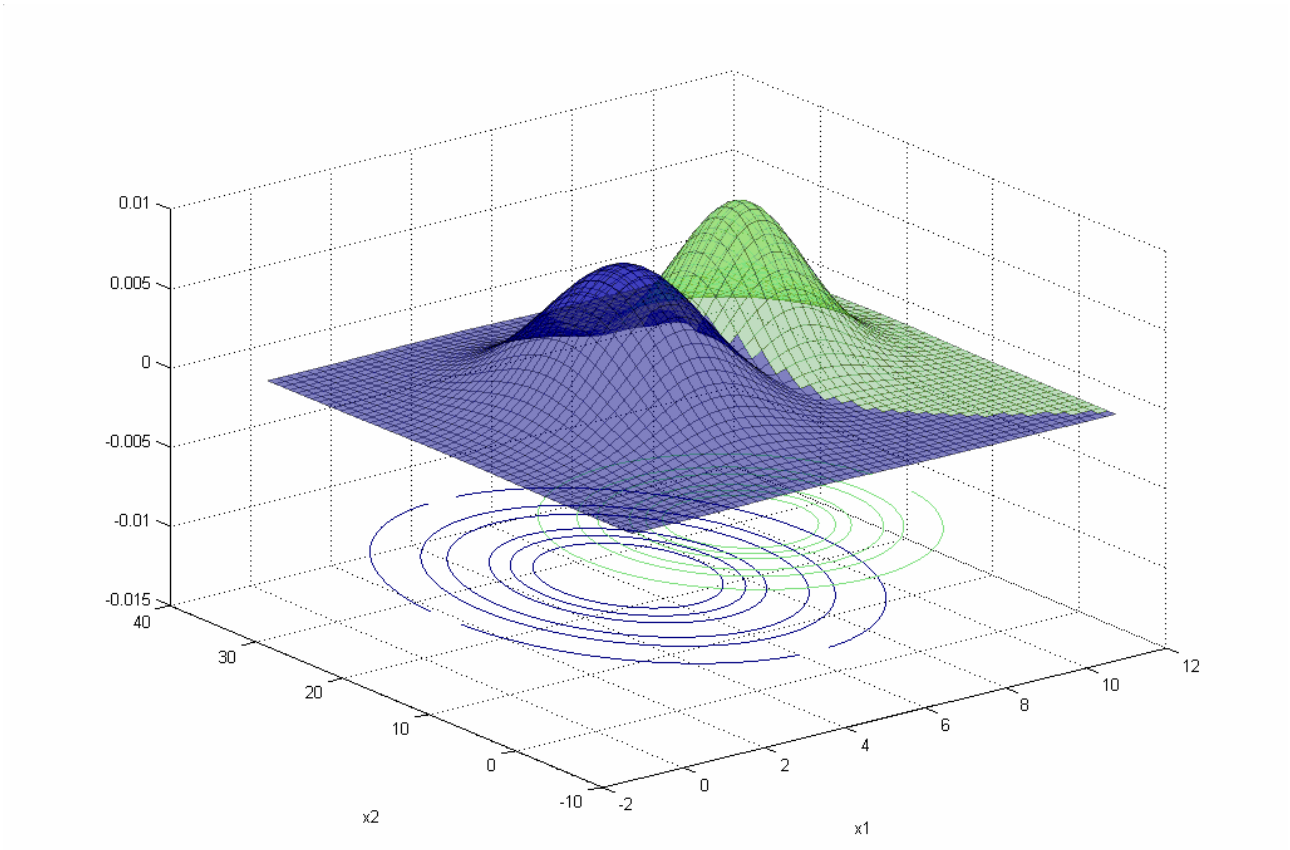
TEORIA DE PROBABILIDADES

Folhas de Apoio da disciplina de Detecção e Estimação



Dezembro de 2003

Isabel Milho

Secção de Análise de Sinais
Departamento de Engenharia de Electrónica e de Telecomunicações e de Computadores
Instituto Superior de Engenharia de Lisboa

Índice

1. Variáveis Aleatórias Discretas

Seja x uma variável aleatória (v.a.) que pode tomar um número finito m de valores diferentes no conjunto $\mathcal{X} = \{v_1, v_2, ..., v_m\}$. Designa-se por p_i a probabilidade de x tomar o valor v_i:

$$p_i = \Pr[x = v_i], \quad i = 1, ..., m. \tag{1}$$

As probabilidades p_i satisfazem as seguintes condições:

$$p_i \geq 0 \qquad \text{e} \qquad \sum_{i=1}^{m} p_i = 1. \tag{2}$$

FUNÇÃO MASSA DE PROBABILIDADE

Por vezes é conveniente exprimir o conjunto de probabilidades $\{p_1, p_2, ..., p_m\}$ em termos da *função de massa de probabilidade* (ou distribuição de probabilidade discreta) $P(x)$, que satisfaz as seguintes condições:

$$P(x) \geq 0 \qquad \text{e} \qquad \sum_{i=1}^{m} P(x) = 1. \tag{3}$$

2. Operador Valor Expectável

MÉDIA

O *valor expectável*, *média* ou *valor médio* da v.a. x é definido por

$$\mathrm{E}[x] = \mu_x = \sum_{x \in \mathcal{X}} xP(x) = \sum_{i=1}^{m} v_i p_i \tag{4}$$

Considerando a função de massa de probabilidade como um conjunto de pontos de massa, sendo p_i a massa concentrada em $x = v_i$, então o valor expectável μ_x é o centro de massa. Alternativamente, pode-se interpretar μ_x como a média aritmética dos valores de um conjunto significativo de amostras de x. Genericamente, se $f(x)$ é uma função de x, o valor esperado de f é definido por

$$\mathrm{E}[f(x)] = \sum_{x \in \mathcal{X}} f(x)P(x). \tag{5}$$

OPERADOR VALOR EXPECTÁVEL

Note que o *operador valor expectável* é linear de tal modo que

$$\mathrm{E}[\alpha_1 f_1(x) + \alpha_2 f_2(x)] = \alpha_1 \mathrm{E}[f_1(x)] + \alpha_2 \mathrm{E}[f_2(x)], \tag{6}$$

sendo α_1 e α_2 constantes arbitrárias.

