

Mestrado em Engenharia de Sistemas e Computação

Processamento Estatístico do Sinal

Sérgio M.M. Jesus

sjesus@ualg.pt

Unidade de Ciências Exactas e Humanas

Universidade do Algarve

1995/96

Última revisão: Setembro 1999

(<ftp://ftp.ualg.pt/sjesus/pub/pes99.ps>)

Programa

1.	Introdução e objectivos	4
2.	Variáveis aleatórias bidimensionais	7
2.1	Definição	7
2.2	Função de distribuição	7
2.3	Densidade de probabilidade	8
2.4	Esperança matemática	9
2.5	Distribuições condicionais	10
2.6	Esperanças condicionais	11
2.7	Variáveis aleatórias discretas	12
2.8	Momentos de segunda ordem: covariância e correlação	12
3.	Vectores aleatórios	15
3.1	Função de distribuição e densidade de probabilidade	15
3.2	Esperança matemática	15
3.3	Matrizes de covariância	16
3.4	Transformações lineares de vectores aleatórios	17
3.5	Distribuições condicionais	19
3.6	Esperança matemática condicional	19
3.7	Funções de vectores aleatórios	21
3.8	Funções características e momentos	22
4.	Vectores aleatórios gaussianos	23
4.1	Teorema central limite	23
4.2	Definição	24
4.3	Propriedades	24
4.4	Transformações lineares	25
4.5	Distribuições condicionais e regressão	25
4.6	Vectores aleatórios gaussianos complexos	27
4.7	Formas quadráticas de VA's normais multivariadas	29
5.	Elementos de álgebra linear	31
5.1	Espaços vectoriais	31
5.2	Independência linear	32
5.3	Sub-espacos vectoriais lineares, bases, dimensões e ordem	32
5.4	Matrizes hermitianas	33
5.5	Decomposição em valores singulares	34
5.6	Projeções, rotações e pseudoinversos	35
5.7	Formas quadráticas	38

6.	Mínimos quadrados (MQ)	39
6.1	O modelo linear	39
6.2	Soluções dos MQ	40
6.3	Decomposição em valores singulares e problemas de MQ	43
6.4	Algoritmos resolventes dos MQ	44
6.4.1	<i>Algoritmo de factorização de Cholesky</i>	44
6.4.2	<i>Algoritmo QR</i>	45
6.4.3	<i>Decomposição em valores singulares</i>	45
6.5	Desempenho dos MQ	46
6.6	Determinação e redução da ordem	47
6.7	MQ recursivos	50
6.8	MQ ponderados	51
6.9	MQ condicionados e condicionamento linear	51
6.10	MQ totais	53
6.11	Problemas inversos e MQ sub-determinados	54
6.11.1	<i>Solução de norma mínima</i>	55
6.11.2	<i>Solução de ordem reduzida</i>	56
6.11.3	<i>Solução do máximo de entropia</i>	56
6.12	Problemas de identificação	56
6.12.1	<i>Identificação do modelo ARMA</i>	57
6.13	Exemplos	58
6.13.1	<i>Modelo de regressão polinomial</i>	58
6.13.2	<i>Modelo de exponenciais complexas</i>	59
6.13.3	<i>Representação de Fourier discreta</i>	59
	Referências	61
	Exames	62

NOTA PRÉVIA

O material contido neste conjunto de apontamentos é cedido a título gratuito e para ser utilizado exclusivamente como texto de apoio à disciplina de Processamento Estatístico do Sinal do curso de Mestrado em Engenharia de Sistemas e Computação da Universidade do Algarve. O autor não pretende de forma alguma que o texto esteja privado de erros involuntários * de cujas consequências não poderá ser responsabilizado. Em caso de perdas ou prejuízos, directos ou indirectos, causados pelo uso deste texto o autor concede uma indemnização igual ao custo do próprio texto, i.e., zero escudos. A sua consulta não dispensa (e aliás aconselha) a de outras obras, nomeadamente as citadas na bibliografia. Boa leitura !...

* Agradecem-se correções e comentários para sjesus@ualg.pt: em troca oferece-se uma cópia electrónica gratuita devidamente zipada !

1. Introdução e objectivos

Transmissão de informação implica troca de energia entre dois pontos. Trata-se, na maior parte dos casos, de um processo que se encontra frequentemente na vida do dia a dia, sendo o mecanismo microscópico de natureza físico-química. Uma descrição matemática do processo de transmissão da informação permite o seu estudo, independentemente do processo de suporte físico-químico, concentrando-se naquilo que hoje se chama *Teoria da Informação*. Os pioneiros da Teoria da Informação foram sem dúvida Shannon e Kolmogorov nos anos 40 e 50. Já desde essa altura foi introduzida a noção de *senal* como suporte da informação e a de *ruído* como perturbação da informação. Nos anos 50, com o aparecimento do transistor, e mais tarde nos anos 60 com o aparecimento dos circuitos integrados e dos computadores electrónicos, a teoria do sinal transformou-se no processamento do sinal e mais tarde no processamento digital do sinal, que não é mais do que um conjunto de métodos e técnicas para emitir, transmitir e receber o máximo da informação eliminando, o mais possível o ruído.

É nesta óptica que se colocam os conceitos de que falaremos nesta disciplina que incluem: filtragem, predição, estimação de sinais em ruído, análise espectral, etc... Os métodos óptimos serão delimitados e as suas implementações práticas serão deduzidas. O melhor uso do conhecimento a priori do sinal a filtrar ou a analisar será determinado.

O material apresentado segue normalmente o programa de formação matemática de base de um engenheiro incluindo probabilidades e estatística e álgebra elementar e necessita de um bom conhecimento introdutório á teoria do sinal e sistemas, tanto no caso contínuo como no caso discreto. Gostaria de começar através de um exemplo típico das noções ou tipos de problemas aos quais tentaremos de responder durante este estudo.

Exemplo

Considere o sinal $x_t, t = 0, \dots, N - 1$ que representa uma série temporal escalar que contém uma função coseno do tempo discreto t a uma frequência angular ω ,

$$x_t = s_t + n_t \tag{1.1a}$$

$$s_t = \cos \omega t = \frac{1}{2}(e^{j\omega t} + e^{-j\omega t}) \tag{1.1b}$$

a sequência s_t encontra-se representada na figura 1.1.

Podemos muito naturalmente acrescentar zeros no sinal x_t sem por isso alterar a informação nele contida, obtendo assim

$$x_t = \begin{cases} s_t + n_t, & t = 0, 1, \dots, N - 1 \\ 0, & t = N, \dots, M - 1. \end{cases} \tag{1.2}$$

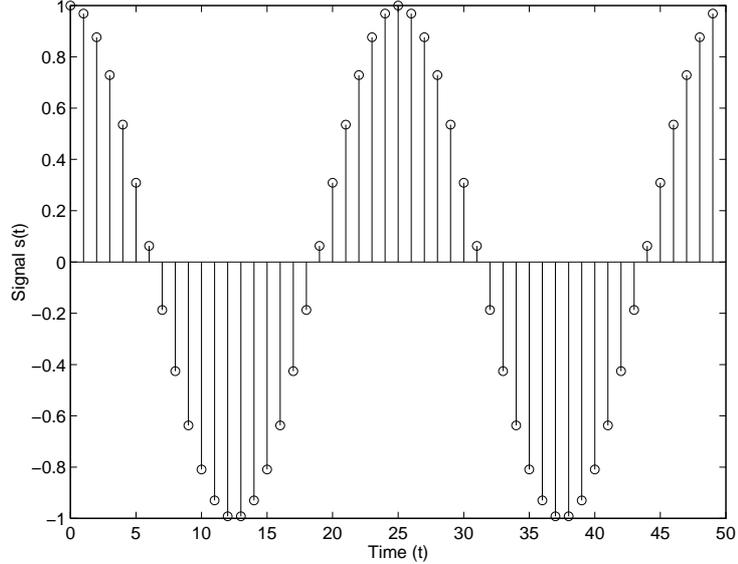


Figura 1.1: função cosinusoidal discreta

A transformada de Fourier discreta (TFD) de x_t permite obter,

$$X_m = \sum_{t=0}^{N-1} x_t e^{-j2\pi mt/M}, \quad m = 0, 1, \dots, M-1 \quad (1.3)$$

substituindo (1.1b) em (1.3) pode-se obter a componente do sinal no domínio da frequência, que se escreve

$$S_m = \frac{1}{2} \sum_{t=0}^{N-1} e^{jt(\omega - 2\pi m/M)} + \frac{1}{2} \sum_{t=0}^{N-1} e^{-jt(\omega + 2\pi m/M)} \quad (1.4)$$

usando a expressão de uma soma finita de termos em progressão geométrica podemos transformar os dois somatórios em quocientes de senos e rescrever o sinal S_m como

$$S(\phi) = \frac{1}{2} e^{j[(N-1)/2](\omega - \phi)} \frac{\sin N(\omega - \phi)}{\sin(\omega - \phi)} + \frac{1}{2} e^{-j[(N-1)/2](\omega + \phi)} \frac{\sin N(\omega + \phi)}{\sin(\omega + \phi)} \quad (1.5)$$

com $\phi = 2\pi(m/M)$ e que podemos representar como exemplificado na figura 1.2.

O número de pontos no pico principal da função é igual ao maior número inteiro inferior ou igual a M/N . Podemos então escolher M suficientemente elevado de modo a poder encontrar m tal que $2\pi m/M$ seja arbitrariamente o mais próximo de ω . Ao mesmo tempo S_m para esse valor de ω é arbitrariamente próximo de $N/2$. Este facto dá-nos a impressão de que, apenas juntando um número elevado de zeros na sequência temporal nos permite estimar com grande precisão o valor de ω através de uma simples observação da sua TFD. Porém, observando mais atentamente, podemos ver que um aumento de M faz com que encontremos um elevado número de valores de S_m cujo valor é próximo de $N/2$, o que indica que pequenas perturbações de ruído farão “saltar” o máximo para os valores vizinhos do verdadeiro valor de ω , mantendo assim uma incerteza na estimativa. Pode-se argumentar mesmo, que o facto de aumentar a resolução numérica da função S_m através

de uma extensão da observação temporal por zeros não reduz a ambiguidade no pico principal da função.

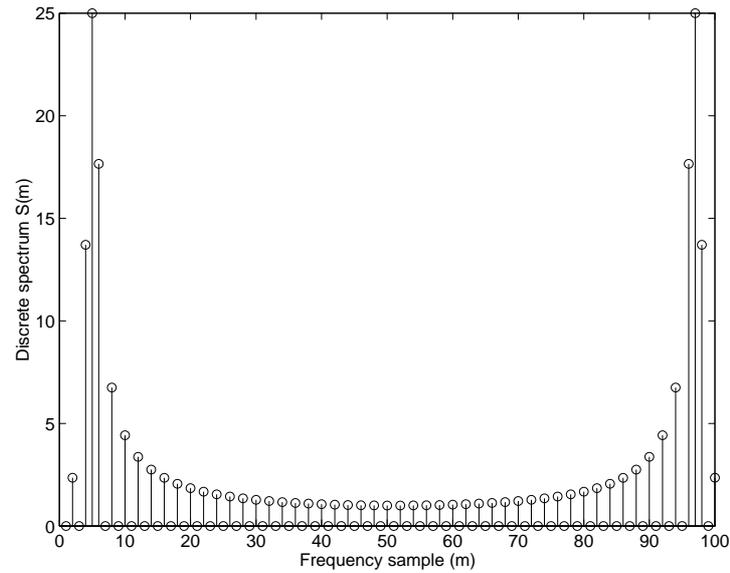


Figura 1.2: TFD da função cosinusoidal discreta

Esta ambiguidade só poderá ser reduzida através do aumento efectivo do intervalo de N amostras original ou através de métodos mais sofisticados. A ambiguidade inerente ao comprimento do sinal $2\pi/N$ é chamado *limite de resolução de Rayleigh* e é fundamental em teoria do sinal.

