Universidade de São Paulo
BRASIL

Análise estatística multivariada aplicada a processos químicos

- **Escola Politécnica**
- **Departamento de Engenharia Química**

- Prof. Dr. Cláudio Augusto Oller do Nascimento
- Prof. Dr. Roberto Guardani

- 2007

Parte 2. Correlação e Covariância

Identificação de dados anômalos ("outliers")

Conceitos Básicos e Definições

Variáveis aleatórias consideradas: $j = 1,...p$ (*p* = número de variáveis)

Observações (medições): $i = 1,...n$ (*n* = número de observações)

Grupos: $g = 1,...G$ (*G* = número de grupos)

Vetor de médias da amostra, $\bar{x}_j$, ou de valores esperados, $E(x_j)$, das variáveis x_j:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} = E(x_j), \qquad j = 1, p \tag{2.1}$$

Para o vetor de médias da população utiliza-se normalmente o símbolo μ_j. Outras formas de expressar a média:

$$\mu_j, \text{ ou } \bar{x}_j = \sum_{i=1}^{n} x_{ij} . \Delta P_i(x_j) \tag{2.2}$$

$$\mu_j, \text{ ou } \bar{x}_j = \int_{-\infty}^{+\infty} x_j . p(x_j) dx_j \tag{2.3}$$

em que $\Delta P_i(x_j)$ representa a probabilidade de ocorrência da *i*-ésima observação da variável x_j e $p(x_j)$ representa a função densidade de probabilidade de x_j.

A *i*-ésima observação da variável aleatória centrada na média X_j é definida como $X_{ij} = x_{ij} - \bar{x}_j$, ou $X_{ij} = x_{ij} - \mu_j$, $j = 1, p$. Portanto, a variável X_j tem média igual a zero e segue a mesma distribuição que x_j.

Variância de uma variável aleatória:

$$s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{GL} = \frac{\sum_{i=1}^{n} (X_{ij})^2}{n-1} = \frac{SS_j}{n-1} \tag{2.4}$$

O símbolo SS_j refere-se à soma dos quadrados (“sum of squares”) da variável centrada na média, X_j.

Desvio padrão de uma variável aleatória:

$$s_j = \sqrt{s_j^2} \tag{2.5}$$

Comumente utiliza-se o símbolo *s* para o desvio padrão da amostra de observações e o símbolo σ para o desvio padrão de toda a população.

i-ésima observação da variável aleatória padronizada z_j:

$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$ (para a amostra), ou $z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$ (para toda a população).

Cada variável padronizada z_j, j=1, p, é adimensional, tem média igual a zero e desvio padrão igual a 1, o que auxilia os procedimentos de análise, porque eliminam-se as diferenças numéricas entre diferentes variáveis.

Covariância entre as variáveis x_j e x_k:

$$s_{jk} = \frac{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{GL} = \frac{\sum_{i=1}^{n}\left(X_{ij}\right)\left(X_{ik}\right)}{n-1} = \frac{SCP_{jk}}{n-1} \qquad (2.6)$$

O símbolo SCP_{jK} refere-se à “sum of cross products”, ou soma dos produtos cruzados das variáveis centradas na média X_j e X_k.

$\rightarrow$ No caso de p variáveis, é mais fácil representar as somas SS_j e SCP_{jk} na forma da matriz de covariância, $\mathbf{\Sigma}$ (pxp), que tem a seguinte forma:

$$\mathrm{cov}(x) = \mathbf{\Sigma} = \begin{bmatrix} s_1^2 & s_{12} & \ldots & s_{1p} \\ s_{21} & s_2^2 & \ldots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_p^2 \end{bmatrix} \qquad (2.7)$$

A matriz de covariância, que é simétrica, contém, na diagonal principal, a variância das p variáveis e, nos demais elementos, as covariâncias das variáveis. Se as p variáveis forem independentes, então a matriz $\mathbf{\Sigma}$ é diagonal.

