\lambda$, com $\lambda = \boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}^{-1}\boldsymbol{\delta} \geq 0$, o quadrado da distância de Mahalanobis do deslocamento.

(c) Uma falha que só multiplica as variâncias por κ ($\boldsymbol{\delta} = \mathbf{0}$, $\mathbf{S}_f = \kappa\mathbf{S}$) eleva $\mathbb{E}[d^2]$ de m para κm.

(d) Uma falha que não altera nem a média nem a covariância dos resíduos deixa $\mathbb{E}[d^2] = m$ e é indistinguível da operação normal por esse índice, qualquer que seja o limiar.

Demonstração. Os itens (a) a (c) seguem por substituição direta em (18), com $\operatorname{tr}(\mathbf{S}^{-1}\kappa\mathbf{S}) = \kappa\,\operatorname{tr}(\mathbf{I}_m) = \kappa m$ em (c). Em (b), $\lambda \geq 0$ porque $\mathbf{S}^{-1}$ é definida positiva. O item (d) é o caso (a) aplicado aos dados de falha: se a distribuição dos resíduos não muda, a de d^2 e a de D também não, e a fração de alarmes permanece igual à fração de falsos alarmes. ■

O Lema 4 responde à pergunta de como um único índice detecta tanto falhas que deslocam a média quanto falhas que alteram a dispersão: os dois efeitos se somam no valor esperado de d^2. A Figura 3 aplica a decomposição aos dados, calculando $\boldsymbol{\delta}$ e $\mathbf{S}_f$ com todas as amostras posteriores ao início de cada falha nas 75 simulações de teste; a soma das duas parcelas coincide, em todas as falhas, com a média empírica de d^2, como (18) exige. Três grupos se destacam:

- *Falhas 3, 9 e 15:* as duas parcelas somam 10,2, 10,3 e 10,9, praticamente $m = 10$, e o termo de média é inferior a 0,05. É o item (d) do Corolário 2 observado nos dados: essas falhas não deixam assinatura nos resíduos, e suas taxas de detecção, de 1,3% a 1,9%, são da ordem da taxa de falsos alarmes.
- *Falhas de média:* na falha 4, degrau na temperatura de entrada da água de refrigeração do reator, o termo de média vale 88,7 e o de dispersão, 9,8, igual ao normal; a falha desloca a média dos resíduos de S1 e nada mais.

- *Falhas de dispersão:* nas falhas 10, 11, 14, 16 e 19, o termo de média é inferior a 1, e toda a assinatura está no termo de dispersão, de 45 a 287. Na falha 14, agarramento da válvula de água de refrigeração do reator, a oscilação provocada pelo atrito eleva a dispersão a 287, sem deslocar a média.

Convém notar que, como $\mathbf{S}_f$ é calculada com todas as amostras posteriores à falha, desvios de média que variam ao longo do tempo, como os transitórios das falhas 1, 5 e 7, aparecem no termo de dispersão. Além disso, um valor esperado alto não garante uma taxa de detecção alta: nas falhas 5 e 7, o valor esperado é elevado pelo transitório inicial, mas, depois que o controle compensa o degrau, os resíduos voltam ao normal, e a fração de amostras em alarme fica em 46% e 67%.