2. Variáveis aleatórias bidimensionais

2.1 Definição

Considere-se um par de variáveis aleatórias (VA's) X e Y que podem ser tomadas como as coordenadas de um vector aleatório \mathbf{V} no espaço R^2 . Este vector e as respectivas coordenadas estão representadas na fig. 2.1.

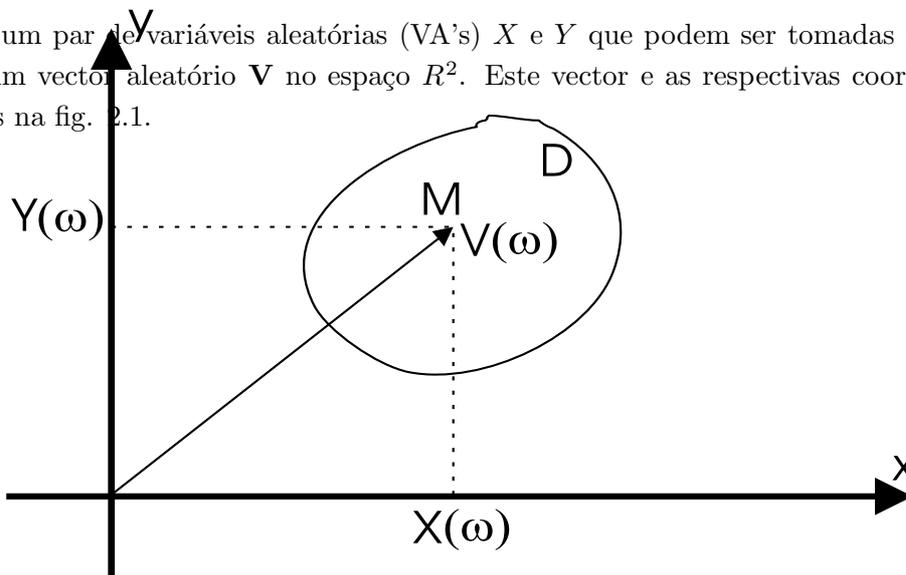


Figura 2.1: VA bidimensional.

Nesta figura também está representado um domínio D . Podemos considerar um acontecimento, $A(D)$, o facto do ponto M , extremidade de \mathbf{V} , pertencer a D . A este acontecimento é possível associar uma probabilidade $P(D)$, que é a medida de probabilidade associada com o acontecimento $A(D)$

$$P(D) = Prob\{\omega | M(\omega) \in D\} \quad (2.1)$$

da mesma forma podemos definir o conjunto de pontos

$$A(x, y) = \{\omega | X(\omega) \leq x \text{ e } Y(\omega) \leq y\} \quad (2.2)$$

A probabilidade associada com o acontecimento de o ponto M pertencer ao conjunto de pontos definido em (2.2) constitui a base das VA bidimensionais a partir do qual a extensão ao caso a n dimensões se obtém sem dificuldade.

2.2 Função de distribuição

Já vimos anteriormente que as probabilidades associadas aos acontecimentos

$$\{X \leq x\} \quad \text{e} \quad \{Y \leq y\} \quad (2.3)$$

nos dão as funções de distribuição $F(x)$ e $F(y)$ das VAs X e Y respectivamente. Estas são chamadas as funções de distribuição marginais por oposição à distribuição bidimensional que vamos considerar

já a seguir. A função de distribuição bidimensional (ou conjunta) das VA's X e Y define-se como a probabilidade do conjunto $A(D)$ definido em (2.2) e que não é mais do que a probabilidade do acontecimento intersecção dos dois acontecimentos elementares em (2.3) que se escreve

$$\{X \leq x \text{ e } Y \leq y\} = \{X \leq x\} \cap \{Y \leq y\} \quad (2.4)$$

Assim

$$F(x, y) = Prob\{X \leq x \text{ e } Y \leq y\} \quad (2.5)$$

é a função de distribuição conjunta das VA's X, Y . (2.5) goza das mesmas propriedades que a função de distribuição unidimensional. Em particular, trata-se de uma função monótona não decrescente de x e de y e, se x e y tenderem para infinito, $F(x, y)$ tende para um. De um modo geral $F(x, y)$ é a probabilidade associada com o domínio D definido na fig 2.1 e podemos assim calcular a probabilidade associada com qualquer domínio arbitrário Δ do plano através de

$$P(\Delta) = \int \int_{\Delta} d^2 F(x, y) \quad (2.6)$$

Se Δ for todo o plano $F(+\infty, +\infty) = 1$. Podemos definir a função de distribuição marginal a partir da função de distribuição conjunta, assim

$$F(x) = F(x, +\infty) \quad F(y) = F(+\infty, y) \quad (2.7)$$

2.3 Densidade de probabilidade

A VA bidimensional (X, Y) é contínua se $F(x, y)$ admite uma derivada ou se pudermos escrever

$$p(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} \quad (2.8)$$

que é a função densidade de probabilidade de (X, Y) . Nesse caso (2.6) escreve-se

$$P(\Delta) = \int \int_{\Delta} p(x, y) dx dy \quad (2.9)$$

e do mesmo modo podemos escrever as funções de distribuição marginais, sabendo que,

$$F(+\infty) = \int_{-\infty}^{+\infty} p(u) du$$

como

$$F(x) = \int_{-\infty}^x \int_{-\infty}^{+\infty} p(u, v) dudv \quad F(y) = \int_{-\infty}^y \int_{-\infty}^{+\infty} p(u, v) dudv \quad (2.10)$$

Derivando ambas as expressões em relação a x obtemos as densidades de probabilidade marginais

$$p(x) = \int p(x, y) dy \quad p(y) = \int p(x, y) dx \quad (2.11)$$

As relações obtidas para VA's contínuas poderiam ter sido demonstradas, com as evidentes modificações, para VA's discretas.

2.4 Esperança matemática

O caso bidimensional pode ser visto, para efeitos do cálculo da esperança matemática, como uma função de duas VA's. Assim se no caso unidimensional escrevíamos

$$E[g(X)] = \int g(x)p(x)dx \quad (2.12)$$

poderemos escrever no caso de uma função h de duas VA's X, Y ,

$$E[h(X, Y)] = \int \int h(x, y)p(x, y)dxdy \quad (2.13)$$

e em particular para calcularmos a esperança matemática de uma função de uma só VA então podemos dizer que

$$E[h(X)] = \int h(x)p(x, y)dx \quad (2.14)$$

Se por exemplo tivermos $h(x, y) = x + y$ então

$$E[X + Y] = E[X] + E[Y] \quad (2.15)$$

o que significa que a esperança matemática da soma de duas VAs é igual á soma das esperanças matemáticas de cada uma das VAs independentemente do facto de elas serem ou não independentes. No entanto se $h(x, y) = xy$ então usando (2.13) temos que em geral

$$E[XY] \neq E[X]E[Y] \quad (2.16)$$

salvo no caso em que as duas integrais são separáveis, ou seja, se

$$p(x, y) = p(x)p(y) \quad (2.17)$$

que só se verifica se $F(x, y) = F(x)F(y)$ e que por sua vez depende do facto dos dois acontecimentos em (2.3) serem independentes. No caso de VAs independentes (2.16) não será verdadeira pois teremos

$$E[XY] = E[X]E[Y] \quad (2.18)$$

O inverso não é verdade pois poderemos ter VA's dependentes para as quais (2.18) é verdadeira. Assim introduziu-se o termo de não correlação para o caso de VA's que admitem (2.18). Consequentemente duas VA's independentes são não correladas mas duas VA's não correladas podem ser dependentes.

2.5 Distribuições condicionais

Recordemos que a probabilidade do acontecimento $\{A \text{ se } B\}$ é dada pela relação fundamental

$$P(A/B) = \frac{P(AB)}{P(B)} \quad (2.19)$$

para $P(B) \neq 0$. Em particular, se A for relacionado com X , por exemplo, $A = \{X \leq x\}$ então a função de distribuição condicional escreve-se

$$F(x/B) = Prob\{X \leq x|B\} \quad (2.20)$$

que é a distribuição de X condicionada ao acontecimento B . Se $F(x/B)$ for diferenciável em todo o domínio, então podemos definir a densidade de probabilidade condicional

$$p(x/B) = \frac{dF(x/B)}{dx} \quad (2.21)$$

Da mesma maneira podemos também definir a esperança condicional

$$E[h(X)/B] = \int h(x)p(x/B)dx \quad (2.22)$$

a partir da qual, substituindo $h(x)$ por x^k , podemos obter os momentos condicionais de ordem k . Finalmente, podemos igualmente introduzir a função característica condicional substituindo em (2.22) $h(x) = \exp(jux)$ que goza de todas as propriedades já citadas para o caso não condicional.

No caso de duas VAs (X, Y) , podemos introduzir a distribuição condicional conjunta

$$F(x, y|B) = Prob\{(X \leq x) \cdot (Y \leq y)|B\} \quad (2.23)$$

Como anteriormente, poderemos definir a função de distribuição condicional marginal, por exemplo para X ,

$$F(x|B) = F(x, +\infty|B) \quad (2.24)$$

e igualmente se a função de distribuição conjunta for continuamente derivável

$$p(x, y|B) = \frac{\partial^2 F(x, y|B)}{\partial x \partial y} \quad (2.25)$$

e obviamente por derivação relativamente a x

$$p(x|B) = \int p(x, y|B)dy \quad (2.26)$$

Frequentemente interessamo-nos mais pela distribuição condicional de uma VA X em relação a uma outra VA Y , por exemplo quando os dois acontecimentos são $X \leq x$ e $Y \leq y$, nesse caso

$$F(x|Y \leq y) = Prob\{(X \leq x)|(Y \leq y)\} = \frac{F(x, y)}{F(+\infty, y)} \quad (2.27)$$

Porém, em muitos casos, estamos mais interessados no acontecimento condicional $Y = y$ do que propriamente $Y \leq y$ e então, admitindo que $F(x, y)$ é diferenciável em todo o domínio, podemos escrever a definição da densidade de probabilidade condicional de $X \leq x$ para $Y = y$

$$p(x|y) = \frac{p(x, y)}{p(y)} \quad (2.28)$$

Note-se que esta relação é uma definição e que é uma relação entre densidades de probabilidade e não uma relação entre probabilidades. De (2.28) podemos escrever

$$p(x, y) = p(x|y)p(y) \quad (2.29)$$

Frequentemente as distribuições condicionais são chamadas distribuições *a posteriori* contrastando com as distribuições marginais que são chamadas distribuições *a priori*.

2.6 Esperanças condicionais

A importância das distribuições condicionais surge completamente quando se calculam médias condicionais, em geral, e no tratamento de problemas de estimação em particular. Assim a esperança matemática condicional

$$E_X[h(X)|y] = \int h(x)p(x|y)dx \quad (2.30)$$

de notar que (2.30) é ainda uma função de y e que portanto poderemos definir a sua esperança matemática em relação a y que se escreve

$$E_Y[E_X[h(X)|y]] = \int E_X[h(X)|y]p(y)dy \quad (2.31)$$

Podemos agora observar que substituindo (2.30) em (2.31) e usando (2.28) se obtém

$$E_Y[E_X[h(X)|y]] = \int \int h(x)p(x, y)dxdy \quad (2.32)$$

expressão na qual podemos usar (2.11) para a densidade de probabilidade marginal $p(x)$ obtendo finalmente

$$E_Y[E_X[h(X)|y]] = \int h(x)p(x)dx = E[h(X)] \quad (2.33)$$

que não é mais do que a esperança matemática marginal. Em conclusão, calculando a esperança matemática em relação á variável de condicionamento obtem-se a esperança matemática marginal. Mais uma vez substituindo $h(x)$ por x^k podemos calcular os momentos condicionais de ordem k . O momento condicional de primeira ordem tem um papel primordial pois é o que normalmente se chama *regressão* $r(y)$ definida por

$$r(y) = E_X[X|y] = E[X|y] \quad (2.34)$$

notemos desde já que uma VA X, Y pode ter média nula sem no entanto ter uma regressão nula.