Em notação matricial, para uma população de n observações das p variáveis e vetor de médias $\boldsymbol{\mu}$ (px1), a matriz de covariância pode ser representada como:

$$\mathbf{\Sigma} = E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right] = E\left(\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_p-\mu_p \end{bmatrix}\begin{bmatrix}(x_1-\mu_1) & (x_2-\mu_2) & \ldots & (x_p-\mu_p)\end{bmatrix}\right) \qquad (2.8)$$

em que o valor esperado de um argumento t, $E(t)$, é definido como nas Eqs. 2.1 a 2.3. No caso, o argumento é o produto dos vetores ($\mathbf{x}$ - $\boldsymbol{\mu}$) e ($\mathbf{x}$ - $\boldsymbol{\mu}$)^T (o símbolo ^{T} indica a matriz transposta) e, no denominador, o valor n é substituído por (n – 1), que é o número de graus de liberdade referente ao cálculo da variância, como na Eq. 2.6.

O cálculo da covariância entre duas variáveis quaisquer pode ser feito também utilizando-se as variáveis padronizadas *z*. Nesse caso, obtém-se o coeficiente de correlação entre as duas variáveis x_j e x_k, r_{jk}:

$$r_{jk} = \frac{\sum_{i=1}^{n}\left(z_{ij}\right)\left(z_{ik}\right)}{n-1} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)\left(x_{ik}-\bar{x}_{ik}\right)}{(n-1).s_j.s_k} = \frac{s_{jk}}{s_j.s_k} \tag{2.9}$$

É comum representar a correlação entre duas variáveis na forma do coeficiente de determinação, que corresponde a $(r_{jk})^2$. Nos casos em que $j = k$, a Eq. 2.9 resulta na divisão da variância por ela mesma:

$$r_{jj} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}{(n-1).s_j^2} = \frac{s_j^2}{s_j^2} = 1 \tag{2.10}$$

O coeficiente de correlação de uma dada variável em relação a si própria vale 1, e a correlação entre duas variáveis quaisquer varia no intervalo de –1 a 1. Devido a essas propriedades, o coeficiente de correlação é um indicador importante da correlação linear entre variáveis aleatórias, pois valores próximos a zero indicam ausência de correlação (variáveis linearmente independentes) e valores próximos a 1 ou –1 indicam variáveis linearmente correlacionadas (positiva ou negativamente). É importante observar que o fato de haver, eventualmente, valor de r_{jk} próximo de zero não significa que as variáveis x_j e x_k não sejam correlacionadas. Tal fato é ilustrado na Fig. 2.1.

Se a matriz de covariância, **Σ**, for construída com variáveis padronizadas, é chamada de matriz de correlação, $\mathbf{R}_{(pxp)}$:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{bmatrix} \tag{2.11}$$

A matriz de correlação é simétrica e tem os elementos da diagonal principal iguais a 1. No caso de variáveis aleatórias linearmente independentes, essa matriz torna-se diagonal. O uso da matriz de correlação é importante nos casos em que as variáveis têm valores numéricos muito diferentes, pois tais valores afetam a matriz de covariância, enquanto a matriz de correlação tem os valores dos elementos limitados entre –1 e 1.