Dois casos especiais de valores expectáveis são o segundo momento (ou *momento de 2ª ordem*) e a *variância*:

MOMENTO DE 2ª ORDEM

$$\mathrm{E}[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x) \tag{7}$$

VARIÂNCIA

$$\mathrm{var}(x) = \sigma^2 = \mathrm{E}[(x - \mu)^2] = \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x), \tag{8}$$

DESVIO-PADRÃO

onde σ_x é o *desvio-padrão* de x. A variância pode ser interpretada como o momento de inércia da função de massa de probabilidade. A variância nunca toma valores negativos e só pode ser nula se a função de massa está centrada num único ponto.

O valor do desvio-padrão é uma medida de dispersão dos valores de x à volta da média. O seu nome sugere que é a quantidade típica expectável que uma saída aleatória de x se desvie, ou difira, de μ. A *desigualdade de Chebyshev* relaciona o desvio-padrão e $|x - \mu|$:

DESIGUALDADE DE *CHEBYSHEV*

$$\Pr\left[|x - \mu| > k\sigma\right] \leq \frac{1}{k^2} \tag{9}$$

Esta desigualdade, independentemente da forma de $P(x)$, dá-nos o majorante do valor da probabilidade. Note que a desigualdade é inútil para $k < 1$. Por exemplo, para k=2, qualquer v.a. x toma valores entre μ–2σ e μ+2σ com probabilidade superior a 0.75. Uma regra de confiança mais prática, que é válida apenas para a distribuição normal (ou gaussiana), garante que 68% dos valores estão contidos no intervalo de um desvio-padrão à volta da média, 95% no intervalo de dois, e 99.7% no de três. Ou seja, quando a v.a. x tem distribuição normal, média μ e desvio-padrão σ (ver Figura 1, na secção 10), tem-se que

$$\Pr\left[|x - \mu| < \sigma\right] = 68\%, \quad \Pr\left[|x - \mu| < 2\sigma\right] = 95\%, \quad \Pr\left[|x - \mu| < 3\sigma\right] = 99.7\%. \tag{10}$$

Apesar do valor $1/k^2$ ser apenas o majorante de $\Pr\left[|x - \mu| > k\sigma\right]$, a desigualdade de Chebyshev mostra a forte ligação entre o desvio-padrão e a dispersão da função $P(x)$. Além disto, sugere que $|x - \mu|/\sigma$ é uma importante medida normalizada da distância de x em relação à média (ver

normalização na secção 10).
Expandindo o quadrado em (8), é fácil de verificar a fórmula

$$\sigma_x^2 = \mathrm{E}[x^2] - \mu_x^2 \tag{11}$$

Note que, ao contrário da média (valor expectável de x), a variância *não* é linear. Em particular, se $y = \alpha x$, onde α é uma constante, então $\mathrm{Var}[y] = \alpha^2 \mathrm{Var}[x]$. Além do mais, a variância da soma de duas v.a. não é igual à soma das variâncias, de um modo geral. No entanto, veremos mais à frente que as variâncias se somam quando as v.a. em causa são estatisticamente independentes.

No caso especial de x ser uma v.a. binária, que toma os valores $v_1 = 0$ e $v_2 = 1$, pode-se obter fórmulas simples para μ e σ. Sendo $p = \Pr[x = 1]$, demonstra-se que

$$\mu = p \qquad \text{e} \qquad \sigma = \sqrt{p(1-p)} \tag{12}$$

3. Probabilidade Conjunta

Sejam x e y duas v.a. que tomam valores em $\mathcal{X} = \{v_1, v_2, ..., v_m\}$ e $\mathcal{Y} = \{w_1, w_2, ..., w_n\}$, respectivamente. Considerando o par (x, y) como um vector do espaço $\mathbb{R}^2$, para cada possível par de valores (v_i, w_j) temos a *probabilidade conjunta* $p_{ij} = \Pr[x = v_i, y = w_j]$. Estas mn probabilidades conjuntas são representadas através da *função de massa de probabilidade conjunta* $P(x, y)$, tal que

PROBABILIDADE CONJUNTA

$$P(x, y) \geq 0 \qquad \text{e} \qquad \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \tag{13}$$

A função de massa de probabilidade conjunta representa completamente o par de v.a. (x, y); ou seja toda a informação das v.a. x e y, individualmente ou em conjunto, pode ser extraída de $P(x, y)$. Em particular, extraem-se as *distribuições marginais* de x e y, através da soma sobre a outra variável:

DISTRIBUIÇÃO MARGINAL

$$P_x(x) = \sum_{y \in \mathcal{Y}} P_{xy}(x, y)$$
$$P_y(x) = \sum_{x \in \mathcal{X}} P_{xy}(x, y) \tag{14}$$

Casualmente usam-se índices, como na eq. (14), para realçar o facto de $P_x(x)$ ter significado diferente de $P_y(y)$. É *comum* escrever simplesmente $P(x)$ e $P(y)$ quando o contexto torna claro que se trata de duas funções diferentes – e não a mesma função com diferentes valores de argumento. Usa-se a mesma regra para a média μ, o desvio-padrão σ e a variância σ^2, como em (11), para realçar o facto de serem medidas da v.a. x.

4. Independência Estatística

As variáveis x e y dizem-se estatisticamente independentes se e só se

$$P(x, y) = P_x(x) P_y(y) \,. \tag{15}$$

Pode-se entender a independência estatística do seguinte modo. Suponha que $p_i = \Pr[x = v_i]$ é fracção de tempo que $x = v_i$ e que $q_j = \Pr[y = w_j]$ é a fracção de tempo que $y = w_j$. Considere as situações em que $x = v_i$. Se continuar a ser verdade que a fracção de tempo em que $y = w_j$ tem o mesmo valor q_j, então conclui-se que conhecer o valor de x não trouxe informação adicional sobre os possíveis valores de y; neste sentido y é independente de x. Finalmente, se x e y são estaticamente independentes, é claro que a fracção de tempo que um específico par de valores (v_i, w_j) ocorre vem igual ao produto das duas fracções $p_i q_j = P(v_i) P(w_j)$ como exploraremos melhor na secção 6.