2.7 Variáveis aleatórias discretas

Como já tivemos ocasião de referir, todos os resultados obtidos nos capítulos anteriores podem ser redemonstrados no caso de VA discretas. No caso discreto a função de distribuição é uma função em

escada que no caso bidimensional é evidentemente uma função de duas variáveis $F(x_i, y_i)$ tomando valores nos pontos (x_i, y_i) do plano com as probabilidades

$$p(i, j) = p_{i,j} = Prob\{(X = x_i) \text{ e } (Y = y_i)\} \quad (2.35)$$

A probabilidade total associada com um determinado domínio Δ do plano (como exemplificado na fig 2.1) escreve-se

$$P(\Delta) = \sum_{\Delta} p(i, j) \quad (2.36)$$

A probabilidade marginal $p(i)$ pode ser obtida facilmente através do somatório na outra variável,

$$p(i) = \sum_j p(i, j) \quad (2.37)$$

Em particular, e em seguimento de (2.18), podemos dizer que para VAs independentes

$$p(i, j) = p(i)p(j) \quad (2.38)$$

A distribuição condicional entre duas VA's X, Y discretas escreve-se

$$p_{i|j} = Prob\{(X = x_i)|(Y = y_i)\} = \frac{p_{i,j}}{p_j} \quad (2.39)$$

2.8 Momentos de segunda ordem: covariância e correlação

Consideremos duas VAs X e Y . Define-se o momento conjunto de ordem $j + k$ por

$$m_{jk} = E[X^j Y^k] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^j y^k p(x, y) dx dy \quad (2.40)$$

de onde se pode deduzir imediatamente que

$$m_{00} = 1 \quad (2.41a)$$

$$m_{10} = E[X] = m_x \quad (2.41b)$$

$$m_{01} = E[Y] = m_y \quad (2.41c)$$

$$m_{n0} = E[X^n] \quad (2.41d)$$

$$m_{0n} = E[Y^n] \quad (2.41e)$$

Os momentos centrais conjuntos escrevem-se então

$$\mu_{jk} = E[(X - m_x)^j (Y - m_y)^k] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - m_x)^j (y - m_y)^k p(x, y) dx dy \quad (2.42)$$

de onde

$$\mu_{00} = 1 \quad (2.43a)$$

$$\mu_{10} = 0 \quad (2.43b)$$

$$\mu_{01} = 0 \quad (2.43c)$$

$$\mu_{20} = E[(X - m_x)^2] = V[X] = \sigma_x^2 \quad (2.43d)$$

$$\mu_{02} = E[(Y - m_y)^2] = V[Y] = \sigma_y^2 \quad (2.43e)$$

O momento central conjunto de ordem um, que toma geralmente o nome de covariância, desempenha um papel importante

$$\mu_{11} = E[(X - m_x)(Y - m_y)] = COV[X, Y] \quad (2.44)$$

A partir de (2.44) podemos escrever

$$COV[X, Y] = m_{11} - m_x m_y \quad (2.45)$$

Suponha-se agora que a VA bidimensional, ou o conjunto das duas VA's X, Y , satisfaz $COV[X, Y] = 0$. Então a partir de (2.44), $m_{11} = m_x m_y$. Já vimos anteriormente que neste caso se pode dizer que X e Y são duas VAs independentes. Podemos introduzir agora o coeficiente de correlação

$$\rho = \frac{COV[X, Y]}{\sigma_x \sigma_y} = \frac{\mu_{11}}{\sqrt{\mu_{20} \mu_{02}}} \quad (2.46)$$

o que significa que para VAs não correladas temos $\rho = 0$. Mais, $-1 \leq \rho \leq 1$, o que nos indica que ρ é uma medida de correlação entre as VAs X e Y .

Uma aplicação do coeficiente de correlação encontra-se na regressão linear. A regressão linear aplica-se quando se pretende determinar o valor de uma VA que depende de outra VA. Uma forma gráfica de mostrar a possível ligação entre as duas VA's faz-se através de um plano X, Y onde se marcam os pontos de coordenadas x, y como exemplificado na fig 2.2. Normalmente se X e Y fossem duas VAs independentes os pontos estariam distribuídos no plano de modo uniforme. Se pelo contrário existir uma certa dependência entre as duas VA's então os pontos estarão distribuídos no plano segundo uma determinada curva característica da dependência. A curva mais simples será uma linha recta. Se suspeitarmos que se trata efectivamente de uma dependência linear, à parte os erros, então procuraremos ajustar *o melhor possível* uma linha recta entre os pontos do plano. O problema consiste em determinar a e b que a partir de qualquer X permitem determinar $Y_e = aX + b$ com o mínimo erro em relação aos Y do plano.

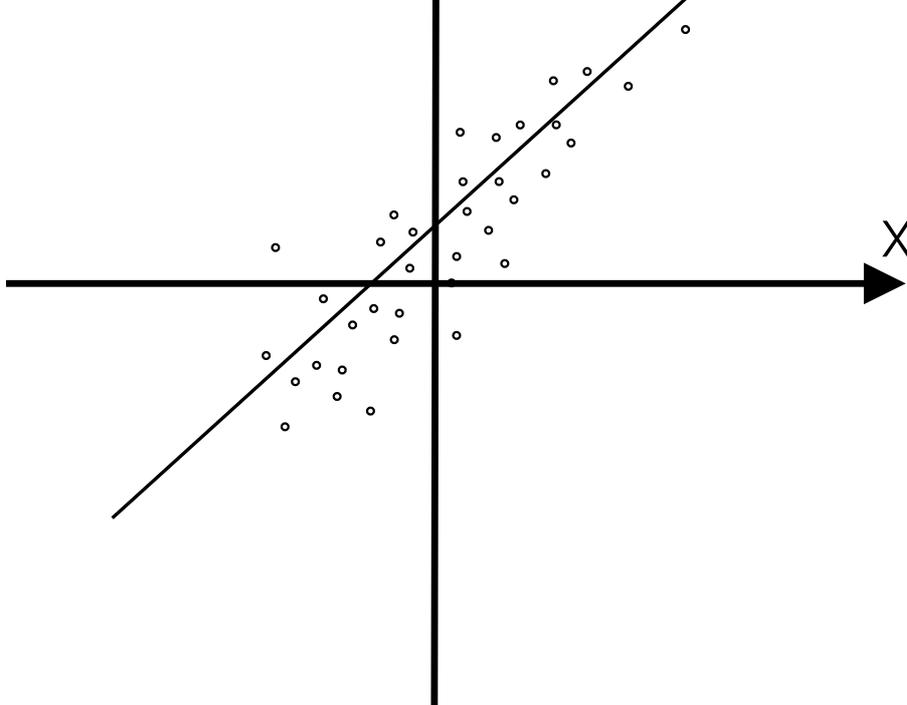


Figura 2.2: regressão linear

Se definirmos o erro através do erro quadrático médio

$$MSE = E[(Y - Y_e)^2] = E[(Y - aX - b)^2] \quad (2.47)$$

Poderemos então calcular a e b derivando (2.47) em relação a a e a b . Igualando as derivadas a zero e resolvendo o sistema de equações obtido (que deixamos como exercício!) obtém-se

$$b = m_y - \frac{\mu_{11}m_x}{\sigma_x^2} \quad (2.48a)$$

$$a = \rho \frac{\sigma_y}{\sigma_x} \quad (2.48b)$$

e portanto a linha de regressão escreve-se

$$Y_e = \rho \frac{\sigma_y}{\sigma_x} (X - m_x) + m_y \quad (2.49)$$

A partir de (2.49) podemos ver que se $\rho = 0$, i.e., se X e Y forem não correladas, que $Y_e = m_y$ e que $MSE = \sigma_y^2$ e que se pelo contrário $|\rho| = 1$ então $MSE = 0$ e todos ou quase todos os pontos (em média) encontrar-se-ão em cima da linha de regressão cuja inclinação será σ_y/σ_x .

3. Vectores aleatórios

Como já referimos acima, a extensão a vectores aleatórios de dimensão n faz-se directamente considerando variáveis aleatórias multidimensionais por extensão do caso bidimensional. Consideremos então n VA's $X_1(\omega), X_2(\omega), \dots, X_n(\omega)$. Estas VA's podem ser consideradas como sendo as componentes de um vector $\mathbf{X}(\omega)$ de R^n , que é nesse caso um vector aleatório. O caso bidimensional discutido anteriormente correspondia a um vector num plano.

Antes de passar à discussão dos métodos a pôr em prática utilizando vectores aleatórios podemos citar como exemplos de utilização o caso da observação de dados multisensores em diversos pontos do espaço e o caso de um sinal temporal $s(t)$ observado a n momentos discretos do tempo $s(t_i)$ formando um vector $\mathbf{s} = [s(t_1), \dots, s(t_n)]$ que pode ser manipulado como um vector aleatório, calculada a sua FFT, etc...

3.1 Função de distribuição e densidade de probabilidade

Como no caso bidimensional admitiremos que um vector aleatório se encontra completamente caracterizado pela sua função de distribuição que no caso a n dimensões se escreve

$$F(\mathbf{x}) = F(x_1, x_2, \dots, x_n) = Prob\{(X_1 \leq x_1) \cdot (X_2 \leq x_2) \dots (X_n \leq x_n)\} \quad (3.1)$$

A função $F(\mathbf{x})$ do vector aleatório \mathbf{x} goza exactamente das mesmas propriedades de qualquer função de distribuição, i.e., é uma função monotóna não decrescente de todas as variáveis x_i , se todas as variáveis x_i tenderem para $-\infty$ ou $+\infty$, $F(\mathbf{x})$ tende para 0 ou 1 respectivamente e se só k variáveis x_i tenderem para $+\infty$, $F(\mathbf{x})$ torna-se uma função de um vector de dimensão $n - k$ o que não é mais do que a definição de função de distribuição marginal. Se $F(\mathbf{x})$ admite uma derivada no ponto x podemos definir a função densidade de probabilidade do vector \mathbf{x} por

$$p(\mathbf{x}) = \frac{\partial^n F}{\partial x_1 \partial x_2 \dots \partial x_n} \quad (3.2)$$

o que implica que

$$F(\mathbf{x}) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} p(\xi_1, \dots, \xi_n) d\xi_1 \dots d\xi_n \quad (3.3)$$

Do mesmo modo a densidade de probabilidade marginal de ordem m escreve-se

$$p(x_1, \dots, x_m) = \frac{\partial^m}{\partial x_1 \dots \partial x_m} F(x_1, \dots, x_m, +\infty, \dots, +\infty) \quad (3.4)$$

3.2 Esperança matemática

Considere uma função vectorial $h(\mathbf{x})$ do vector aleatório \mathbf{x} . Podemos escrever a média de $h(\mathbf{X})$ como

$$E[h(\mathbf{X})] = \int h(\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad (3.5)$$

que obviamente é um vector da mesma maneira que a esperança matemática $\mathbf{m} = E[\mathbf{X}]$ do vector aleatório \mathbf{X} ,

$$\mathbf{m} = E[\mathbf{X}] = \int \mathbf{x}p(\mathbf{x})d\mathbf{x} \quad (3.6)$$

As propriedades de segunda ordem cruzadas podem ser deduzidas através da esperança matemática do produto externo do vector aleatório centrado. Assim

$$\mathbf{R}_x = COV[\mathbf{X}] = E[(\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T] \quad (3.7)$$

onde $\mathbf{m} = E[\mathbf{X}]$ e \mathbf{R}_x é uma matriz chamada matriz de covariância do vector aleatório \mathbf{X} . Alternativamente podemos escrever \mathbf{R}_x como

$$\mathbf{R}_x = E[\mathbf{X}\mathbf{X}^T] - \mathbf{m}\mathbf{m}^T \quad (3.8)$$

Podemos notar que os termos da diagonal desta matriz não são mais do que as variâncias de \mathbf{X} . Muitas vezes o primeiro termo de (3.8) é chamada matriz de correlação o que significa que para um vector de média nula, $\mathbf{m} = \mathbf{0}$ as matrizes de correlação e de covariância são idênticas.