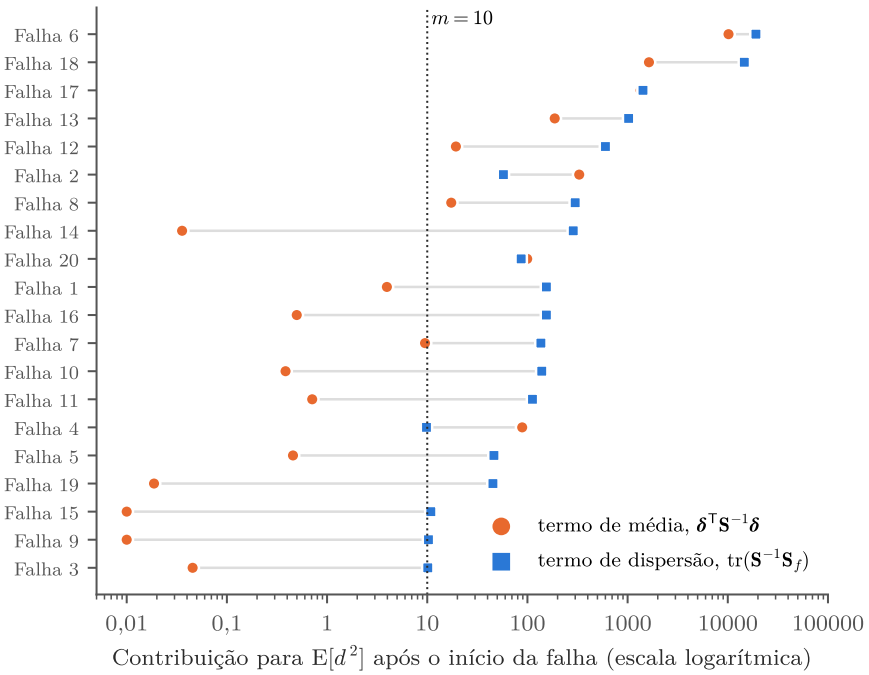


Figura 3: Decomposição (18) do valor esperado de d^2 após o início de cada falha (vetor de dez resíduos ARX, 75 simulações de teste), em escala logarítmica; a linha pontilhada marca o valor normal, $m = 10$. Termos de média menores que 0,01 são desenhados em 0,01.

4.4 O papel da janela da média móvel

O valor esperado não basta para detectar: o que importa é se a distribuição de $D(k)$ em falha se separa da distribuição normal o suficiente para ultrapassar o limiar. É aqui que a janela W atua.

Proposição 3 (Efeito da janela)**.** *A média móvel não altera o valor esperado,* $\mathbb{E}[D(k)] = \mathbb{E}[d^2(k)]$*, e sua variância é*

$$\mathrm{Var}\big[D(k)\big] = \frac{1}{W^2}\sum_{i=0}^{W-1}\sum_{j=0}^{W-1}\gamma(i-j), \qquad \gamma(h) = \mathrm{Cov}\big[d^2(k), d^2(k-h)\big], \tag{19}$$

supondo-se d^2 estacionário. Se os valores de d^2 forem independentes no tempo, $\mathrm{Var}[D] = \mathrm{Var}[d^2]/W$*; se, além disso, os resíduos forem gaussianos com deslocamento $\boldsymbol{\delta}$ constante e $\mathbf{S}_f = \kappa\mathbf{S}$, então $WD(k)/\kappa$ segue a distribuição qui-quadrado não central com mW graus de liberdade e parâmetro de não centralidade $W\lambda/\kappa$.*

Demonstração. A primeira afirmação é a linearidade da esperança, já usada no Lema 3. Para a variância, escreve-se $D = \frac{1}{W}\sum_i X_i$, com $X_i = d^2(k-i)$; pela bilinearidade da covariância,

$$\operatorname{Var}[D] = \operatorname{Cov}\Big[\frac{1}{W}\sum_i X_i,\ \frac{1}{W}\sum_j X_j\Big] = \frac{1}{W^2}\sum_i\sum_j \operatorname{Cov}[X_i,X_j] = \frac{1}{W^2}\sum_i\sum_j \gamma(i-j),$$

que é (19). Se os X_i forem independentes, $\gamma(h) = 0$ para $h \neq 0$ e só restam os W termos diagonais, iguais a $\gamma(0) = \operatorname{Var}[d^2]$, donde $\operatorname{Var}[D] = W\operatorname{Var}[d^2]/W^2 = \operatorname{Var}[d^2]/W$. No caso gaussiano, com $\mathbf{L}$ como na prova do Lema 3, $\mathbf{w}(k) = \mathbf{L}^{-1}[\mathbf{r}(k) - \boldsymbol{\mu}]/\sqrt{\kappa}$ tem covariância $\mathbf{L}^{-1}\kappa\mathbf{S}(\mathbf{L}^{-1})^{\mathsf{T}}/\kappa = \mathbf{I}_m$ e média $\mathbf{L}^{-1}\boldsymbol{\delta}/\sqrt{\kappa}$, de modo que $d^2(k)/\kappa = \mathbf{w}(k)^{\mathsf{T}}\mathbf{w}(k)$ é a soma dos quadrados de m normais independentes de variância unitária, cujas médias têm soma de quadrados $\boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}^{-1}\boldsymbol{\delta}/\kappa = \lambda/\kappa$. Somando-se W instantes independentes, $WD/\kappa = \sum_j \mathbf{w}(k-j)^{\mathsf{T}}\mathbf{w}(k-j)$ é a soma dos quadrados de mW normais independentes de variância unitária, com soma dos quadrados das médias $W\lambda/\kappa$, que é a definição da qui-quadrado não central indicada. ∎