5. Valores Expectáveis de Funções de 2 Variáveis Aleatórias

Como expansão da secção 2, define-se o valor expectável da função $f(x,y)$ de duas v.a. x e y por

$$\mathrm{E}[f(x, y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x, y) P(x, y) \,, \tag{16}$$

e, como visto anteriormente, o operador valor expectável é linear:

$$\mathrm{E}[\alpha_1 f_1(x, y) + \alpha_2 f_2(x, y)] = \alpha_1 \mathrm{E}[f_1(x, y)] + \alpha_2 \mathrm{E}[f_2(x, y)] \,. \tag{17}$$

Os valores médios (momentos de 1ª ordem) e as variâncias (momentos de 2ª ordem) são:

$$\begin{aligned}
\mu_x &= \mathrm{E}[x] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xP(x,y) \\
\mu_y &= \mathrm{E}[y] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} yP(x,y) \\
\sigma_x^2 &= \mathrm{E}[(x-\mu_x)^2] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} (x-\mu_x)^2 P(x,y) \\
\sigma_y^2 &= \mathrm{E}[(y-\mu_y)^2] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} (y-\mu_y)^2 P(x,y).
\end{aligned} \tag{18}$$

COVARIÂNCIA

Outro valor expectável (momento cruzado) é designado por *covariância* de x e y:

$$\begin{aligned}
\mathrm{cov}(x,y) = \sigma_{xy} &= \mathrm{E}[(x-\mu_x)(y-\mu_y)] \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} (x-\mu_x)(y-\mu_y)P(x,y) \\
&= \mathrm{E}[xy] - \mu_x\mu_y .
\end{aligned} \tag{19}$$

Usando notação vectorial, pode-se abreviar as equações (18) e (19) como

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{x}] = \sum_{\mathbf{x}\in\{\mathcal{XY}\}} \mathbf{x}P(\mathbf{x}) \tag{20}$$

$$\boldsymbol{\Sigma} = \mathrm{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}], \tag{21}$$

onde $\{\mathcal{XY}\}$ representa o espaço de todos os valores possíveis para todas as componentes de $\mathbf{x}$ e $\boldsymbol{\Sigma}$ é a matriz de covariância (ver secção 11).

A covariância é uma medida da dependência estatística entre x e y. Se x e y são estatisticamente independentes, então $\sigma_{xy} = 0$. Se $\sigma_{xy} = 0$, as variáveis dizem-se *incorrelacionadas*.

INCORRELAÇÃO

Note que *incorrelação* entre variáveis não implica independência estatística – a covariância é apenas uma medida de dependência. No entanto, para v.a. com distribuição gaussiana (ver secção 10) é um facto que se forem incorrelacionadas então são estatisticamente independentes. Na prática, é comum tratar v.a. incorrelacionadas como se fossem independentes. Se α for uma constante e $y = \alpha x$, que é o caso de forte dependência estatística, é fácil verificar que $\sigma_{xy} = \alpha\,\sigma_x^2$. Assim, a covariância é positiva se x e y crescem ou decrescem conjuntamente, e negativa se y decresce quando x cresce.

DESIGUALDADE DE *CAUCHY-SCHWARZ*

A *desigualdade de Cauchy-Schwarz* relaciona as variâncias σ_x^2 e σ_y^2 com a covariância σ_{xy}. Pode ser derivada observando que a variância de uma v.a. nunca é negativa, logo a variância de $\lambda x + y$ não pode ser negativa, independentemente do valor de λ. Assim, temos a importante desigualdade

$$\sigma_{xy}^2 \le \sigma_x^2\sigma_y^2 \tag{22}$$

que é análoga à desigualdade vectorial $(\mathbf{x}^{\mathrm{T}}\mathbf{y})^2 \le \|\mathbf{x}\|^2\|\mathbf{y}\|^2$, frequentemente usada em álgebra linear.

COEFICIENTE DE CORRELAÇÃO

O *coeficiente de correlação*, definido como

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}, \tag{23}$$

é a covariância normalizada, e toma valores entre -1 e +1. Se ρ = +1, então x e y são correlacionadas positivamente ao máximo, enquanto que se ρ = -1, são correlacionadas negativamente ao máximo. Se $\rho = 0$, as v.a. são incorrelacionadas. Em casos práticos, é comum considerar as v.a. incorrelacionadas quando o coeficiente de correlação estiver abaixo de determinado valor, por exemplo 0.05, embora a escolha deste valor dependa do caso real.

Se x e y são estatisticamente independentes, então para quaisquer duas funções f e g obtemos

$$\mathrm{E}[f(x)g(y)] = \mathrm{E}[f(x)]\mathrm{E}[g(y)], \tag{24}$$

resultado que deriva da definição de independência estatística e operador valor expectável. Note que se $f(x) = x - \mu_x$ e $g(y) = y - \mu_y$, este teorema mostra que a covariância $\sigma_{xy} = \mathrm{E}[(x - \mu_x)(y - \mu_y)]$ é nula se x e y são estatisticamente independentes.

6. Probabilidade Condicional

Quando duas v.a. são estatisticamente dependentes, sabendo o valor de uma delas permite-nos obter uma melhor estimativa do valor da outra. Este conhecimento vem expresso pela seguinte definição de *probabilidade condicional* de x dado y:

$$\Pr[x = v_i \,|\, y = w_j] = \frac{\Pr[x = v_i, y = w_j]}{\Pr[y = w_j]}, \tag{25}$$

ou, em termos das funções de *massa de probabilidade*,

$$P(x|y) = \frac{P(x,y)}{P(y)}. \tag{26}$$

Note que, se x e y são estatisticamente independentes então $P(x|y) = P(x)$. Ou seja, quando x e y são independentes, conhecer o valor de y não nos fornece mais informação acerca de x além da que já tínhamos através da sua distribuição marginal $P(x)$.

Considere o exemplo de duas v.a. binárias x e y em que ambas tomam os valores 0 ou 1. Suponha que são produzidas aleatoriamente, em elevado número, n amostras dos pares (x,y). Seja n_{ij} o número de pares $(x=i, y=j)$, ou seja, n_{00} é o número de vezes que saiu o par (0,0), n_{10} é o número de vezes que saiu o par (1,0), e assim sucessivamente, tal que $n_{00}+n_{10}+n_{01}+n_{11}=n$. Se considerarmos apenas os pares em que $y=1$ - isto é, os pares (0,1) e (1,1) – então a fracção dos casos em x também vem igual a 1 é

$$\frac{n_{11}}{n_{01}+n_{11}} = \frac{n_{11}/n}{(n_{01}+n_{11})/n} \tag{27}$$

Intuitivamente, o valor desta fracção é o que gostaríamos de obter para $P(x|y)$ quando $y=1$ e n elevado. De facto, é o que se obtém pois n_{11}/n é aproximadamente igual a $P(x,y)$ e $(n_{01}+ n_{11})/n$ aproximadamente igual a $P(y)$ para valores elevados de n.