Impõe-se aqui uma pequena nota respeitante á terminologia de correlação e covariância que neste caso não é completamente satisfatória. Com efeito, já vimos num capítulo anterior que em termos de variáveis aleatórias indepêndencia implicava não correlação, mas que o inverso não é forçosamente verdade. Porém adoptando a terminologia acima duas variáveis aleatórias de média nula e não correladas implicaria uma covariância nula o que não seria verdade. A conclusão é que chamar $E[\mathbf{X}\mathbf{X}^T]$ matriz de correlação é enganador e pode induzir em graves erros.

A equação (3.8) permite calcular a matriz de covariância do vector \mathbf{X} com ele próprio e é portanto, em termos de vectores, uma espécie de auto-covariância. É frequentemente necessário determinar as características estatísticas entrelaçadas entre dois vectores aleatórios \mathbf{X} e \mathbf{Y} . Podemos então definir a função de distribuição conjunta

$$F(\mathbf{x}, \mathbf{y}) = Prob\left\{\prod_{i=1}^m (X_i \leq x_i) \prod_{j=1}^n (Y_j \leq y_j)\right\} \quad (3.9)$$

onde \mathbf{X} e \mathbf{Y} são de dimensão m e n respectivamente. Neste caso a matriz de covariância entre os dois vectores define-se por

$$\mathbf{R}_{x,y} = E[(\mathbf{X} - \mathbf{m}_x)(\mathbf{Y} - \mathbf{m}_y)^T] \quad (3.10)$$

que é chamada muitas vezes matriz de covariância cruzada (cross-covariance matrix) e é de dimensões $m \times n$. Devido à importância de que se reveste a análise de matrizes de covariância em processamento de sinais multidimensionais o próximo capítulo será inteiramente dedicado ao estudo das suas propriedades.

3.3 Matrizes de covariância

A propriedade principal de qualquer matriz de covariância é que é uma matriz definida não negativa (DNN). Inversamente qualquer matriz DNN pode ser considerada como a matriz de covariância de

um vector aleatório \mathbf{X} . Podemos além disso provar que a soma de duas matrizes de covariância é ainda uma matriz de covariância. O produto directo

$$\mathbf{R}_{m,n} = \mathbf{R}_{1;m,n} \mathbf{R}_{2;m,n} \quad (3.11)$$

de duas matrizes de covariância é também uma matriz de covariância e finalmente sendo \mathbf{A} uma matriz qualquer, $\mathbf{A}\mathbf{R}\mathbf{A}^H$ é também uma matriz de covariância.

Uma matriz de covariância é de ordem completa se todos os seus valores próprios λ_i forem positivos. Por outro lado se p valores próprios são nulos a ordem da matriz será $n - p$.

A decomposição espectral da matriz de covariância permite uma análise aprofundada da sua estrutura. Assim podemos escrever a matriz de covariância \mathbf{R} como

$$\mathbf{R} = COV[\mathbf{X}] = \sum_{i=1}^n \lambda_i \mathbf{e}_i \mathbf{e}_i^H = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^H \quad (3.12)$$

onde $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_n]$ é uma matriz cujas colunas são os vectores próprios de \mathbf{R} e os λ_i são os valores próprios correspondentes colocados na diagonal de $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$. Isto significa que o vector aleatório \mathbf{X} se pode escrever na base formada pelos vectores próprios

$$\mathbf{X} = \sum_{i=1}^n X_i \mathbf{e}_i \quad ; \quad X_i = \mathbf{e}_i^H \mathbf{X} \quad (3.13)$$

que se obtém devido ao facto de os vectores próprios serem ortonormais entre eles. Nesta base, formada pelos vectores próprios, as componentes X_i são descorreladas

$$E[X_i X_j^*] = \mathbf{e}_i^H \mathbf{R}_x \mathbf{e}_j = \lambda_i \delta_{i,j} \quad (3.14)$$

A equação (3.13) é muitas vezes chamada a expansão Karhunen-Loève do vector aleatório \mathbf{X} . Se a ordem da matriz \mathbf{R}_x for inferior a n isto quer dizer que alguns dos valores próprios são nulos, o que a partir de (3.14) significa que a variância de algumas componentes de \mathbf{X} é igual a zero. Variância igual a zero significa que essas componentes perderam o seu carácter aleatório e são em princípio iguais à média, que se esta for igual a zero, quer dizer que as componentes são elas mesmas quase de certeza iguais a zero. Como consequência o vector \mathbf{X} encontra-se completamente definido por uma expansão de ordem $n - p$ se houverem p valores próprios nulos e portanto está restrito a um sub-espaco de C^{n-p} de dimensão $n - p$. Mais, como a base formada pelos vectores próprios é ortogonal, pode-se dizer que o sub-espaco de \mathbf{X} de dimensão $n - p$ é ortogonal ao sub-espaco formado pelos p vectores próprios associados com os valores próprios nulos. Em processamento do sinal estas considerações são extremamente importantes e quando esta situação acontece é geralmente recomendado de suprimir o sub-espaco associado com os valores próprios nulos de forma a reobter uma matriz de ordem completa. Voltaremos a insistir neste assunto mais á frente.

3.4 Transformações lineares de vectores aleatórios

Consideremos dois vectores aleatórios \mathbf{X} e \mathbf{Y} de dimensões m e n respectivamente e \mathbf{A} uma matriz determinística de dimensões $m \times n$ tal que

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \quad (3.15)$$

Para começar é fácil deduzir que

$$\bar{\mathbf{Y}} = E[\mathbf{Y}] = \mathbf{A}E[\mathbf{X}] = \mathbf{A}\bar{\mathbf{X}} \quad (3.16)$$

A matriz de covariância escreve-se então

$$\mathbf{R}_y = E[(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^H] = \mathbf{A}\mathbf{R}_x\mathbf{A}^H \quad (3.17)$$

onde \mathbf{R}_x é a matriz de covariância do vector aleatório \mathbf{X} . O problema inverso é também interessante, i.e., a partir de uma dada matriz de covariância \mathbf{R}_x e uma matriz desejada \mathbf{R}_y determinar qual a transformação que permite passar de uma para a outra. A solução deste problema encontra-se na factorização de matrizes de covariância.

Considere um vector aleatório \mathbf{X} de média nula e cujas componentes são mutuamente descorreladas e de variância σ^2 , i.e., $E[X_i X_j^*] = \sigma^2 \delta_{i,j}$. A matriz de covariância de \mathbf{X} escreve-se nesse caso

$$\mathbf{R}_x = E[\mathbf{X}\mathbf{X}^H] = \sigma^2 \mathbf{I} \quad (3.18)$$

onde \mathbf{I} é a matriz identidade. Podemos além disso considerar, para simplificar, que $\sigma^2 = 1$. Supunhamos que \mathbf{X} é transformado em \mathbf{Y} por (3.15). A matriz de covariância de \mathbf{Y} escreve-se segundo (3.17)

$$\mathbf{R}_y = \mathbf{A}\mathbf{A}^H \quad (3.19)$$

podemos dizer então que \mathbf{R}_y foi *factorizada* usando a matriz \mathbf{A} . Esta matriz não é única pois para qualquer matriz \mathbf{B} tal que $\mathbf{B}\mathbf{B}^H = \mathbf{I}$, definindo $\mathbf{C} = \mathbf{A}\mathbf{B}$, $\mathbf{R}_y = \mathbf{C}\mathbf{C}^H$ é também uma factorização de \mathbf{R}_y . Podemos calcular a matriz \mathbf{A} de (3.19), algumas vezes chamada raiz de \mathbf{R}_y , se considerarmos que é uma matriz triangular inferior com elementos não negativos na diagonal. Esta suposição conduz a escrever os elementos de \mathbf{Y} como

$$Y_i = \sum_{j=1}^i a_{ij} X_j \quad (3.20)$$

pois os elementos $a_{i,j} = 0; j > i$. O primeiro elemento para $i = 1$ calcula-se directamente através de

$$Y_1 = a_{11} X_1 \Rightarrow E[|Y_1|^2] = R_{y;11} = |a_{11}|^2 \quad (3.21)$$

visto que $a_{11} \geq 0$ que $a_{11} = \sqrt{R_{y;11}}$. Os outros elementos da matriz \mathbf{A} podem ser calculados recursivamente. Vejamos para a linha k supondo que os a_{ij} são conhecidos para $1 \leq j \leq i$ e $1 \leq j \leq k-1$ e que desejamos calcular a_{km} . Como os X_i são descorrelados temos que

$$R_{y;kp} = E[Y_k Y_p^*] = a_{k1} a_{p1}^* + a_{k2} a_{p2}^* + \dots + a_{kp} a_{pp}^* \quad (3.22)$$

no caso em que $p < k$. Para $p = 1$ temos que $R_{y;k1} = a_{k1} a_{11}^*$ e visto que conhecemos a_{11} podemos calcular a_{k1} . Para $p = 2$ temos que $R_{y;k2} = a_{k1} a_{21}^* + a_{k2} a_{22}^*$ de onde podemos tirar a_{k2} visto que conhecemos todos os outros elementos. Recursivamente podemos chegar ao elemento a_{km} . Este procedimento recursivo de factorização de matrizes DNN é chamado factorização de Cholesky. Se

além disso os elementos da diagonal de \mathbf{A} forem não nulos \mathbf{A} admite um inverso, que é também triangular inferior e que permite escrever,

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y} \quad (3.23)$$

permitindo obter um vector aleatório descorrelado a partir de qualquer vector aleatório \mathbf{Y} . Esta transformação linear (inversa) é chamada filtragem de branqueamento (whitening filtering) por razões que veremos mais adiante.

Finalmente a factorização (3.19) permite-nos determinar a transformação linear a aplicar a uma matriz de covariância \mathbf{R}_x para obter uma outra matriz pré definida \mathbf{R}_y , assim se escrevermos

$$\mathbf{R}_x = \mathbf{A}_x \mathbf{A}_x^H \quad (3.24a)$$

$$\mathbf{R}_y = \mathbf{A}_y \mathbf{A}_y^H \quad (3.24b)$$

e substituindo em (3.17) temos que

$$\mathbf{A}_y = \mathbf{A} \mathbf{A}_x \quad (3.25)$$

ou seja que

$$\mathbf{A} = \mathbf{A}_y \mathbf{A}_x^{-1} \quad (3.26)$$

que só existe se \mathbf{R}_x for DNN.

3.5 Distribuições condicionais

A densidade de probabilidade condicional no caso de vectores aleatórios escreve-se por extensão do caso bidimensional (2.28) da seguinte forma

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} \quad (3.27)$$

onde a densidade de probabilidade $p(\mathbf{y})$ que aparece no denominador é a densidade de probabilidade marginal em \mathbf{y} que se obtém por integração sobre \mathbf{x} como em (2.11) mas agora sob todo o espaço R^m igual á dimensão de \mathbf{x} . Da mesma maneira que precedentemente para dois vectores aleatórios independentes

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}) \quad (3.28)$$

3.6 Esperança matemática condicional

A densidade de probabilidade condicional $p(\mathbf{x}|\mathbf{y})$ definida em (3.27) permite-nos introduzir a noção de esperança matemática condicional de um vector \mathbf{x} sabendo que \mathbf{y} é conhecido, $E[\mathbf{X}|\mathbf{Y}]$. Aplicam-se exactamente as mesmas relações que no caso bidimensional a saber que

$$E[h(\mathbf{X})|\mathbf{Y}] = \int h(\mathbf{x})p(\mathbf{x}|\mathbf{y})d\mathbf{x} \quad (3.29)$$

sabendo porém que $E[h(\mathbf{X})|\mathbf{Y}]$ é neste caso um vector aleatório função do vector \mathbf{Y} . Como no caso bidimensional, podemos calcular a esperança condicional de (3.29) relativamente a \mathbf{Y} que se escreve

$$E[E[h(\mathbf{X}|\mathbf{Y})]] = E[h(\mathbf{X})] \quad (3.30)$$

que não é mais do que a esperança marginal relativamente a \mathbf{X} .