A consequência prática é que, ao passar de $W = 1$ para W amostras, a média do índice em falha continua afastada da média normal pela mesma quantidade, mas o espalhamento de cada distribuição cai, no caso independente, pelo fator $\sqrt{W}$; as duas distribuições se separam, e a fração que ultrapassa o limiar cresce. A Tabela 2 quantifica o efeito no caso ideal da Proposição 3, com $m = 10$ e limiar teórico de 1% de falsos alarmes: uma falha que aumenta as variâncias em 50% ($\kappa = 1{,}5$) é detectada em 11,6% das amostras com $W = 1$ e em 74,0% com $W = 10$.

Tabela 2: Probabilidade de detecção no caso ideal (resíduos gaussianos e independentes no tempo, $m = 10$, FAR de 1%), para falhas que só alteram a dispersão (κ) ou só deslocam a média (λ)

W	Limiar $\chi^2_{mW;\,0{,}99}/W$	Dispersão		Média	
		$\kappa = 1{,}5$	$\kappa = 2$	$\lambda = 5$	$\lambda = 10$
1	23,21	11,6%	31,2%	10,6%	30,6%
5	15,23	44,3%	89,2%	44,4%	92,4%
10	13,58	74,0%	99,4%	75,5%	99,8%
20	12,47	96,1%	100%	96,9%	100%

Os dados reais seguem o mesmo padrão, embora com ganho menor que o ideal, pois a autocorrelação dos resíduos faz $\gamma(h) \neq 0$ em (19): em operação normal, a variância de D cai de 21,0 com $W = 1$ para 8,4 com $W = 10$, e não para 2,1, como no caso independente. Ainda assim, na falha 19, cujo efeito está todo na dispersão (Figura 3), a taxa de detecção sobe de 75,5% com $W = 1$ para 98,7% com $W = 10$. A Figura 4 mostra o mecanismo: com $W = 1$, as distribuições normal e em falha se sobrepõem amplamente; com $W = 10$, ambas se estreitam e a distribuição em falha passa quase inteira para a direita do limiar.

4.5 Ordem de grandeza do atraso

A janela tem um custo: o índice reage a uma mudança abrupta só gradualmente, à medida que as amostras em falha entram na janela.

Proposição 4 (Atraso em média para um deslocamento abrupto)**.** *Suponha-se que, a partir de k_0, $\mathbb{E}[d^2]$ passe de m para $m + \lambda$, com $\lambda > 0$. Então, para $n = 1,\ldots,W$ amostras em falha dentro da janela,*

$$\mathbb{E}\big[D(k_0 + n - 1)\big] = m + \frac{n}{W}\,\lambda, \tag{20}$$

e o valor esperado do índice ultrapassa o limiar a partir de

$$n^* = \left\lceil \frac{W(\tau - m)}{\lambda} \right\rceil \ \text{amostras, desde que } \lambda > \tau - m. \tag{21}$$