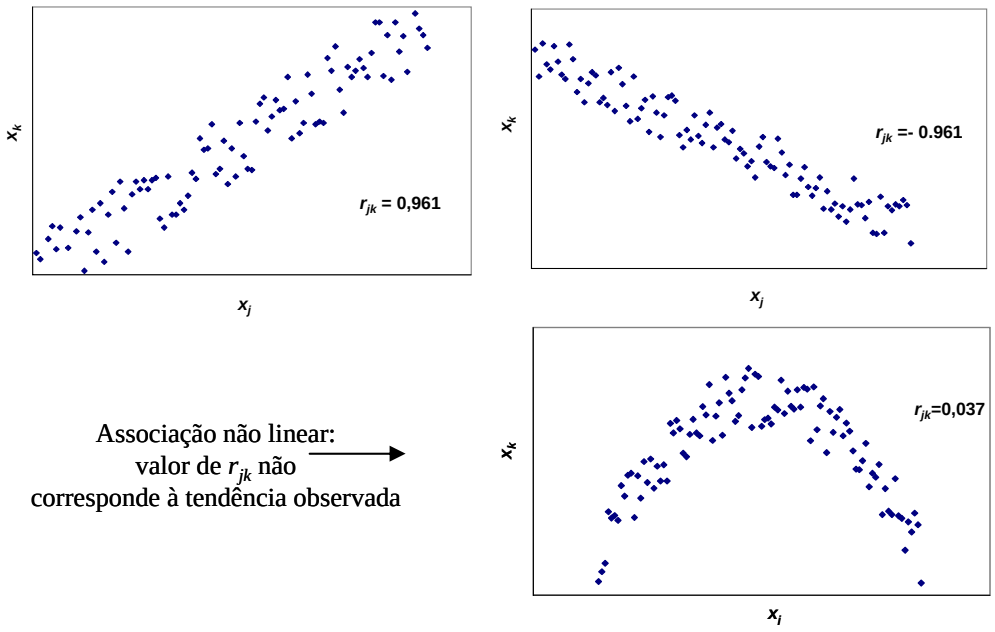

Figura 2.1. Exemplos de gráficos de dispersão de valores de duas variáveis, x_j e x_k, com os respectivos valores de r_{jk}.

Definições de distância

Muitos métodos estatísticos baseiam-se na distância entre observações, ou grupos de observações, ou, alternativamente, na distância entre variáveis, ou grupos de variáveis. A distância mais comumente usada é a distância euclideana. Dadas duas observações, *A* e *B* nas variáveis x_1 e x_2, conforme mostra a Figura 2.2, a distância euclideana entre os pontos é definida como:

$$D_{AB} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \qquad (2.12)$$

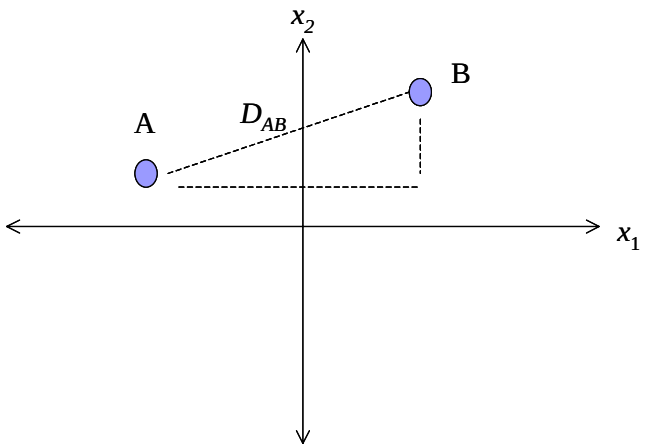

Figura 2.2. Ilustração da definição de distância euclideana entre duas observações *A* (coordenadas a_1, a_2)e *B* (coordenadas b_1, b_2) nas variáveis x_1 e x_2.

No caso de um sistema com *p* variáveis, a distância euclideana é:

$$D_{AB} = \sqrt{\sum_{j=1}^{p} \left(a_j - b_j\right)^2} \tag{2.13}$$

Distância estatística entre duas observações quaisquer, *i* e *h*, de uma variável:

$$SD_{ih}^2 = \left(\frac{x_i - x_h}{s} \right)^2 \tag{2.14}.$$

Distância estatística entre duas observações quaisquer, *i* e *h*, para *p* variáveis:

$$SD_{ih}^2 = \sum_{j=1}^{p} \left(\frac{x_{ij} - x_{hj}}{s_j} \right)^2 \tag{2.15}$$

No cálculo da distância estatística são consideradas as dispersões nas medidas das variáveis, dividindo-se a distância quadrática pela variância de cada variável.