Voltemos agora á regressão expressa em (2.34) para o caso bidimensional e por

$$\mathbf{r}(\mathbf{Y}) = E[\mathbf{X}|\mathbf{y}] \quad (3.31)$$

no caso multidimensional. A regressão $\mathbf{r}(\mathbf{Y})$ tem uma propriedade muito interessante que usaremos mais tarde em estimação. Com efeito, consideremos um vector aleatório \mathbf{Y} e uma função $g(\mathbf{Y})$ desse vector aleatório. Vamos provar que o vector aleatório $\mathbf{X} - \mathbf{r}(\mathbf{Y})$ e $\mathbf{g}(\mathbf{Y})$ são *não correladas e ortogonais*.

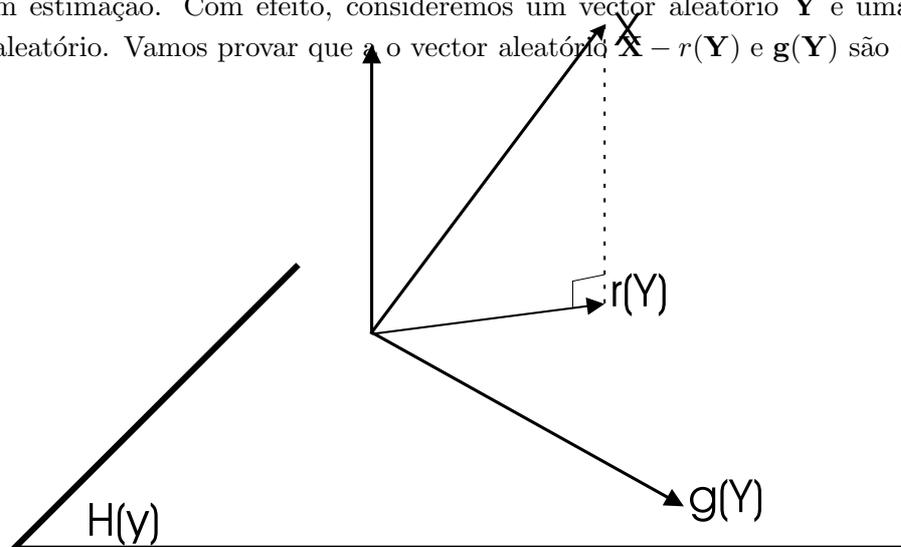


Figura 3.1: regressão e projeção ortogonal

Antes de mais notemos que $E[\mathbf{X} - \mathbf{r}(\mathbf{Y})] = E[\mathbf{X}] - E[\mathbf{r}(\mathbf{Y})] = 0$ devido a (3.30). A propriedade de ortogonalidade e não correlação (para vectores aleatórios de média nula) prova-se demonstrando que

$$E[(\mathbf{X} - \mathbf{r}(\mathbf{Y}))^T \mathbf{g}(\mathbf{Y})] = E\left[\sum_{i=1}^n (X_i - r_i(\mathbf{Y}))g_i(\mathbf{Y})\right] = 0 \quad (3.32)$$

e portanto devemos provar que

$$E[(X_i - r_i(\mathbf{Y}))g_i(\mathbf{Y})] = 0 \quad \forall i = 1, \dots, n \quad (3.33)$$

o que é relativamente fácil sabendo que $g_i(\mathbf{Y}) = g(Y_i)$ e portanto que

$$E[(X_i - r_i(\mathbf{Y}))g_i(\mathbf{Y})] = \int \int [x_i - r_i(y)]g(y_i)p(x_i, y_i)dx_i dy_i \quad (3.34)$$

ou ainda que

$$E[(X_i - r_i(\mathbf{Y}))g_i(\mathbf{Y})] = \int \int [x_i - r_i(\mathbf{y})]p(x_i|y_i)g(y_i)p(y_i)dx_i dy_i \quad (3.35)$$

no entanto a integração desta expressão em relação a x_i é precisamente igual a $E[X_i|\mathbf{y}]$ e portanto igual a $r_i(\mathbf{y})$ obtendo-se $r_i(\mathbf{y}) - r_i(\mathbf{y})$ e provando então (3.33). O conceito importante que quisemos demonstrar é que qualquer que seja a função g temos que $\mathbf{X} - \mathbf{r}(\mathbf{Y})$ e $\mathbf{g}(\mathbf{Y})$ são ortogonais o que indica geometricamente que $\mathbf{r}(\mathbf{Y})$ é a projeção ortogonal do vector aleatório \mathbf{X} no sub-espaço gerado pelos vectores aleatórios $\mathbf{g}(\mathbf{Y})$, segundo representado na fig. 3.1. Esta noção é de grande importância em estimação, em particular nos estimadores chamados da esperança condicional.

3.7 Funções de vectores aleatórios

Voltemos ao problema das funções de VAs que no caso de vectores aleatórios

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}) \quad (3.36)$$

O cálculo da função de distribuição de \mathbf{Y} em função da função de distribuição de \mathbf{X} é um problema clássico de mudança de variável numa função multivariada. Notando $D_{\mathbf{y}}$ o domínio de definição do vector aleatório \mathbf{y} podemos dizer que

$$D_{\mathbf{y}} = \{\mathbf{x} | \prod_{i=1}^n [g_i(\mathbf{x}) \leq y_i]\} \quad (3.37)$$

A partir do domínio de \mathbf{Y} podemos deduzir a função de distribuição de \mathbf{Y} a partir da de \mathbf{X} através de

$$F_{\mathbf{Y}}(\mathbf{y}) = \int_{D_{\mathbf{y}}} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \quad (3.38)$$

O cálculo resultante pode ser extremamente complexo dependendo da função g e do domínio $D_{\mathbf{y}}$. Como exemplo podemos lembrar que no caso vectorial o problema $Y = X^2$ se escreve $Y = \mathbf{X}^T \mathbf{X}$, o que significa que sendo o vector \mathbf{X} de dimensão m qualquer o vector aleatório \mathbf{Y} é um escalar de dimensão $n = 1$. É óbvio que neste caso será difícil calcular $\mathbf{x} = g^{-1}\mathbf{y}$ visto que \mathbf{x} e \mathbf{y} não são de mesma dimensão. O caso particular em que g é invertível requer $m = n$. O cálculo da integral múltipla de (3.38) pode ser levado a cabo graças á introdução do determinante do Jacobiano da transformação $\mathbf{y} = \mathbf{g}(\mathbf{x})$. Se a transformação em questão for uma correspondência de um para um podemos exprimir x_k em função dos y_l através do determinante da matriz

$$\mathbf{M} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix} \quad (3.39)$$

O determinante $J = \det(\mathbf{M})$ é uma função de \mathbf{y} e podemos escrever (3.38) como

$$F_{\mathbf{Y}}(\mathbf{y}) = \int_{D_{\mathbf{g}(\mathbf{x})}} p_{\mathbf{X}}[\mathbf{g}^{-1}(\mathbf{y})] |J(\mathbf{y})| d\mathbf{y} \quad (3.40)$$

3.8 Funções características e momentos

No caso multidimensional a função característica de um vector aleatório \mathbf{X} é definida de forma idêntica que em (3.5) com $h(\mathbf{X}) = \exp(j\mathbf{u}^T \mathbf{X})$. Assim

$$\phi(\mathbf{u}) = E[\exp(j\mathbf{u}^T \mathbf{X})] = \int_{R^n} \exp(j\mathbf{u}^T \mathbf{X}) p(\mathbf{x}) d\mathbf{x} \quad (3.41)$$

No caso em que as componentes do vector \mathbf{X} são mutuamente independentes temos que

$$\phi(\mathbf{u}) = \int_{R^n} \exp j(u_1 x_1 + \dots + u_n x_n) p(\mathbf{x}) d\mathbf{x} \quad (3.42)$$

e então, visto que $p(\mathbf{x})$ se pode escrever como um produto das densidades marginais, então

$$\phi(\mathbf{u}) = \prod_{i=1}^n \phi(u_i) \quad (3.43)$$

Existe uma série de propiedades da função característica, semelhantes às obtidas para o caso escalar e que serão ilustradas através de exemplos. Em particular a relação entre a função característica e o momento de ordem k que se escreve no caso vectorial

$$\phi(\mathbf{u}) = \sum_{k=0}^{\infty} \frac{j^k}{k!} E[(\mathbf{u}^T \mathbf{X})^k] \quad (3.44)$$

Um último caso em que a função característica é bastante útil é no cálculo dos momentos através de

$$j \frac{\partial}{\partial \mathbf{u}} \phi(\mathbf{u} = \mathbf{0}) = \mathbf{m} \quad (3.45a)$$

$$j \frac{\partial^2}{\partial \mathbf{u}^2} \phi(\mathbf{u} = \mathbf{0}) = \mathbf{R} - \mathbf{m}\mathbf{m}^T \quad (3.45b)$$

4. Vectores aleatórios gaussianos

As variáveis aleatórias gaussianas foram introduzidas anteriormente. A sua densidade de probabilidade é caracterizada por dois parâmetros: a média e a variância. Neste capítulo vamos estudar conjuntos de variáveis aleatórias que tem a propriedade de serem conjuntamente gaussianas.

A importância da distribuição gaussiana repousa no facto que muitos fenómenos naturais podem ser descritos, com uma boa aproximação, por esta distribuição. Uma das razões que explica este facto é o teorema central limite. Essencialmente, quando muitas pequenas contribuições contribuem para um dado resultado, a flutuação aleatória deste pode ser aproximada por uma VA gaussiana, ou no caso multidimensional, por um vector aleatório gaussiano. No caso multidimensional as quantidades que caracterizam um vector aleatório gaussiano são o vector média e a matriz de covariância.

4.1 Teorema central limite

Considere um conjunto de n VAs estatisticamente independentes

$$X_1, X_2, \dots, X_n \quad (4.1)$$

e a VA definida por

$$Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \quad (4.2)$$

O teorema central limite postula que, sob um série de condições pouco severas, a densidade de probabilidade da VA Y aproxima-se de uma distribuição gaussiana quando $n \rightarrow \infty$ independentemente da distribuição das VA's X_i , i.e.,

$$p(y) \rightarrow \frac{1}{\sigma_y \sqrt{2\pi}} \exp - \frac{(y - \bar{y})^2}{2\sigma_y^2} \quad (4.3)$$

onde a média e a variância de Y , \bar{y} e σ_y^2 respectivamente se escrevem

$$\bar{y} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{x}_i \quad (4.4a)$$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n \sigma_{x_i}^2 \quad (4.4b)$$

onde \bar{x}_i e $\sigma_{x_i}^2$ são a média e a variância de X_i respectivamente. Uma condição suficiente para a validade do teorema para grandes valores de n é que

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E[(x_i - \bar{x}_i)^{2+\delta}]}{\left(\sum_{i=1}^n \sigma_{x_i}^2\right)^{1+\delta/2}} = 0 \quad (4.5)$$

para $\delta > 0$. Podemos interpretar esta condição da seguinte forma: suponhamos que $\delta = 0$ e nesse caso a fração é 1. Se $\delta > 0$, o numerador aumentará mais lentamente do que o denominador se

houver uma pequena tendência de x_i para se desviar da média e comparativamente á variância e portanto a fração tenderá para zero quando $n \rightarrow \infty$. Frequentemente a condição relativa á independência dos X_i pode ser aliviada mas nesse caso a convergência será menos rápida.