7. A Lei da Probabilidade Total e a Regra de Bayes

PROBABILIDADE TOTAL

A *Lei da Probabilidade Total* diz que se um acontecimento A ocorrer em m condições diferentes $A_1, A_2, ..., A_m$ e estes m sub-acontecimentos forem mutuamente exclusivos – ou seja, não ocorrem em simultâneo – então a probabilidade de ocorrer A é a soma das probabilidades dos sub-acontecimentos A_i. Em particular, a v.a. y pode tomar o valor y em m condições diferentes – com $x=v_1, x=v_2, ..., x=v_m$. Porque estas condições são mutuamente exclusivas, deduz-se da *Lei da Probabilidade Total* que $P(y)$ é a soma das probabilidades conjuntas $P(x,y)$ sobre todos os valores possíveis de x. Formalmente tem-se que

$$P(y) = \sum_{x \in \mathcal{X}} P(x,y). \tag{28}$$

Como, através da definição de probabilidade condicional $P(y|x)$, tem-se

$$P(x,y) = P(y \mid x)P(x), \tag{29}$$

então, rescrevendo a eq. (28), a $P(y)$ vem igual a

$$P(y) = \sum_{x \in \mathcal{X}} P(y \mid x)P(x). \tag{30}$$

Substituindo na eq. (26) as probabilidades $P(x,y)$ e $P(y)$, definidas respectivamente nas eqs. (28) e (30), vem

REGRA DE *BAYES*

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{\sum_{x \in \mathcal{X}} P(y \mid x)P(x)}. \tag{31}$$

Por outras palavras, tem-se

$$\text{posteriori} = \frac{\textit{likelihood} \times \text{priori}}{\text{evidência}}$$

onde estes termos são os usados na área de Reconhecimento de Padrões (ver Bibliografia) e que serão explicados em seguida.

A eq. (31) é designada por *Regra de Bayes*. Note que o denominador, que é $P(y)$, é obtido pela soma do numerador para todos os valores de x. Escrevendo o denominador desta forma dá-se ênfase ao facto de que todos os termos do lado direito da equação são condicionados por x. Se considerarmos x uma v.a. importante, então podemos dizer que a forma da distribuição $P(x|y)$ depende apenas do numerador $P(y|x)P(x)$; o denominador é o factor de normalização, por vezes designado por *evidência*, para garantir que a soma de $P(x|y)$ seja igual a um.

EVIDÊNCIA

A interpretação mais frequente da regra de *Bayes* é a de inverter ligações estatísticas, tornando $P(y|x)$ em $P(x|y)$. Considere que x é uma "causa" e y um "efeito" da causa x. Assumindo que a causa x está presente, é fácil de determinar a probabilidade do efeito y ser observado; a função de probabilidade condicional $P(y|x)$ – *função de verosemelhança* (*likelihood* em inglês) – representa esta probabilidade explicitamente. Ao contrário, se observarmos o efeito y, pode não ser tão fácil de determinar a causa x, pois haverá diferentes causas, podendo cada uma delas produzir o mesmo efeito observado. No entanto, a regra de *Bayes* torna fácil a determinação de $P(x|y)$, considerando que são conhecidas $P(y|x)$ e $P(x)$, designada por *probabilidade a priori* e que exprime a

VEROSEMELHANÇA

A PRIORI

DISTRIBUIÇÃO A POSTERIORI

probabilidade de x antes de observarmos qualquer valor de y. Ou seja, a regra de *Bayes* mostra como a distribuição de probabilidade de x se altera desde *distribuição a priori* $P(x)$, antes se observar y, até *distribuição a posteriori* $P(x|y)$, depois de se observar o valor de y.

8. Variáveis Aleatórias Contínuas

DENSIDADE DE PROBABILIDADE

Quando uma v.a x pode tomar valores no domínio contínuo (infinitos valores), não faz sentido falar da probabilidade de x ser igual a determinado valor, por exemplo Pr[x = 2.15], pois a probabilidade de um valor em particular é sempre nula (ou quase sempre). Assim, faz sentido falar da probabilidade de x tomar valores num determinado intervalo [a, b]; em vez de termos a função de massa de probabilidade $P(x)$, temos a *função de densidade de probabilidade* $p(x)$. Esta função tem a propriedade de

$$\Pr[x \in [a,b]] = \int_a^a p(x)dx\,. \tag{32}$$

O nome *densidade* vem da analogia a densidade de massa. Se considerarmos um intervalo pequeno [a, a+Δx] sobre o qual a função $p(x)$ é essencialmente constante, tendo o valor $p(a)$, vemos que $p(a) = \Pr[x \in [a,b]]/\Delta x$. Ou seja, a densidade de probabilidade em $x = a$ é a massa de probabilidade $\Pr[x \in [a,b]]$ por unidade de distância. Assim, a função densidade de probabilidade satisfaz as condições

$$p(x) \geq 0 \qquad \text{e} \qquad \int_{-\infty}^{+\infty} p(x)dx = 1\,. \tag{33}$$

De modo geral, a maioria das definições e das fórmulas para as v.a. discretas mantêm-se para as v.a. contínuas, com os somatórios substituídos por integrais. Em particular, o operador valor expectável, a média e a variância de uma v.a. contínua são definidos por

$$\begin{aligned} \mathrm{E}[f(x)] &= \int_{-\infty}^{+\infty} f(x)\,p(x)dx \\ \mu_x = \mathrm{E}[x] &= \int_{-\infty}^{+\infty} x\,p(x)dx \\ \mathrm{var}(x) = \sigma^2 = \mathrm{E}[(x-\sigma)^2] &= \int_{-\infty}^{+\infty} (x-\sigma)^2\,p(x)dx, \end{aligned} \tag{34}$$

e, como em (11), a variância verifica a igualdade $\sigma_x^2 = \mathrm{E}[x^2] - \mu_x^2$.

O caso M-dimensional (densidade multivariada) é semelhante para vectores de v.a. contínuas.

As funções densidade de probabilidade condicionais são definidas como as funções de massa de probabilidade condicionais. Assim, por exemplo, a densidade de x dado y é dada por

$$p(x \mid y) = \frac{p(x,y)}{p(y)} \tag{35}$$

e a regra de *Bayes* vem

$$p(x \mid y) = \frac{p(y \mid x)\,p(x)}{\int_{-\infty}^{+\infty} p(y \mid x)\,p(x)dx} \tag{36}$$

e o mesmo para o caso de vectores de v.a.