Distância estatística generalizada entre duas observações pode também ser expressa em notação matricial:

$$SD_{ik} = \left(\mathbf{x}_i - \mathbf{x}_k\right)^T \mathbf{\Sigma}^{-1} \left(\mathbf{x}_i - \mathbf{x}_k\right) \tag{2.16}$$

em que **x** é um vetor vertical de dimensão *p*.

Densidade de distribuição normal de probabilidade multivariada: é obtida substituindo-se a distância estatística no expoente da função no caso de uma única variável, Eq. 1.14, pela distância estatística entre cada ponto e a média, na forma generalizada (Eq. 2.16). O termo pré-exponencial deve ser alterado para que o hipervolume sob a curva tenha área total igual a 1, ficando na forma da Eq. 2.17:

$$p(x) = \frac{1}{(2\pi)^{p/2} \left|\Sigma\right|^{-1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})/2} \tag{2.17}$$

Identificação de dados anômalos (“outliers”)

Dados anômalos são dados presentes em uma população ou amostra que correspondem a erros grosseiros em relação à média, não representando a tendência geral observada nos valores de uma dada variável, ou conjunto de variáveis, de um processo. Podem ser causados por falhas em instrumentos, ou procedimentos de medida, ou no registro de dados de processo. Há várias técnicas para detecção de dados anômalos, sendo os mais comuns baseados no comportamento individual de cada variável. Há, também, técnicas de detecção de dados anômalos multivariados, que serão discutidas em outras partes do curso.

Entre as técnicas mais comuns estão aquelas baseadas em cartas de controle de processo, que são séries temporais de cada variável. Como exemplo, a Figura 2.3a mostra uma série de dados registrados de vazão em um processo industrial. Na Figura 2.3b, foram acrescentados à mesma série algumas referências, ou seja, a média da população de valores e dois limites de controle: superior (LSC) e inferior (LIC). Esses valores correspondem à média mais uma faixa de tolerância, normalmente baseada na dispersão dos dados e em na tolerância baseada na qualidade especificada para o processo. No caso, esses limites correspondem a uma tolerância de $\pm 3\sigma$, resultando em uma faixa total de 6σ, o que, para uma distribuição normal, deve conter 99,73% dos valores da variável. Observa-se que há dois pontos situados for a dos limites, o que é um forte indício (com 99,73% de probabilidade) de que se tratam de dados com erros. O mesmo pode ser deduzido na Figura 2.4, que apresenta um histograma dos mesmos dados, com pontos isolados nos extremos. Em estudos baseados em dados de processo, esses dados devem ser eliminados para não trazer erros. Uma das formas mais eficientes de detecção de dados anômalos é o exame cuidadoso de séries e distribuições dos dados, como nas Figuras 2.3 e 2.4, e com base no conhecimento sobre o processo.

O exame de cartas de controle é usado, também, na detecção de tendências tanto nos valores médios (calculando-se a média móvel) quanto na dispersão dos dados.

A Figura 2.5 ilustra um caso de “outlier” bivariado, para o qual a distância euclideana daria resultados incorretos, uma vez que o ponto C está mais próximo do ponto A do que o ponto B pela distância euclideana, enquanto, pela distância estatística, estaria fora da região dos demais pontos. Na Figura 2.6, são mostrados dois exemplos de “outliers”: mono e bivariado.

A detecção de “outliers” mono e bivariados é relativamente simples, uma vez que o procedimento envolve gráficos de séries temporais, ou bivariados, como nos exemplos da Figura 2.6. No entanto, a identificação é muito mais difícil quando o número de variáveis é alto.