4.2 Definição

Considere-se o vector aleatório $\mathbf{X}^T = [X_1, X_2, \dots, X_n]$ composto por n variáveis aleatórias de média m_1, m_2, \dots, m_n e uma matriz de covariância $R_{x;ij} = COV[X_i, X_j]$.

As VA's X_1, \dots, X_n são consideradas conjuntamente gaussianas se a sua densidade de probabilidade conjunta se escrever

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{R}_x)}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{m})^T \mathbf{R}_x^{-1}(\mathbf{x} - \mathbf{m})\right] \quad (4.6)$$

ou de forma equivalente se a sua função característica conjunta se escrever

$$\phi(\mathbf{u}) = \exp\left[j\mathbf{u}^T \mathbf{m} - \frac{1}{2}\mathbf{u}^T \mathbf{R}_x \mathbf{u}\right] \quad (4.7)$$

onde $\mathbf{x}^T = [x_1, x_2, \dots, x_n]$, $\mathbf{m}^T = [m_1, m_2, \dots, m_n]$ e $\mathbf{u}^T = [u_1, u_2, \dots, u_n]$. É usual dizer que neste caso \mathbf{X} é um vector aleatorio gaussiano ou abreviadamente que $\mathbf{X} : N(\mathbf{m}, \mathbf{R}_x)$.

4.3 Propriedades

Se as VAs X_1, X_2, \dots, X_n forem não correladas então temos que a sua variância $E[X_i X_j] = 0$ com $i \neq j$ e portanto a matriz de covariância $\mathbf{R}_x = COV[\mathbf{X}\mathbf{X}^T]$ é uma matriz diagonal. Nesse caso é óbvio que podemos factorizar (4.6) sob a forma

$$p(\mathbf{x}) = p(x_1)p(x_2) \dots p(x_n) \quad (4.8)$$

o mesmo acontecendo para (4.7),

$$\phi(\mathbf{u}) = \phi(u_1)\phi(u_2) \dots \phi(u_n) \quad (4.9)$$

e portanto podemos deduzir o importante resultado que, contrariamente ao caso geral, para VA's gaussianas não correlação implica independência.

Como já referimos atrás a distribuição gaussiana tem a particularidade de se encontrar completamente especificada a partir dos seus dois primeiros momentos, média e variância o que no caso vectorial significa o vector média e a matriz de covariância. Este facto aliado com a frequência com a qual os fenómenos físicos podem ser modelizados por distribuições gaussianas demonstra a importância do estudo dos dois primeiros momentos em geral.

Note-se que em (4.6) entra o inverso da matriz de covariância. O inverso de uma matriz de covariância só existe quando está é definida positiva, i.e., se todos os seus valores próprios são

estricamente positivos. Na prática podemos porém falar de $X : N(\mathbf{m}, \mathbf{R}_x)$ mesmo se alguns dos valores próprios de \mathbf{R}_x são nulos. Físicamente isso significa que algumas das componentes de \mathbf{X} são combinações lineares das outras componentes. Se a ordem de \mathbf{R}_x for $p \leq n$ o vector \mathbf{X} é gaussiano num espaço a p dimensões.

O uso de (4.7) como definição de distribuição gaussiana é mais prática pois, entre outros, não necessita o cálculo do inverso da matriz de covariância. Além disso pode-se demonstrar que a função característica é equivalente á transformada de Fourier o que, dado a forma de (4.7) significa que para vectores aleatórios de média nula a Transformada de Fourier de uma VA gaussiana é ainda uma VA gaussiana. O mesmo podemos dizer para o caso vectorial.

4.4 Transformações lineares

Uma das mais interessantes propriedades da distribuição gaussiana é que o resultado de uma transformação linear de um conjunto de VA's conjuntamente gaussianas é ainda um conjunto de VA's conjuntamente gaussianas. Óbviamente isto traduz-se no caso vectorial que uma transformação linear de um vector aleatório gaussiano é ainda um vector aleatório gaussiano.

Assim considerando o vector $\mathbf{X} : N(\mathbf{m}_x, \mathbf{R}_x)$ e o vector $\mathbf{Y} = \mathbf{A}\mathbf{X}$ temos que

$$\phi_Y(\mathbf{w}) = E[\exp(j\mathbf{w}^T \mathbf{Y})] = E[\exp(j\mathbf{v}^T \mathbf{X})] \quad (4.10)$$

onde $\mathbf{w}^T \mathbf{A} = \mathbf{v}^T$, i.e., $\mathbf{v} = \mathbf{A}^T \mathbf{w}$. Usando (4.7) como função característica de \mathbf{X} podemos escrever

$$\phi_Y(\mathbf{w}) = \exp[j\mathbf{w}^T \mathbf{A}\mathbf{m}_x - \frac{1}{2}\mathbf{w}^T \mathbf{A}\mathbf{R}_x \mathbf{A}^T \mathbf{w}] \quad (4.11)$$

que não é outra que a função característica de um vector aleatório gaussiano \mathbf{Y} de média e covariância

$$\begin{aligned} \mathbf{m}_y &= \mathbf{A}\mathbf{m}_x \\ \mathbf{R}_y &= \mathbf{A}\mathbf{R}_x \mathbf{A}^T \end{aligned}$$

Este resultado é fundamental devido á abundância da distribuição gaussiana na descrição de sinais aleatórios, por razões que já referimos, e da importância dos sistemas lineares em processamento do sinal. Devemos no entanto frisar que este resultado só é válido se as componentes do vector forem conjuntamente gaussianas. Isto significa que duas VAs X e Y podem ser marginalmente gaussianas sem ser conjuntamente gaussianas. Nesse caso uma VA bidimensional $Z = X + Y$ por exemplo pode não ser gaussiana. Notemos ainda que se duas VA's marginalmente gaussianas podem não ter uma distribuição conjunta gaussiana o inverso não é verdade, i.e., uma densidade conjunta gaussiana implica densidades marginais gaussianas.

4.5 Distribuições condicionais e regressão

Considere dois vectores aleatórios \mathbf{X} e \mathbf{Y} de média nula e conjuntamente gaussianos. A densidade de probabilidade conjunta do par \mathbf{X}, \mathbf{Y} escreve-se

$$p(\mathbf{x}, \mathbf{y}) = (2\pi)^{-n/2} [\det(\mathbf{R})]^{-1/2} \exp[-(1/2)(\mathbf{x}^T \mathbf{M}_x \mathbf{x} + 2\mathbf{x}^T \mathbf{M}_{x,y} \mathbf{y} + \mathbf{y}^T \mathbf{M}_y \mathbf{y})] \quad (4.12)$$

onde \mathbf{R} é simétrica

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{pmatrix} \quad (4.13)$$

e onde as matrizes \mathbf{M} podem ser calculadas a partir de \mathbf{R} . A densidade de probabilidade condicional de \mathbf{X} conhecendo \mathbf{y} é dada por (3.27) onde $p(\mathbf{y})$ é a densidade de probabilidade marginal em relação a \mathbf{Y} que é também uma densidade gaussiana visto que a densidade conjunta do par \mathbf{X}, \mathbf{Y} é gaussiana. Assim $\mathbf{Y} : N(\mathbf{0}, \mathbf{R}_{yy})$. Como o quociente de duas exponenciais dá ainda uma exponencial onde os expoentes se subtraem, usando (3.27), deduz-se que a densidade de probabilidade condicional $p(\mathbf{X}|\mathbf{y})$ tem uma forma gaussiana onde o expoente da exponencial é a diferença entre duas formas quadráticas que é ainda uma forma quadrática. Portanto considerando o vector aleatório $\mathbf{x}|\mathbf{y}$ como $N(\mathbf{m}(\mathbf{y}), \mathbf{R}_{\mathbf{x}|\mathbf{y}}(\mathbf{y}))$ que é obviamente uma função de \mathbf{y} .

A média condicional é a regressão $\mathbf{r}(\mathbf{y}) = E[\mathbf{x}|\mathbf{y}]$ que é uma função do vector \mathbf{y} . Substituindo a média \mathbf{m} por $\mathbf{r}(\mathbf{y})$ em (4.6) deverá obter-se uma forma quadrática em \mathbf{x}, \mathbf{y} semelhante a (4.12). Poderemos na generalidade escrever que $\mathbf{r}(\mathbf{y})$ é uma *função* linear de \mathbf{y} , i.e., que

$$\mathbf{r}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{c} \quad (4.14)$$

Calculando a esperança matemática em relação a \mathbf{y} de ambos os lados de (4.14) obtemos que

$$E_Y[\mathbf{r}(\mathbf{y})] = \mathbf{A}E[\mathbf{y}] + \mathbf{c} \quad (4.15)$$

sabendo que $E_Y[\mathbf{r}(\mathbf{y})] = E[\mathbf{x}]$ segundo (3.30) e como no nosso caso $E[\mathbf{x}] = E[\mathbf{y}] = \mathbf{0}$ obtemos que $\mathbf{c} = \mathbf{0}$. Podemos agora determinar a matriz \mathbf{A} usando a propriedade de decorrelação entre $\mathbf{x} - \mathbf{r}(\mathbf{y})$ e qualquer função $\mathbf{g}(\mathbf{y})$ e em particular com $\mathbf{g}(\mathbf{y}) = \mathbf{y}$ que se representa dizendo que

$$E\{[\mathbf{x} - \mathbf{r}(\mathbf{y})]\mathbf{y}^T\} = \mathbf{0} \quad (4.16)$$

substituindo $\mathbf{r}(\mathbf{y}) = \mathbf{A}\mathbf{y}$ obtemos facilmente que

$$\mathbf{A} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} \quad (4.17)$$

Em conclusão chegamos ao resultado que a esperança matemática condicional entre dois vectores aleatórios gaussianos é linear e escreve-se

$$\mathbf{r}(\mathbf{y}) = E[\mathbf{x}|\mathbf{y}] = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{y} \quad (4.18)$$

A matriz de covariância condicional é a outra característica (com a média) da distribuição condicional de vectores aleatórios gaussianos. Esta matriz pode ser definida por

$$V[\mathbf{x}|\mathbf{y}] = E[(\mathbf{x} - \mathbf{A}\mathbf{y})(\mathbf{x} - \mathbf{A}\mathbf{y})^T|\mathbf{y}] \quad (4.19)$$

como acabamos de demonstrar em (4.16) que $\mathbf{x} - \mathbf{A}\mathbf{y}$ e \mathbf{y} são decorrelacionados então temos que $V[\mathbf{x}|\mathbf{y}]$ é independente de \mathbf{y} e portanto

$$V[\mathbf{x}|\mathbf{y}] = E[(\mathbf{x} - \mathbf{A}\mathbf{y})(\mathbf{x} - \mathbf{A}\mathbf{y})^T] = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} \quad (4.20)$$

Em resumo pode-se dizer que quando dois vectores \mathbf{x} e \mathbf{y} são conjuntamente gaussianos de média nula a sua densidade de probabilidade condicional é ainda gaussiana e toma a forma

$$N(\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}, \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}).$$