Ocasionalmente é necessário determinar o valor expectável em relação a um subconjunto de v.a., e neste caso usamos a notação de valor expectável com um índice, por exemplo

$$\mathrm{E}_{x_1}[f(x_1,x_2)] = \int_{-\infty}^{+\infty} f(x_1,x_2)\,p(x)dx\,. \tag{37}$$

9. Distribuição da Soma de Variáveis Aleatórias Independentes

Acontece frequentemente conhecermos as densidades de probabilidade de duas v.a. independentes x e y, e precisarmos de conhecer a função densidade de probabilidade da sua soma $z = x + y$. Para obter a média e a variância da soma, fazemos

$$\begin{aligned} \mu_z &= \mathrm{E}[z] = \mathrm{E}[x+y] = \mathrm{E}[x] + \mathrm{E}[y] = \mu_x + \mu_y, \\ \sigma_z^2 &= \mathrm{E}[(z-\mu_z)^2] = \mathrm{E}[(x+y-(\mu_x+\mu_y))^2] = \mathrm{E}[((x-\mu_x)+(y-\mu_y))^2] \\ &= \mathrm{E}[(x-\mu_x)^2] + 2\underbrace{\mathrm{E}[(x-\mu_x)(y-\mu_y)]}_{\mathrm{cov}(x,y)=0} + \mathrm{E}[(y-\mu_y)^2] \\ &= \sigma_x^2 + \sigma_y^2, \end{aligned} \tag{38}$$

onde usamos o conhecimento de independência das v.a. x e y na determinação da variância de z, ao anularmos o termo igual à cov(x,y) . Note que, se as v.a. x e y não forem independentes, este termo não se anula e variância de z não vem igual à soma das variâncias de x e y. Na determinação do valor médio de z não necessitamos de usar a independência de x e y para chegarmos à conclusão de que a média vem igual à soma das médias de x e y.

Para obter a função densidade de probabilidade de z a partir das densidades de x e y vamos analisar a probabilidade de z tomar valores no intervalo infinitesimal $[a, a+\Delta z]$,

$$\Pr[a < z < a + \Delta z] = \int_a^{a+\Delta z} p(z)dz = p(a)\Delta z \,, \tag{39}$$

que, sabendo que $z = x + y$, se obtém integrando a função conjunta $p(x,y)$ sobre os intervalos de valores de x e y tal que a soma esteja no intervalo $[a, a+\Delta z]$:

$$\Pr[a < x + y < a + \Delta z] = \Pr[a - x < y < a + \Delta z - x]$$

$$= \Pr[-\infty < x < +\infty, a - x < y < a - x + \Delta z] = \int_{-\infty}^{+\infty} \int_{a-x}^{a-x+\Delta z} \underbrace{p(x,y)}_{p_x(x)\cdot p_y(y)}\, dydx \tag{40}$$

$$= \int_{-\infty}^{+\infty} p_x(x) \underbrace{\int_{a-x}^{a-x+\Delta z} p_y(y)dy}_{p_y(a-x)\Delta z}\, dx = \left[\int_{-\infty}^{+\infty} p_x(x)\, p_y(a-x)dx\right]\Delta z.$$

CONVOLUÇÃO

Comparando os resultados obtidos em (39) e (40), conclui-se que o primeiro termo em (40) é igual a $p(a)$. Logo a função de densidade de probabilidade da soma $z = x + y$ é igual à *convolução* das funções de densidade de probabilidade das duas v.a. (desde que x e y sejam independentes):

$$p(z) = p_x(x) * p_y(y) = \int_{-\infty}^{+\infty} p_x(x)\, p_y(z-a)\, dx. \tag{41}$$

Estes resultados generalizam-se para a soma de N v.a. independentes, $x_1, x_2, ..., x_N$:

A média da soma é a soma das médias. (De facto, as v.a. não necessitam ser independentes para a soma das médias se verificar.)

A variância da soma é a soma das variâncias.

A função de densidade de probabilidade da soma é a convolução das densidades isoladas:

$$p(z) = p(x_1) * p(x_2) * ... * p(x_N) \tag{42}$$

10. Distribuição Normal

TEOREMA DO LIMITE CENTRAL

O *Teorema do Limite Central* diz que a forma da distribuição da soma de N v.a. independentes, com distribuições de probabilidade arbitrárias e não relevantes em relação às outras, no limite (com $N \to \infty$) aproxima-se da *distribuição normal*. Assim, a função de densidade de probabilidade *normal* (ou *gaussiana*) é muito importante, e utilizada em muitos casos, tanto na teoria como na prática. A uma dimensão, a função vem definida por

GAUSSIANA

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-(x-\mu)^2}{2\sigma^2}} \,. \tag{43}$$

A forma da função é completamente definida com dois parâmetros: os valores de μ (média) e de σ^2 (variância). Por isso, normalmente usa-se a notação $N(\mu,\sigma^2)$ que é lida como "x é uma v.a. normal (ou gaussiana) com média μ e variância σ^2". A distribuição é simétrica em relação à média, com o máximo em $x=\mu$ e largura proporcional ao desvio-padrão σ.

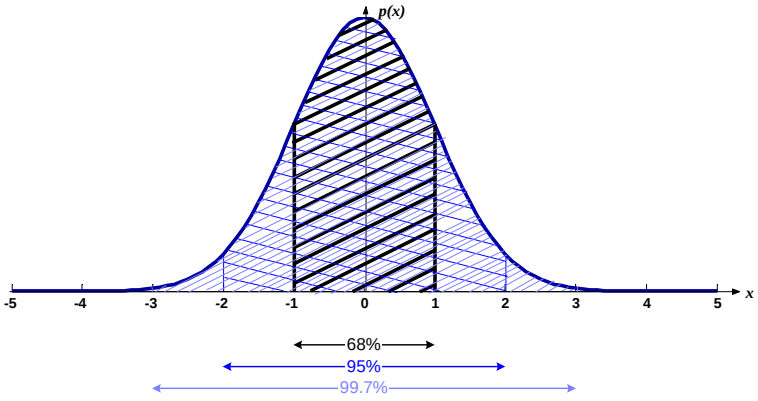


Figura 1. Distribuição normal $p(x) = N(0,1)$ com: 68% da massa de probabilidade no intervalo $|x| \le 1$; 95% no intervalo $|x| \le 2$; e 99.7% no intervalo $|x| \le 3$.