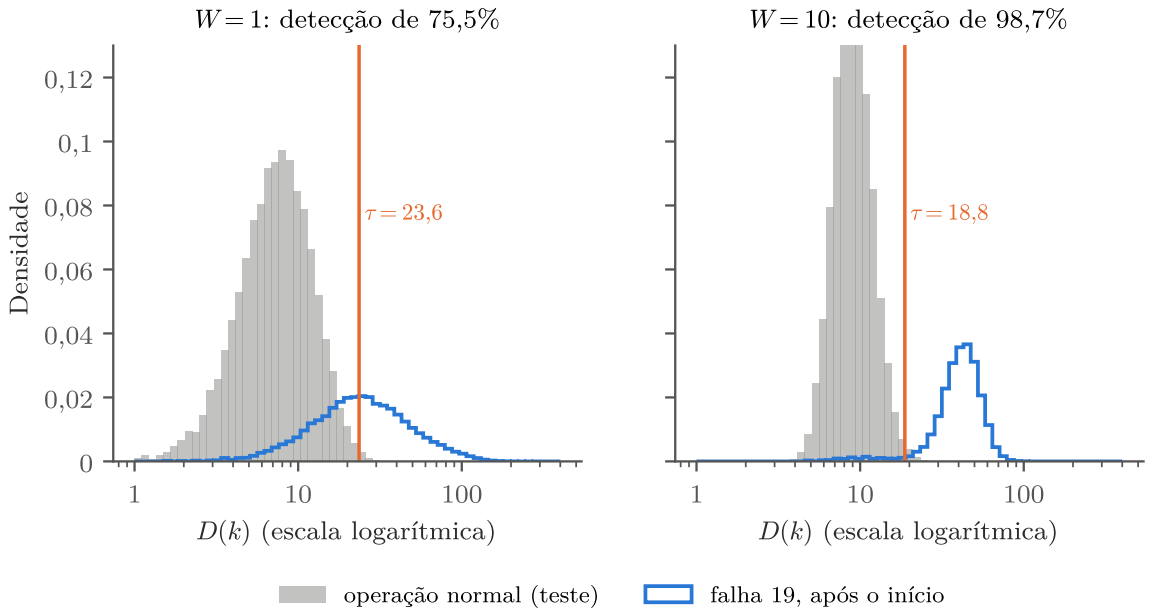


Figura 4: Distribuição de $D(k)$ nas 75 simulações normais de teste e nas 75 simulações com a falha 19, após o início da falha, para $W = 1$ e $W = 10$; a linha vertical é o limiar calibrado na validação.

Demonstração. No instante $k_0 + n - 1$, a janela contém as n amostras $k_0, \ldots, k_0 + n - 1$, em falha, e as $W - n$ anteriores, normais. Pela linearidade da esperança,

$$\mathbb{E}[D] = \frac{1}{W}\big[n(m + \lambda) + (W - n)\, m\big] = \frac{1}{W}\big[Wm + n\lambda\big] = m + \frac{n}{W}\lambda.$$

A condição $m + n\lambda/W > \tau$ equivale a $n > W(\tau - m)/\lambda$, e o menor inteiro que a satisfaz é (21), a menos do caso de igualdade, irrelevante na prática. Como $n \leq W$, a ultrapassagem em média só ocorre se $W(\tau - m)/\lambda < W$, isto é, $\lambda > \tau - m$. ■

Com os valores do artigo, $W = 10$, $\tau = 18{,}76$ e $m = 10$, tem-se $\tau - m = 8{,}76$. Na falha 4, $\lambda \approx 88{,}7$ e (21) dá $n^* = 1$: o índice cruza o limiar já na primeira amostra em falha, o que se vê na Figura 5. Uma falha com $\lambda = 20$ levaria $n^* = 5$ amostras, isto é, 15 min. Uma falha com $\lambda < 8{,}76$ não faz o valor esperado cruzar o limiar, embora possa ser detectada numa fração das amostras pelas flutuações do índice. A escolha de W é, portanto, um compromisso: janelas maiores aumentam a sensibilidade a falhas pequenas e persistentes (Proposição 3) e atrasam a resposta a falhas abruptas (Proposição 4). O valor $W = 10$ (30 min) foi indicado pela análise exploratória, e, no artigo, retirar a média móvel custou de 4 a 17 pontos percentuais de taxa de detecção.

5 Uso do índice na detecção de falhas

5.1 Procedimento

O uso do índice divide-se numa fase de projeto, feita uma única vez com dados de operação normal, e numa fase de operação, executada a cada nova amostra.