4.6 Vectores aleatórios gaussianos complexos

Por definição podemos dizer que um vector aleatório complexo \mathbf{z} é gaussiano se as suas partes real e imaginária, respectivamente \mathbf{x} e \mathbf{y} são conjuntamente gaussianas. Se ambas as partes real e imaginária forem de média nula então poderemos dizer que o par de vectores \mathbf{x}, \mathbf{y} se encontra completamente definido através da matriz de covariância (4.13). Se os vectores \mathbf{x} e \mathbf{y} forem ambos de dimensão n então \mathbf{R} será de dimensão $2n$ e pode ser separada em quatro matrizes reais de dimensão $n \times n$. Se definirmos $\mathbf{z} = \mathbf{x} + j\mathbf{y}$ então $\mathbf{R}_z = E[\mathbf{z}\mathbf{z}^H]$ é igual a

$$\mathbf{R}_z = \mathbf{A} + j\mathbf{B} = \mathbf{R}_x + \mathbf{R}_y + j(-\mathbf{R}_{xy} + \mathbf{R}_{yx}) \quad (4.21)$$

podemos desde logo notar que o conhecimento das matrizes \mathbf{A} e \mathbf{B} não é suficiente para podermos determinar sem ambiguidade as matrizes $\mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_{xy}$ e \mathbf{R}_{yx} , i.e., um vector aleatório complexo não se encontra completamente determinado pela sua matriz de covariância pois temos de introduzir um segundo momento de segunda ordem que é

$$\mathbf{C} = E[\mathbf{z}\mathbf{z}^T] = \mathbf{P} + j\mathbf{Q} \quad (4.22)$$

de onde se deduz facilmente que $\mathbf{P} = \mathbf{R}_x - \mathbf{R}_y$ e $\mathbf{Q} = \mathbf{R}_{xy} + \mathbf{R}_{yx}$. Inversamente a partir de \mathbf{C} e \mathbf{R}_z podemos retirar

$$\mathbf{R}_x = \frac{1}{2}(\mathbf{A} + \mathbf{P}) \quad (4.23a)$$

$$\mathbf{R}_y = \frac{1}{2}(\mathbf{A} - \mathbf{P}) \quad (4.23b)$$

$$\mathbf{R}_{xy} = \frac{1}{2}(\mathbf{Q} - \mathbf{B}) \quad (4.23c)$$

e obviamente $\mathbf{R}_{xy} = \mathbf{R}_{yx}^T$. Em muitos casos teremos que $\mathbf{C} = \mathbf{0}$ o que simplifica muito as expressões acima

$$\mathbf{R}_x = \mathbf{R}_y = \frac{1}{2}\mathbf{A} \quad (4.24a)$$

$$\mathbf{R}_{xy} = -\mathbf{R}_{yx} = -\frac{1}{2}\mathbf{B} \quad (4.24b)$$

Vejamos o mesmo problema sob outro aspecto que é o de determinar qual a condição para que a densidade de probabilidade e a função característica que se escrevem em geral como funções $p(\mathbf{x}, \mathbf{y})$ e $\phi(\mathbf{u}, \mathbf{v})$ respectivamente

$$p(\mathbf{x}, \mathbf{y}) = (2\pi)^{-n/2} [\det(\mathbf{R})]^{-1/2} \exp [-(1/2)(\mathbf{x}^T \mathbf{M}_x \mathbf{x} + 2\mathbf{x}^T \mathbf{M}_{x,y} \mathbf{y} + \mathbf{y}^T \mathbf{M}_y \mathbf{y})] \quad (4.25)$$

e

$$\phi(\mathbf{u}, \mathbf{v}) = \exp\left[-\frac{1}{2}\mathbf{u}^T \mathbf{R}_x \mathbf{u} + \mathbf{v}^T \mathbf{R}_y \mathbf{v} + 2\mathbf{u}^T \mathbf{R}_{xy} \mathbf{v}\right] \quad (4.26)$$

possam tomar uma forma quadrática do tipo de (4.6) e (4.7) respectivamente, usando os vectores complexos $\mathbf{z} = \mathbf{x} + j\mathbf{y}$ e $\mathbf{w} = \mathbf{u} + j\mathbf{v}$. Comparando o expoente de (4.26) com a forma quadrática desejada

$$Q(\mathbf{w}) = \alpha \mathbf{w}^H \mathbf{R}_z \mathbf{w} = \alpha[\mathbf{u}^T \mathbf{A} \mathbf{u} + \mathbf{v}^T \mathbf{A} \mathbf{v} - 2\mathbf{u}^T \mathbf{B} \mathbf{v}] \quad (4.27)$$

deduzimos directamente que se e só se (4.24) se verificar é que poderemos igualar o expoente de (4.26) com (4.27) fazendo $\alpha = 1/2$. Por outras palavras se $\mathbf{C} = E[\mathbf{z}\mathbf{z}^T] = \mathbf{0}$ então poderemos escrever a função característica de \mathbf{z} como

$$\phi(\mathbf{w}) = \exp\left\{-(1/4)\mathbf{w}^H \mathbf{R}_z \mathbf{w}\right\} \quad (4.28)$$

Poderemos agora demonstrar que a redução á forma quadrática em função do vector complexo \mathbf{z} da densidade de probabilidade se faz mediante a mesma condição em $\mathbf{C} = \mathbf{0}$. Porém a demonstração é um pouco mais difícil pois necessita o calculo de \mathbf{R}_z^{-1} . A hipótese de $\mathbf{C} = \mathbf{0}$ leva a escrever usando (4.24)

$$\mathbf{R} = \frac{1}{2} \begin{pmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix} \quad (4.29)$$

sabendo que $\mathbf{R}\mathbf{R}^{-1} = \mathbf{I}$ podemos escrever

$$\mathbf{R}^{-1} = 2 \begin{pmatrix} \mathbf{V} & \mathbf{M} \\ \mathbf{M}^T & \mathbf{V} \end{pmatrix} \quad (4.30)$$

com

$$\begin{aligned} \mathbf{V} &= (\mathbf{A} + \mathbf{B}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ \mathbf{M} &= \mathbf{A}^{-1}\mathbf{B}\mathbf{V} \end{aligned}$$

com estes valores das sub-matrizes de \mathbf{R} obtemos o expoente da densidade de probabilidade $p(\mathbf{x}, \mathbf{y})$ sob a forma

$$q(\mathbf{x}, \mathbf{y}) = 2(\mathbf{x}^T \mathbf{V} \mathbf{x} + \mathbf{y}^T \mathbf{V} \mathbf{y} + 2\mathbf{x}^T \mathbf{M} \mathbf{y}) \quad (4.31)$$

como no caso da função característica (4.31) pode ser escrita em função do inverso da matriz \mathbf{R}_z que pode ser

$$\mathbf{R}_z^{-1} = \mathbf{T} + j\mathbf{S} \quad (4.32)$$

por comparação com (4.29) temos que $\mathbf{T} = \mathbf{V}$ e $\mathbf{S} = -\mathbf{M}$ e portanto podemos com estas definições escrever (4.31) sob a forma quadrática $q(\mathbf{x}, \mathbf{y}) = 2\mathbf{z}^H \mathbf{R}_z^{-1} \mathbf{z}$ onde $\mathbf{z} = \mathbf{x} + j\mathbf{y}$. O último passo necessário para podermos escrever finalmente a densidade de probabilidade do vector complexo \mathbf{z} será de se assegurar de que a forma obtida representa bem uma densidade de probabilidade, i.e., que $\int p(z)dz = 1$. Finalmente

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{z}) = \pi^{-n} [\det \mathbf{R}_z]^{-1} \exp(-\mathbf{z}^H \mathbf{R}_z^{-1} \mathbf{z}) \quad (4.33)$$

que é a densidade de probabilidade de um vector gaussiano complexo de média nula de covariância $E[\mathbf{z}\mathbf{z}^H] = \mathbf{R}_z$ e com $E[\mathbf{z}\mathbf{z}^T] = \mathbf{0}$. Todos os vectores gaussianos complexos que verificam esta última propriedade são chamados *circulares*.

4.7 Formas quadráticas de VA's normais multivariadas

Já vimos que transformações lineares de vectores aleatórios normais são ainda vectores aleatórios normais. E transformações quadráticas de vectores aleatórios normais? A resposta a esta pergunta reveste-se da maior importância na análise de funções energéticas de séries temporais de sinais aleatórios distribuídos normalmente.

Considere \mathbf{x} um vector aleatório de dimensão N e de média \mathbf{m} e covariância \mathbf{R} , que usualmente se nota $N(\mathbf{m}, \mathbf{R})$. Prova-se que a distribuição da forma quadrática

$$Q = (\mathbf{x} - \mathbf{m})^T \mathbf{R}^{-1} (\mathbf{x} - \mathbf{m}) \quad (4.34)$$

é chi-quadrado χ_N^2 . Para o efeito basta provar (o que se deixa como exercício) que a sua função característica se escreve

$$\phi(\omega) = E[e^{j\omega Q}] = \frac{1}{(1 - 2j\omega)^{N/2}} \quad (4.35)$$

o que é com efeito a função característica de uma variável distribuída segundo a lei do chi-quadrado com N graus de liberdade.

No caso mais geral da forma quadrática

$$Q = (\mathbf{x} - \mathbf{m})^T \mathbf{P} (\mathbf{x} - \mathbf{m}) \quad (4.36)$$

onde \mathbf{P} é uma matriz simétrica pode-se calcular a função característica de Q como

$$\phi(\omega) = \int e^{j\omega(\mathbf{x}-\mathbf{m})^T \mathbf{P}(\mathbf{x}-\mathbf{m})} (2\pi)^{N/2} (\det \mathbf{R})^{-1/2} \times \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^T \mathbf{R}^{-1} (\mathbf{x} - \mathbf{m}) \right\} d\mathbf{x} \quad (4.37)$$

reagrupando os termos do expoente da exponencial podemos escrever

$$\phi(\omega) = \int (2\pi)^{N/2} (\det \mathbf{R})^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^T (\mathbf{I} - 2j\omega \mathbf{P}\mathbf{R}) \mathbf{R}^{-1} (\mathbf{x} - \mathbf{m}) \right\} d\mathbf{x} \quad (4.38)$$

de forma a obter um resultado unitário do integral temos que multiplicar e dividir pela raiz quadrada do determinante do factor matricial existente no expoente da exponencial, i.e., $\det (\mathbf{I} - 2j\omega \mathbf{P}\mathbf{R})$ que em seguida sai do integral formando o resultado final

$$\phi(\omega) = \frac{1}{[\det (\mathbf{I} - 2j\omega \mathbf{P}\mathbf{R})]^{1/2}} \quad (4.39)$$

daqui, e usando (3.45), podemos determinar os primeiros momentos de Q

$$E[Q] = \text{tr } \mathbf{P}\mathbf{R} \quad (4.40a)$$

$$V[Q] = 2 \operatorname{tr} (\mathbf{PR})^2 \quad (4.40b)$$

No caso particular e importante em que \mathbf{PR} é simétrica a função característica pode-se escrever

$$\phi(\omega) = \frac{1}{\prod_{n=1}^N (1 - 2j\omega\lambda_n)^{1/2}} \quad (4.41)$$

onde os λ_n são os valores próprios de \mathbf{PR} . Comparando (4.41) com (4.35) podemos notar que (4.41) só será a função característica de uma distribuição χ_r^2 se

$$\lambda_n = \begin{cases} 1, & n = 1, 2, \dots, r \\ 0, & n = r + 1, \dots, N \end{cases} \quad (4.42)$$

o que significa que neste caso \mathbf{PR} é ortogonalmente similar a uma matriz identidade de ordem deficiente e que \mathbf{PR} é idempotente, i.e., que $(\mathbf{PR})(\mathbf{PR}) = \mathbf{PR}$. Uma tal matriz é chamada uma matriz de projecção (ver próximo capítulo). Em particular, se $\mathbf{R} = \mathbf{I}$, i.e., se as componentes de \mathbf{x} forem independentes então Q é χ_r^2 se e só se \mathbf{P} for idempotente, $\mathbf{P}^2 = \mathbf{P}$.

Em resumo, temos que a transformação linear

$$\mathbf{y} = \mathbf{P}\mathbf{x} \quad (4.43)$$

é tal que $\mathbf{y} : N(\mathbf{0}, \mathbf{P})$ e a forma quadrática

$$\mathbf{y}^T \mathbf{y} = \mathbf{x}^T \mathbf{P}\mathbf{x} \quad (4.44)$$

segue uma distribuição χ_r^2 quando $\mathbf{x} : N(\mathbf{0}, \mathbf{I})$ e \mathbf{P} é uma matriz de projecção de ordem r .