As amostras de uma v.a. com distribuição normal estão concentradas à volta da média conforme se observa na Figura 1. Como já se referiu anteriormente, a propósito da desigualdade de *Chebyshev* (secção 2, eq. (10)), quando a v.a. x tem distribuição normal, média μ e desvio-padrão σ, tem-se que

$$\begin{aligned} \Pr\left[|x-\mu| \le \sigma\right] &\simeq 0.68 \\ \Pr\left[|x-\mu| \le 2\sigma\right] &\simeq 0.95 \\ \Pr\left[|x-\mu| \le 3\sigma\right] &\simeq 0.997 , \end{aligned} \tag{44}$$

como se ilustra na Figura 1.

Uma medida natural da distância de um valor de x em relação à média μ é a distância $|\, x - \mu \,|$, medida em unidades de desvio-padrão σ:

$$k = \frac{|x-\mu|}{\sigma} , \tag{45}$$

DISTÂNCIA DE *MAHALANOBIS*

e que se designa pela *distância de Mahalanobis* de x a μ. Por exemplo, a probabilidade da distância de *Mahalanobis* de x a μ ser inferior a 2 é aproximadamente igual a 0.95, ou seja $\Pr\left[|x-\mu| \le k\sigma\right] \simeq 0.95$, com $k = 2$. As probabilidades em (44) representam a probabilidade de x se encontrar afastado da média μ (distância de *Mahalanobis* de x a μ), no máximo até 1, 2 e 3 unidades, respectivamente. Por isso, se modificarmos uma v.a. x, a) subtraindo a sua média e b) dividindo pelo seu desvio-padrão, diz-se que procedemos à *normalização* de x. Ou seja, uma v.a. gaussiana normalizada $u = (x - \mu) / \sigma$, tem média nula e desvio-padrão unitário. Isto é,

NORMALIZAÇÃO

$$p(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2} , \tag{46}$$

que se representa por $p(u) = N(0,1)$. Ver, na Figura 1, a função $p(x) = N(0,1)$ que representa a função densidade de probabilidade da v.a. gaussiana x que já se encontra normalizada. Para determinar a probabilidade de u tomar valores superiores a determinado valor k sabemos que:

$$\Pr[u > k] = \int_k^{+\infty} p(u)du = \frac{1}{\sqrt{2\pi}} \int_k^{+\infty} \mathrm{e}^{-u^2/2}\, du \tag{47}$$

Como a função $\mathrm{e}^{-u^2/2}$ não tem primitiva, para determinar o integral em (47), normalmente usam-se tabelas, aproximações ou integração numérica. A Tabela 1 mostra os valores deste integral, para vários valores de k.

k	Q(k)	k	Q(k)	k	Q(k)
0.0	0.50000000	1.0	0.15865525	2.0	0.02275013
0.1	0.46017216	1.1	0.13566606	2.1	0.01786442
0.2	0.42074029	1.2	0.11506967	2.3	0.01072411
0.3	0.38208858	1.3	0.09680048	2.5	0.00620967
0.4	0.34457826	1.4	0.08075666	3.0	0.00134990
0.5	0.30853754	1.5	0.06680720	3.3	0.00048342
0.6	0.27425312	1.6	0.05479929	3.5	0.00023263
0.7	0.24196365	1.7	0.04456546	4.0	0.00003167
0.8	0.21185540	1.8	0.03593032	5.0	0.00000029
0.9	0.18406013	1.9	0.02871656	5.5	0.00000002

Tabela 1. Valores da função Q(k)

FUNÇÃO Q(K)

Esta função que designamos por *função* Q(k),

$$\mathrm{Q}(k) = \frac{1}{\sqrt{2\pi}} \int_k^{+\infty} \mathrm{e}^{-u^2/2}\, du = \Pr[u > k] = \Pr[u < k] , \tag{48}$$

pode ser obtida através de outra função, normalmente mais conhecida, definida como

$$\mathrm{erf}(u) = \frac{2}{\sqrt{\pi}} \int_k^{+\infty} \mathrm{e}^{-x^2/2}\, dx = 1 - \mathrm{erfc}(u) \tag{49}$$

ERROR FUNCTION

e designada por *error function* -ver Figura 2. Esta função relaciona-se com a função Q(k) através da sua função complementar, tal que:

$$\mathrm{Q}(k) = \frac{1}{2} \mathrm{erfc}\left(\frac{k}{\sqrt{2}} \right) . \tag{50}$$

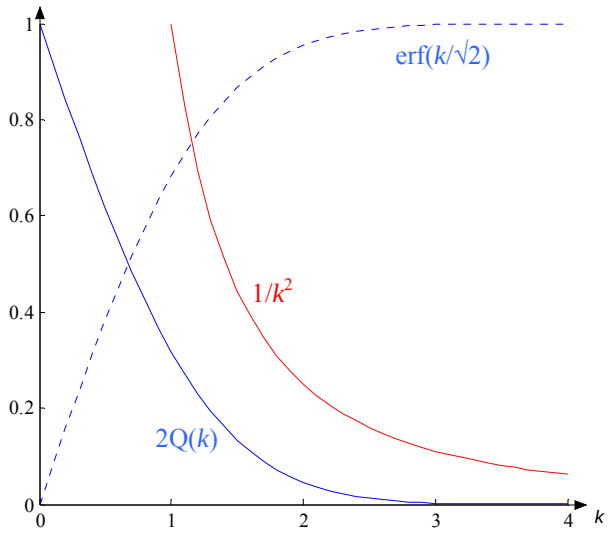


Figura 2. A função erf($k/\sqrt{2}$) corresponde à área de uma gaussiana normalizada entre $-k$ e k; ou seja, se x for uma v.a gaussiana normalizada, $\Pr[|x|\leq k] = \text{erf}(k/\sqrt{2})$. Assim, a probabilidade complementar, $1 - \text{erf}(k/\sqrt{2}) = 2Q(k)$, é a probabilidade de x, em módulo, tomar valores superiores a k. A desigualdade de *Chebyshev* diz que, para uma distribuição arbitrária com média nula e desvio-padrão unitário, a probabilidade $\Pr[|x|<k]$ é menor que $1/k^2$, por isso a curva $2Q(k)$ é limitada pela curva $1/k^2$. Como se observa, este limite é débil para a distribuição gaussiana.