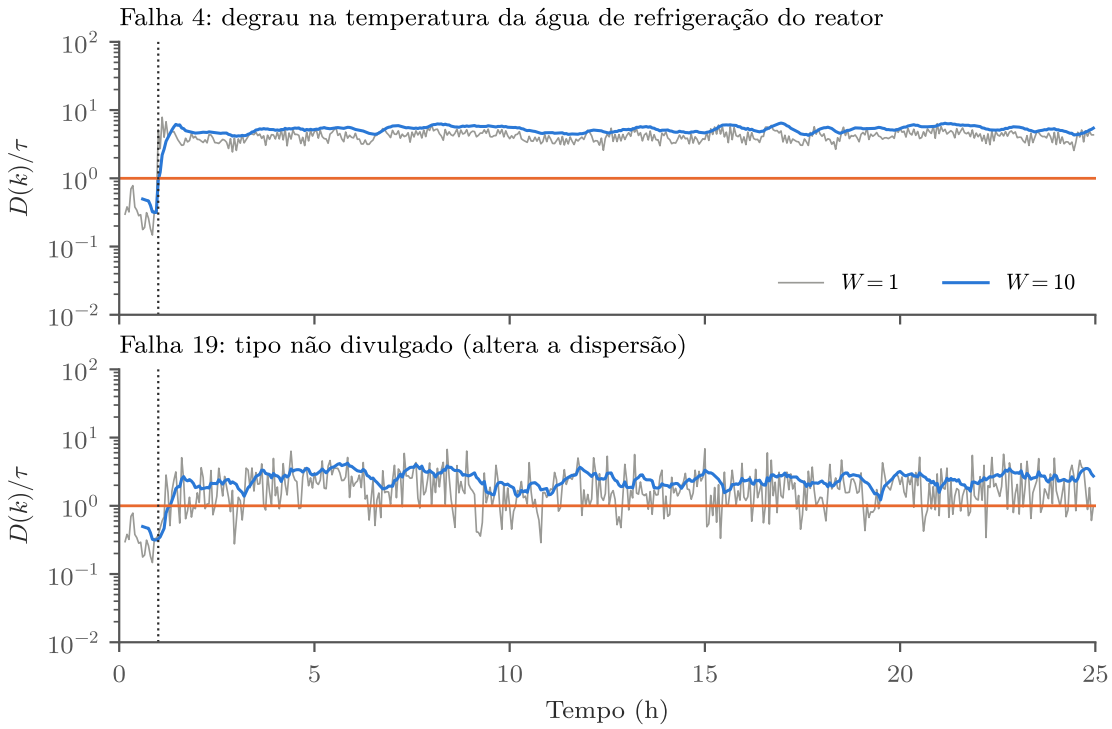


Figura 5: Índice $D(k)$ normalizado pelo limiar na simulação de teste 426, com $W = 1$ e $W = 10$, para a falha 4 (deslocamento de média) e a falha 19 (alteração da dispersão); há alarme quando a curva está acima de 1.

Fase de projeto (com dados de operação normal)

1. Separar as simulações normais em treino, validação e teste (no artigo, 350, 75 e 75 simulações).
2. Padronizar as variáveis pela média e pelo desvio padrão do treino.
3. Identificar, com o treino, o modelo de cada subsistema (1).
4. Calcular, nas simulações de validação, as predições (4) e (5), os resíduos (6) e o vetor $\mathbf{r}(k)$ (10).
5. Estimar $\boldsymbol{\mu}$ e $\mathbf{S}$ por (11) e calcular $\mathbf{S}^{-1}$ (ou $\mathbf{S}^{+}$).
6. Calcular $D(k)$ (13) na validação e fixar τ pelo quantil (14).
7. Conferir, nas simulações normais de teste, que a fração de alarmes fica próxima da nominal.

Fase de operação (a cada amostra k)

1. Medir as saídas $y_i(k)$ e as entradas $u_{ij}(k)$ e padronizá-las.
2. Para cada subsistema, calcular $\hat{y}_i(k \mid k-1)$ com as saídas medidas e atualizar $\hat{y}_{s,i}(k)$ com a recursão (5).
3. Formar $\mathbf{r}(k)$ com as diferenças $y_i(k) - \hat{y}_i(k \mid k-1)$ e $y_i(k) - \hat{y}_{s,i}(k)$.
4. Calcular $d^2(k)$ (12) e atualizar a média móvel, $D(k) = D(k-1) + \big[d^2(k) - d^2(k-W)\big]/W$.
5. Declarar alarme se $D(k) > \tau$; para reduzir alarmes isolados, pode-se exigir, como no cálculo do atraso no artigo, três alarmes consecutivos.