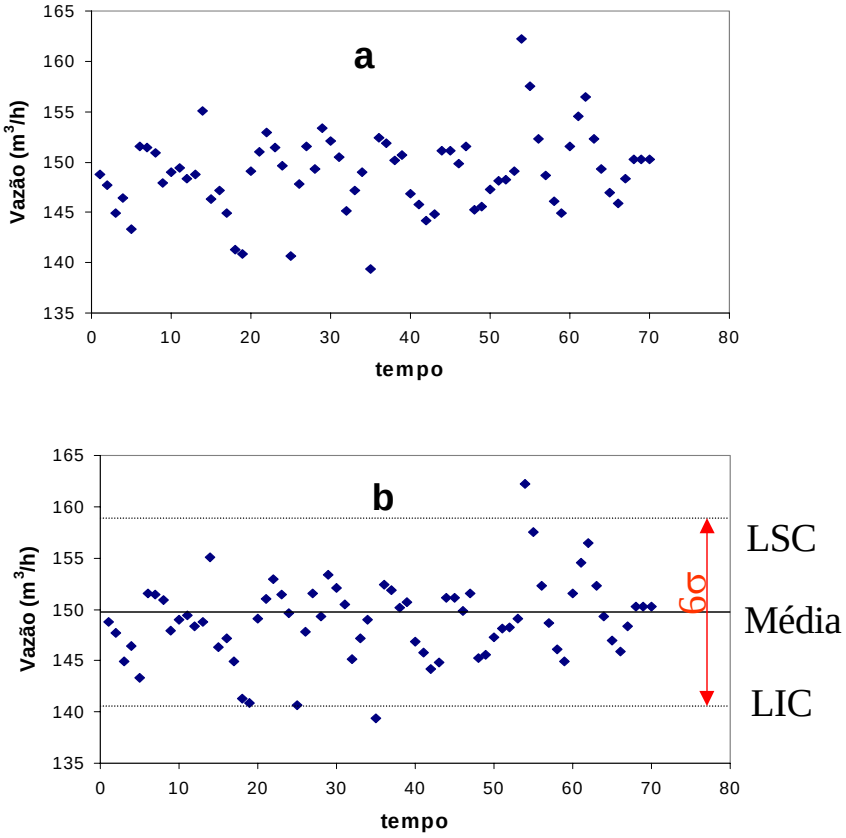

Figura 2.3. a) Exemplo de série temporal de dados de processo, com valores anômalos. B) A mesma série, como carta de controle.

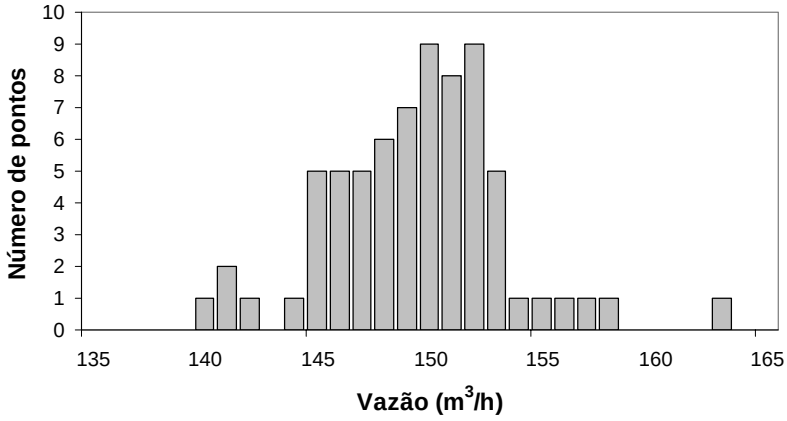

Figura 2.4. Histograma da série temporal da Figura 3.2.