Assim, tirando partido da normalização de uma v.a. e da distância de *Mahalanobis*, através da função $Q(k)$ podemos determinar a probabilidade de uma v.a. x tomar valores num determinado intervalo definido em função da distância de *Mahalanobis* de x em relação à média μ, definida em (45), tal que

$$\Pr[x > \mu + k\sigma] = \Pr[x < \mu - k\sigma] = \Pr[|x-u| > k\sigma] = Q(k) \tag{51}$$

e

$$\Pr[\mu - k\sigma < x < \mu + k\mu] = \Pr[|x-\mu| < k\sigma] = 1 - 2\,Q(k)\,. \tag{52}$$

11. Vectores de Variáveis Aleatórias

Para expandir os resultados de duas v.a. x e y para M variáveis $x_1, x_2, ..., x_M$, é conveniente usar a notação vectorial, como já o fizemos nas equações (20) e (21). A função de probabilidade conjunta $P(\mathbf{x})$ satisfaz as condições $P(\mathbf{x}) \geq 0$ e $\sum P(\mathbf{x}) = 1$, como em (13), onde o somatório é expandido para todos os valores possíveis do vector $\mathbf{x}$. Note que $P(\mathbf{x})$ é função de M variáveis, ou seja é uma função multi-dimensional. No entanto, se as v.a. x_i forem estatisticamente independentes, reduz-se ao produto

$$P(x) = P_{x_1}(x_1)P_{x_2}(x_2)...P_{x_M}(x_M) = \prod_{i=1}^{M} P_{x_i}(x_i)\,, \tag{53}$$

onde usámos os índices para enfatizar o facto de as distribuições marginais terem formas diferentes, de modo geral. As distribuições marginais $P_{xi}(x_i)$ podem ser obtidas através da soma da distribuição conjunta sobre as outras v.a., como em (14). Além destas marginais univariadas, outras distribuições marginais podem ser obtidas usando a Lei da Probabilidade Total. Por exemplo, se tivermos $P(x_1, x_2, x_3, x_4, x_5)$ e quisermos $P(x_1, x_4)$, fazemos

$$P(x_1, x_4) = \sum_{x_2}\sum_{x_3}\sum_{x_5} P(x_1, x_2, x_3, x_4, x_5)\,. \tag{54}$$

Definem-se várias distribuições condicionais, como $P(x_1, x_2 \mid x_3)$ ou $P(x_2 \mid x_1, x_4, x_5)$. Por exemplo,

$$P(x_1, x_2 \mid x_3) = \frac{P(x_1, x_2, x_3)}{P(x_3)}\,, \tag{55}$$

onde todas as distribuições conjuntas podem ser obtidas de $P(\mathbf{x})$ através da soma sobre todas as outras variáveis não pretendidas. Se cada uma das v.a. não forem escalares mas vectores de v.a., então estas distribuições podem ser escritas como

$$P(\mathbf{x}_1 \mid \mathbf{x}_2) = \frac{P(\mathbf{x}_1, \mathbf{x}_2)}{P(\mathbf{x}_2)}\,, \tag{56}$$

e do mesmo modo, em forma vectorial, a regra de *Bayes* vem

$$P(\mathbf{x}_1 \mid \mathbf{x}_2) = \frac{P(\mathbf{x}_2 \mid \mathbf{x}_1)P(\mathbf{x}_1)}{\sum_{\mathbf{x}_1} P(\mathbf{x}_2 \mid \mathbf{x}_1)P(\mathbf{x}_1)} \,. \tag{57}$$

12. Valores Expectáveis, Vectores de Média e Matrizes de Covariância

O valor expectável de um vector de v.a. é definido como um vector cujas componentes são os valores expectáveis das componentes do vector. Assim, se **f**(**x**) é um vector cujas componentes são funções do vector aleatório M-dimensional **x**,

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{bmatrix}, \tag{58}$$

então o valor expectável de **f** é definido por

$$\mathrm{E}[\mathbf{f}] = \begin{bmatrix} \mathrm{E}[f_1(\mathbf{x})] \\ \mathrm{E}[f_2(\mathbf{x})] \\ \vdots \\ \mathrm{E}[f_M(\mathbf{x})] \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{f}(\mathbf{x})P(\mathbf{x}) \,. \tag{59}$$

VECTOR DE MÉDIA

Em particular, o vector M-dimensional designado por *vector de média* **μ** é definido por

$$\mathrm{E}[\mathbf{x}] = \begin{bmatrix} \mathrm{E}[x_1] \\ \mathrm{E}[x_2] \\ \vdots \\ \mathrm{E}[x_M] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_M \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{x}\, P(\mathbf{x}) \tag{60}$$

MATRIZ DE COVARIÂNCIA

Do mesmo modo, a *matriz de covariância* **Σ** é definida como a matriz quadrada cujo elemento genérico σ_{ij} é a covariância de x_i e x_j:

$$\sigma_{ij} = \mathrm{cov}(x_i, x_j) = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)] \qquad i, j = 1 \ldots M \,, \tag{61}$$

como vimos no caso de duas variáveis em (19). Assim, na sua forma expandida, a matriz de covariância vem

$$\begin{aligned} \mathbf{\Sigma} &= \begin{bmatrix} \mathrm{E}[(x_1 - \mu_1)(x_1 - \mu_1)] & \mathrm{E}[(x_1 - \mu_1)(x_2 - \mu_2)] & \cdots & \mathrm{E}[(x_1 - \mu_1)(x_M - \mu_M)] \\ \mathrm{E}[(x_2 - \mu_2)(x_1 - \mu_1)] & \mathrm{E}[(x_2 - \mu_2)(x_2 - \mu_2)] & \cdots & \mathrm{E}[(x_2 - \mu_2)(x_M - \mu_M)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(x_M - \mu_M)(x_1 - \mu_1)] & \mathrm{E}[(x_M - \mu_M)(x_2 - \mu_2)] & \cdots & \mathrm{E}[(x_M - \mu_M)(x_M - \mu_M)] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_M^2 \end{bmatrix}. \end{aligned} \tag{62}$$

Usa-se o produto matricial $(\mathbf{x} - \boldsymbol{\mu})\,(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}$ para escrever a matriz de covariância como

$$\mathbf{\Sigma} = \mathrm{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\right] \,. \tag{63}$$

A matriz Σ é simétrica e os elementos da diagonal são as variâncias das v.a. do vector **x**, que nunca podem ser negativas; os elementos fora da diagonal são as covariâncias, que podem ser positivas ou negativas. Se as variáveis são estatisticamente independentes, as covariâncias são nulas e a matriz de covariância é diagonal.