O custo por amostra é pequeno: algumas dezenas de multiplicações por subsistema para as predições e uma forma quadrática $m \times m$, com $m = 10$. A atualização recursiva da média móvel, no passo 4, segue diretamente de (13), pois $WD(k) - WD(k-1) = d^2(k) - d^2(k-W)$, já que as $W - 1$ parcelas comuns se cancelam.

5.2 Implementação

O código abaixo é o da versão MATLAB do projeto (funções `detector_ajustar` e `detector_indice`); os resíduos são organizados num arranjo de dimensões simulações × amostras × resíduos.

```
function det = detector_ajustar(Rval, W, q)
% Rval: residuos das simulacoes normais de validacao (R x T x m)
m = size(Rval, 3);
v = reshape(Rval, [], m);  v = v(all(isfinite(v), 2), :);
det.mu = mean(v, 1);                % vetor de medias (1 x m)
det.Si = pinv(cov(v));              % pseudoinversa de S
det.W  = W;
det.tau = quantil_nan(detector_indice(det, Rval), q);   % limiar
end

function D = detector_indice(det, Rs)
[R, T, m] = size(Rs);
d = reshape(Rs, [], m) - det.mu;          % r(k) - mu
Q = sum((d * det.Si) .* d, 2);            % d2(k), forma quadratica
D = media_movel(reshape(Q, R, T), det.W);  % media movel causal de W amostras
end
```

O uso, com os resíduos de validação `Rval` e os de uma condição qualquer `Rteste`, resume-se a:

```
det    = detector_ajustar(Rval, 10, 0.99);   % W = 10, FAR nominal de 1%
D      = detector_indice(det, Rteste);
alarme = D > det.tau;
FDR    = mean(alarme(:, 21:end), 'all', 'omitnan');   % falha inicia na amostra 21
```

5.3 Do alarme ao subsistema

O índice global responde se há falha, mas não onde. Para o isolamento, o artigo aplica a mesma construção a cada subsistema separadamente, com o resíduo de simulação livre $r_{\mathrm{sim},i}(k)$ de um só subsistema. Com $m = 1$, $\mathbf{S}$ reduz-se à variância σ_i^2, e, pelo Lema 1,

$$D_i(k) = \frac{1}{W} \sum_{j=0}^{W-1} \left(\frac{r_{\mathrm{sim},i}(k-j) - \mu_i}{\sigma_i} \right)^2, \tag{22}$$

com limiar próprio τ_i, também calibrado no quantil de 99% da validação. A fração de amostras em que $D_i(k) > \tau_i$ indica quais subsistemas foram afetados; no artigo, ela apontou corretamente S1 nas falhas 4, 11 e 14 e S2 na falha 10. O isolamento indica onde a relação entre entradas e saída se alterou, o que não é necessariamente onde a falha se originou, pois as falhas se propagam pela planta e pelas malhas de controle.

6 Limitações e cuidados

- **Representatividade da operação normal.** O índice mede afastamento em relação ao comportamento dos resíduos na validação. Uma mudança de ponto de operação, de carga ou de sintonia dos controladores que não esteja representada nesses dados produz alarmes sem que haja falha, e exige recalibrar $\boldsymbol{\mu}$, $\mathbf{S}$ e τ ou reidentificar os modelos.
- **Falsos alarmes acima do nominal.** No artigo, a FAR medida no teste ficou entre 1,1% e 1,4%, um pouco acima de 1%, porque o limiar é estimado com amostras autocorrelacionadas e, portanto, com menos informação efetiva do que o número de amostras sugere.
- **Falhas sem assinatura.** Pelo Corolário 2(d), nenhuma escolha de limiar detecta falhas que não alteram a distribuição dos resíduos, como as falhas 3, 9 e 15 do TEP; reduzir o limiar só aumenta igualmente a detecção e os falsos alarmes.
- **Compensação pelo controle.** Quando as malhas de controle restabelecem as relações modeladas após um transitório, como nas falhas 5 e 7, os resíduos voltam ao normal, e a falha só é visível durante o transitório; índices que observam todas as variáveis, como a PCA, complementam o detector, e a fusão dos dois obteve a maior taxa de detecção no artigo.
- **Hipóteses das proposições.** As Proposições 1 e 2 valem exatamente para modelos ARX; a Proposição 3, na forma qui-quadrado, e a Tabela 2 supõem resíduos gaussianos e independentes no tempo e servem para orientar a escolha de W, não para fixar o limiar, que deve ser empírico.