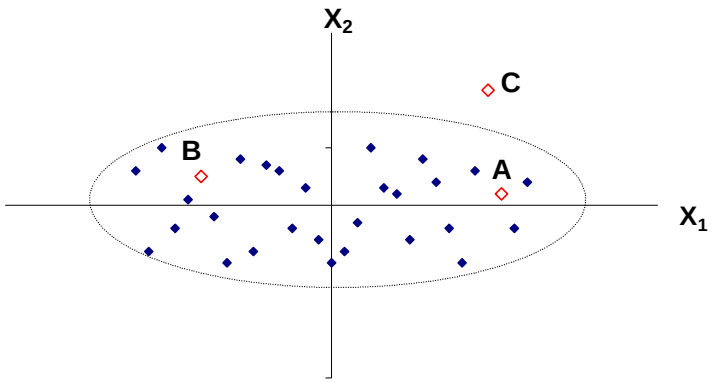

Figura 2.5. Exemplo de “outlier” bivariado.

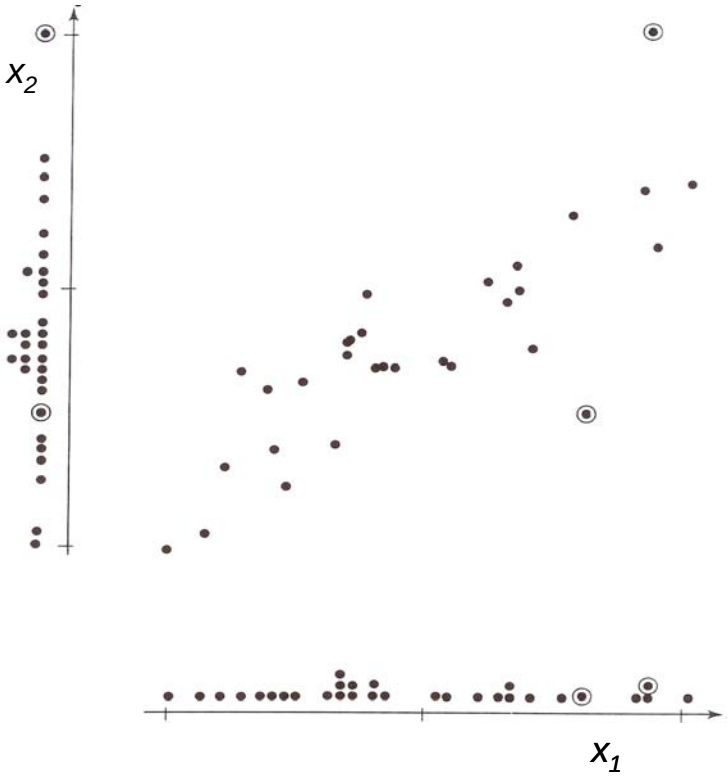

Figura 2.6. Exemplo de “outliers” mono e bivariado.

Procedimento sugerido para detecção de “outliers” em bases de dados multivariados.

No caso de sistemas com número de variáveis *p* maior que 2 fica difícil a identificação de “outliers” baseada apenas em séries temporais. A seqüência de testes apresentada a seguir tem-se mostrado eficiente, nesses casos.

1) Exame de séries temporais de cada variável, x_j, j = 1,...p. O exame deve envolver: gráficos de séries temporais e de histogramas.
2) Exame de gráficos de dispersão bivariados (como na Fig. 2.6).
3) Transformação das variáveis originais x_j em variáveis padronizadas z_j. Examinar os valores extremos dessa variável para as *n* observações. Supondo-se uma distribuição normal dos dados, a probabilidade de haver valores de *z* maiores que +3, ou menores que -3 é de 0,27%. Pode-se adotar também intervalos de confiança para estabelecer os valores de corte de *z*.
4) Cálculo da distância estatística para cada observação das *p* variáveis (Eq. 2.16). A vantagem deste critério é que leva em consideração o padrão de distância quadrática e a dispersão das medidas de todas as *p* variáveis em cada observação. O critério de corte, no caso dessa distância quadrática, pode ser o valor da variável χ_ν^2 (qui-quadrado), com número de graus de liberdade ν igual a *p*. Assim, estabelecido um grau de significância, valores da distância estatística maiores que χ_ν^2 podem ser considerados “outliers” e devem ser examinados com cuidado.