13. Distribuição Gaussiana Multivariada

A forma geral da distribuição normal multivariada, *M* dimensional, vem escrita como

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{M/2}\,|\mathbf{\Sigma}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{64}$$

onde **x** é um vector coluna com *M* componentes, **μ** é o *vector de média* com *M* componentes, Σ é a *matriz de covariância* de dimensão *M* por *M*, e $|\Sigma|$ e Σ^{-1} são o determinante e a matriz inversa, respectivamente. Igualmente, $(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}$ denota a transporta de $(\mathbf{x} - \boldsymbol{\mu})$. Por simplicidade, a equação (64) é frequentemente abreviada como $p(\mathbf{x}) = N(\boldsymbol{\mu},\mathbf{\Sigma})$.

Note que, se os elementos σ_{ij} da matriz de covariancia forem nulos, exceptuando os da diagonal, significa que as componentes de **x** são incorrelacionadas e a expressão de $p(\mathbf{x})$ em (64) reduz-se

ao produto das distribuições univariadas das componentes de **x** - que significa que estas são estatisticamente independentes (ver (15)).

NORMAL BIVARIADA

Um caso particular da distribuição normal multivariada, quando **x** é um vector coluna com duas componentes x_1 e x_2, é a *normal bivariada*,

$$p(\mathbf{x}) = \frac{1}{2\pi |\boldsymbol{\Sigma}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{65}$$

onde

$$|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}$$

$$\boldsymbol{\Sigma}^{-1} = \frac{\mathrm{adj}\,\boldsymbol{\Sigma}}{|\boldsymbol{\Sigma}|} = \frac{1}{|\boldsymbol{\Sigma}|}\begin{bmatrix} \sigma_{22} & -\sigma_{21} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}^{\mathrm{T}} = \frac{1}{|\boldsymbol{\Sigma}|}\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix}. \tag{66}$$

Neste caso, podemos visualizar a função na Figura 3, atribuindo valores para o vector de média $\boldsymbol{\mu}$ e a matriz de covariância $\boldsymbol{\Sigma}$.

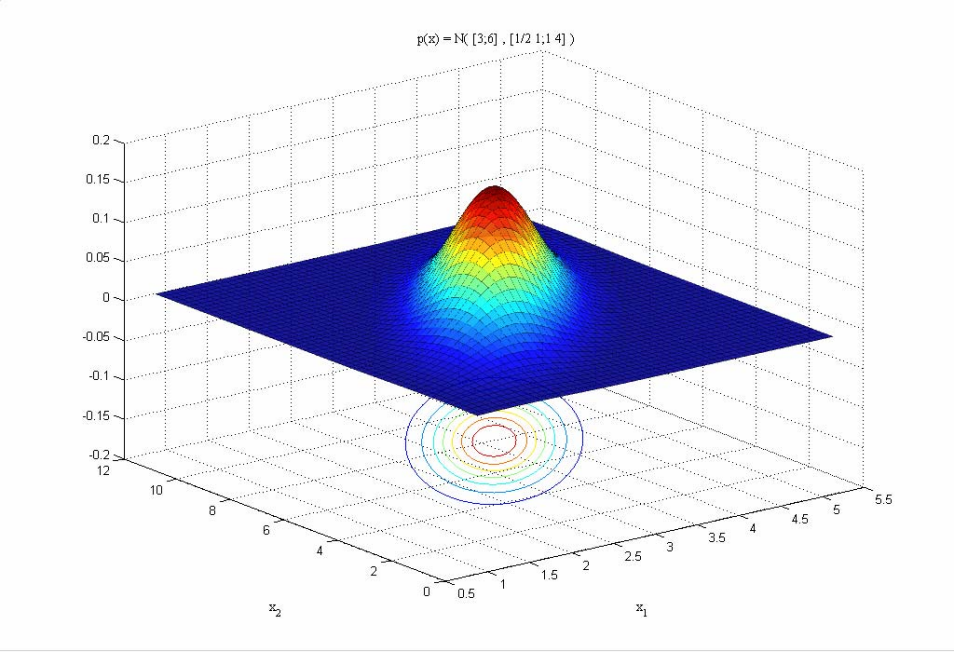


Figura 3. Distribuição normal bivariada: $p(x) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ com $\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$ e $\boldsymbol{\Sigma} = \begin{bmatrix} 1/2 & 1 \\ 1 & 4 \end{bmatrix}$.

Na Figura 4 observam-se as curvas de nível da função de distribuição $p(\mathbf{x})$ da Figura 3 e que podem ser determinadas fazendo $p(\mathbf{x}) = n_i$, onde n_i é a constante respectiva para cada nível.

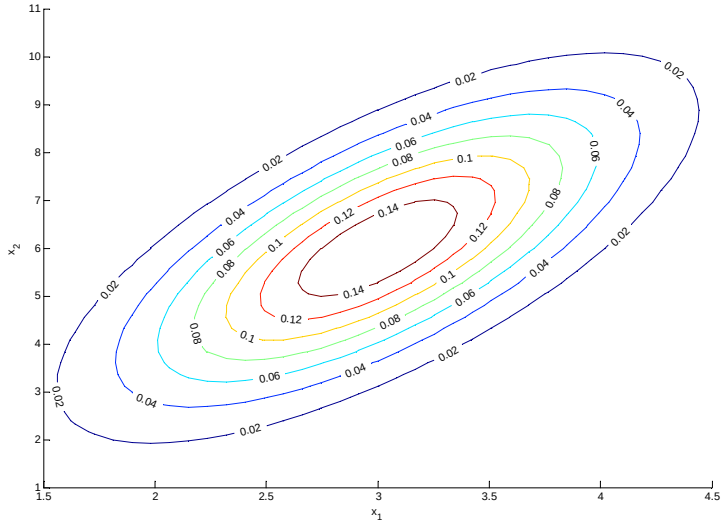


Figura 4. Curvas de nível da função $p(x) = N(\mu, \Sigma)$ com $\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$ e $\boldsymbol{\Sigma} = \begin{bmatrix} 1/2 & 1 \\ 1 & 4 \end{bmatrix}$.

Bibliografia

[1] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification, 2nd edition*, John Wiley & Sons, 2001.

[2] J. S. Marques, *Reconhecimento de Padrões: Métodos Estatísticos e Neuronais*, IST Press, 1999.

[3] V. Barroso, *Sinais Aleatórios em Tempo Contínuo. Parte I: Espaço de Probabilidade e Variáveis Aleatórias*, Folhas de Apoio, IST, 1999.

[4] A. B. Carlson, P. B. Crilly, J. C. Rutledge, *Communication Systems*, , *4th edition*, McGraw-Hill, 2002.