7 Observações finais

O índice $D(k)$ transforma o par formado pela saída medida e pela saída predita de cada subsistema num único número, cuja leitura é simples: em operação normal, ele flutua em torno do número de resíduos m, e uma falha o eleva pela soma de duas contribuições, uma do deslocamento da média dos resíduos e outra da alteração da sua dispersão (Lema 4). Os ingredientes do índice têm justificativas independentes e complementares: os dois tipos de resíduo reagem de forma diferente a desvios rápidos e persistentes, e são, no caso ARX, um a versão filtrada do outro (Proposições 1 e 2); a matriz $\mathbf{S}^{-1}$ torna o índice invariante a escalas e evita contar duas vezes a mesma informação (Lemas 1 e 2); a média móvel separa as distribuições normal e em falha à custa de um atraso controlado (Proposições 3 e 4); e o limiar empírico absorve o que as hipóteses idealizadas não descrevem.

A decomposição (18) sugere ainda duas extensões naturais: acompanhar separadamente os termos de média e de dispersão, o que ajudaria a caracterizar o tipo de falha, e usar o vetor das contribuições de cada resíduo, ou as frações de alarme por subsistema, como assinatura para um classificador de diagnóstico, etapa que não foi abordada no projeto.

Referências

[1] M. Basseville e I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs: Prentice Hall, 1993.

[2] R. V. Beard, *Failure Accommodation in Linear Systems through Self-Reorganization*. Tese (Doutorado), Massachusetts Institute of Technology, 1971.

[3] S. A. Billings, *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. Chichester: Wiley, 2013.

[4] L. H. Chiang, E. L. Russell e R. D. Braatz, *Fault Detection and Diagnosis in Industrial Systems*. Londres: Springer, 2001.

[5] E. Y. Chow e A. S. Willsky, "Analytical redundancy and the design of robust failure detection systems," *IEEE Transactions on Automatic Control*, v. 29, n. 7, p. 603–614, 1984.

[6] J. J. Downs e E. F. Vogel, "A plant-wide industrial process control problem," *Computers & Chemical Engineering*, v. 17, n. 3, p. 245–255, 1993.

[7] P. M. Frank, "Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: a survey and some new results," *Automatica*, v. 26, n. 3, p. 459–474, 1990.

[8] J. J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*. Nova York: Marcel Dekker, 1998.

[9] H. Hotelling, "Multivariate quality control, illustrated by the air testing of sample bombsights," em C. Eisenhart, M. W. Hastay e W. A. Wallis (Eds.), *Techniques of Statistical Analysis*. Nova York: McGraw-Hill, 1947, p. 111–184.

[10] R. J. Hyndman e Y. Fan, "Sample quantiles in statistical packages," *The American Statistician*, v. 50, n. 4, p. 361–365, 1996.

[11] R. Isermann, *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Berlim: Springer, 2006.

[12] H. L. Jones, *Failure Detection in Linear Systems*. Tese (Doutorado), Massachusetts Institute of Technology, 1973.

[13] C. A. Lowry, W. H. Woodall, C. W. Champ e S. E. Rigdon, "A multivariate exponentially weighted moving average control chart," *Technometrics*, v. 34, n. 1, p. 46–53, 1992.

[14] P. C. Mahalanobis, "On the generalised distance in statistics," *Proceedings of the National Institute of Sciences of India*, v. 2, n. 1, p. 49–55, 1936.

[15] R. K. Mehra e J. Peschon, "An innovations approach to fault detection and diagnosis in dynamic systems," *Automatica*, v. 7, n. 5, p. 637–640, 1971.

[16] R. Penrose, "A generalized inverse for matrices," *Mathematical Proceedings of the Cambridge Philosophical Society*, v. 51, n. 3, p. 406–413, 1955.

[17] C. A. Rieth, B. D. Amsel, R. Tran e M. B. Cook, *Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation*. Harvard Dataverse, 2017. doi:10.7910/DVN/6C3JR1.

[18] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, v. 12, n. 6, p. 601–611, 1976.