5. Elementos de álgebra linear

Desde que a discretização da informação passou a ser um factor essencial na sua análise e manipulação por computadores as observações passaram a ser expressas sobre forma de vectores e matrizes. Assim a algebra matricial passou a desempenhar um papel crucial na sua manipulação e inferência de propriedades espaciais, temporais e frequências.

Neste capítulo passaremos em revista as noções de algebra elementar necessárias para a abordagem desta disciplina.

5.1 Espaços vectoriais

Frequentemente lidamos com conjuntos ordenados de pontos da forma (x_1, x_2, \dots, x_n) que podem, por exemplo, designar as coordenadas dum ponto no espaço a n dimensões. Correntemente designamos estes conjuntos por vectores que neste caso seria,

$$\mathbf{x} = [x_1, x_2, \dots, x_n]^T \quad (5.1)$$

Notaremos \mathfrak{R}^n o conjunto de todos os vectores reais e diremos que $\mathbf{x} \in \mathfrak{R}^n$. As propriedades que caracterizam um espaço vectorial linear e real são, para qualquer escalar $a \in \mathfrak{R}$ e vectores \mathbf{x} e \mathbf{y} :

$$\mathbf{x} \in \mathfrak{R}^n \rightarrow a\mathbf{x} \in \mathfrak{R}^n$$

$$\mathbf{x} \text{ e } \mathbf{y} \in \mathfrak{R}^n \rightarrow \mathbf{x} + \mathbf{y} \in \mathfrak{R}^n$$

Da mesma forma uma matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ cujas colunas são formadas por p vectores de dimensão n diz-se que pertence ao espaço vectorial linear $\mathfrak{R}^{n \times p}$.

Espaço Euclidiano

A noção de espaço Euclidiano obtém-se simplesmente através do espaço vectorial real e linear no qual se define a norma Euclidiana

$$\|\mathbf{x}\| = (\mathbf{x}^T \mathbf{x})^{1/2} = \left(\sum_{i=1}^n x_i^2 \right)^{1/2} \quad (5.2)$$

Espaço de Hilbert

Trata-se aqui do espaço vectorial que servirá de suporte para todo o trabalho que efectuaremos em teoria do sinal. Um espaço de Hilbert é um espaço vectorial Euclidiano no qual todas as sequências de Cauchy convergem e no qual é definido o produto interno,

$$\mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i \quad (5.3)$$

quando $\mathbf{x}^T \mathbf{y} = 0$ dizemos que os vectores \mathbf{x} e \mathbf{y} são ortogonais.

5.2 Independência linear

Podemos definir um vector \mathbf{x} como uma combinação linear de outros vectores do mesmo espaço vectorial (obtendo um vector ainda no mesmo espaço segundo 5.1) e assim

$$\mathbf{x} = a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + \dots + a_p \mathbf{x}_p \quad (5.4)$$

A noção de independência linear está intimamente ligada ao facto de que se um dos vectores nesta soma for, ele mesmo, uma combinação linear de um ou mais dos outros vectores, \mathbf{x} pode ser gerado sem ele e diz-se então que os vectores $\mathbf{x}_i, i = 1, \dots, p$ são linearmente dependentes.

Formalmente, o conjunto de vectores $\mathbf{x}_i, i = 1, \dots, p$ são ditos linearmente independentes quando

$$a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + \dots + a_p \mathbf{x}_p = \mathbf{0} \quad (5.5)$$

implique $a_i = 0, i = 1, \dots, p$.

Na prática podemos verificar se um determinado conjunto de vectores é linearmente independente calculando o determinante da matriz de Gram, G , definida por

$$\mathbf{G} = \mathbf{X}^T \mathbf{X} \quad (5.6)$$

onde a matriz \mathbf{X} , $n \times p$ é definida por $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p]$. Se $\det \mathbf{G} \neq 0$ então os vectores $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ são linearmente independentes.

Outro resultado muito útil é o teorema que diz que todo o vector em \mathfrak{R}^n pode ser representado como uma combinação linear de n vectores linearmente independentes em \mathfrak{R}^n . Um corolário deste teorema é de que não podem existir mais do que n vectores linearmente independentes em \mathfrak{R}^n .

5.3 Sub-espços vectoriais lineares, bases, dimensões e ordem

Sub-espços vectoriais

Considere um conjunto de vectores $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$, não necessariamente linearmente independentes. O sub-espço gerado por este conjunto de vectores é o conjunto de vectores $\mathbf{x} \in \mathfrak{R}^n$ que podem ser gerados por combinações lineares do conjunto,

$$S(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p) = \left(\mathbf{x} : \mathbf{x} = \sum_{i=1}^p a_i \mathbf{x}_i \right) \quad (5.7)$$

Como é óbvio, um sub-espço vectorial é ele mesmo um espaço vectorial pois obedece aos lemas expostos em 5.1.

Bases

Se, do conjunto de vectores $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$, forem retirados todos aqueles que não forem linearmente independentes (através de um procedimento tipo *ortogonalização de Gram-Schmidt* por exemplo) podemos dizer que os r vectores restantes com $r \leq p$ formam uma base do sub-espaço S . Assim todo e qualquer vector

$$\mathbf{x} \in S \quad \text{tal que} \quad \mathbf{x} = \sum_{i=1}^r a_i \mathbf{x}_i \quad (5.8)$$

Se os vectores que formam a base são mútualmente ortogonais então dizemos que estes formam uma base ortogonal do sub-espaço S .

Podemos ainda dizer que a representação de qualquer vector no sub-espaço S é única, i.e., existe um e só um conjunto de coeficientes a_1, a_2, \dots, a_r tal que

$$\mathbf{x} = \sum_{i=1}^r a_i \mathbf{x}_i \quad (5.9)$$

Dimensões e ordem

Considere o sub-espaço S gerado pelo conjunto de vectores $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$ tais que

$$S = \left\{ \mathbf{x} : \mathbf{x} = \sum_{i=1}^p a_i \mathbf{x}_i \right\} \quad (5.10)$$

A dimensão do sub-espaço S , notada como $\dim S$ é definida pelo número de vectores linearmente independentes no conjunto $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$.

De forma similar, se formarmos a matriz

$$\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_p]$$

a dimensão do espaço S é tão sómente igual ao número de colunas de \mathbf{X} linearmente independentes que toma o nome de *ordem* da matriz \mathbf{X} e escreve-se $r(\mathbf{X})$ (rank of \mathbf{X}).

5.4 Matrizes hermitianas

As matrizes hermitianas são caracterizadas por uma propriedade: o seu transposto conjugado é igual a ela mesma. Assim a matriz hermitiana \mathbf{R} é tal que $(\mathbf{R}^T)^* = \mathbf{R}$. Frequentemente nota-se transposto-conjugado por $(\mathbf{R}^T)^* = \mathbf{R}^H$. Algumas propriedades:

1) os valores próprios de uma matriz hermitiana são reais: de facto uma matriz hermitiana deverá ter a sua diagonal principal real. Na decomposição

$$\mathbf{R}\mathbf{u} = \lambda\mathbf{u} \quad (5.11)$$

sendo \mathbf{R} uma matriz hermitiana implica $\lambda \in \Re$.

2) os vectores próprios de uma matriz hermitiana são ortogonais, i.e., na decomposição

$$\mathbf{R}\mathbf{u}_i = \lambda_i\mathbf{u}_i \quad (5.12a)$$

e

$$\mathbf{R}\mathbf{u}_j = \lambda_j\mathbf{u}_j \quad (5.12b)$$

temos que $\mathbf{u}_i^H\mathbf{u}_j = 0$ se $\lambda_i \neq \lambda_j$.

3) uma matriz hermitiana é diagonalizável, i.e., pode-se por sob a forma

$$\mathbf{U}^H\mathbf{R}\mathbf{U} = \mathbf{\Lambda} \quad (5.13)$$

onde $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ e $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n]$ tal que

$$\mathbf{R}\mathbf{u}_i = \lambda_i\mathbf{u}_i; i = 1, \dots, n \quad (5.14)$$

De forma similar podemos escrever a decomposição em valores próprios da matriz hermitiana \mathbf{R} como

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H = \sum_{i=1}^n \lambda_i\mathbf{u}_i\mathbf{u}_i^H \quad (5.15)$$

5.5 Decomposição em valores singulares

A decomposição em valores singulares (Singular Value Decomposition = SVD) tem o mesmo papel para matrizes rectangulares do que a decomposição em valores próprios para matrizes quadradas. Neste sentido a SVD tem um papel importantíssimo na definição da característica de uma matriz singular.

Considere uma matriz rectangular \mathbf{H} de dimensão $n \times p$. Vamos supor, no caso geral, que a ordem de \mathbf{H} é $r(\mathbf{H}) \leq p \leq n$, assim \mathbf{H} é uma matriz de característica deficiente e portanto singular. A SVD permite escrever

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H \quad (5.16)$$

onde

$$\mathbf{U} = [\mathbf{U}_1\mathbf{U}_2] \quad \mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2] \quad (5.17)$$

pode-se provar que ambas as matrizes \mathbf{U} e \mathbf{V} são matrizes unitárias de dimensões respectivas $n \times n$ e $p \times p$ e que \mathbf{V} não é mais do que a matriz cujas colunas são os vectores próprios da matriz de Gram $\mathbf{H}^H\mathbf{H}$. A matriz $\mathbf{\Sigma}$ é obviamente de dimensão $n \times p$

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & \mathbf{\Sigma}_2 \end{bmatrix} \quad (5.18)$$

com $\mathbf{\Sigma}_1 = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)$ e $\mathbf{\Sigma}_2 = \text{diag}(0, \dots, 0)$. Os σ_i são obviamente os valores singulares de \mathbf{H} .

Frequentemente é utilizada uma SVD, dita "económica", definida por

$$\mathbf{H} = \mathbf{U}_e \mathbf{\Sigma} \mathbf{V}_e^H \quad (5.19)$$

No caso sobre determinado, *i.e.* $n > p$, temos que as dimensões respectivas de \mathbf{U}_e , $\mathbf{\Sigma}_e$ e \mathbf{V}_e^H são $n \times p$, $p \times p$ e $p \times p$. Neste caso temos que $\mathbf{V}_e = \mathbf{V}$, que $\mathbf{\Sigma}_e = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p)$ e que $\mathbf{U}_e^H \mathbf{U}_e = \mathbf{V}^H \mathbf{V} = \mathbf{V} \mathbf{V}^H = \mathbf{I}$. No caso sub determinado, *i.e.* $n < p$, temos que as dimensões respectivas de \mathbf{U}_e , $\mathbf{\Sigma}_e$ e \mathbf{V}_e^H são $n \times n$, $n \times n$ e $n \times p$. Neste caso temos que $\mathbf{U}_e = \mathbf{U}$, que $\mathbf{\Sigma}_e = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ e que $\mathbf{U}^H \mathbf{U} = \mathbf{U} \mathbf{U}^H = \mathbf{V}_e^H \mathbf{V}_e = \mathbf{I}$.

A decomposição em valores singulares - SVD - enunciada acima, permite resolver um problema muito importante. Considere uma matriz \mathbf{H} de ordem p e suponha que queremos determinar a matriz \mathbf{H}_r de ordem r mais "próxima" de \mathbf{H} ou por outras palavras que minimiza o erro quadrático (soma dos quadrados dos elementos)

$$e^2 = \text{tr}(\mathbf{H} - \mathbf{H}_r)^H (\mathbf{H} - \mathbf{H}_r) \quad (5.19)$$

se substituirmos as duas matrizes pelas respectivas SVDs temos que

$$\begin{aligned} e^2 &= \text{tr}(\mathbf{V}(\mathbf{\Sigma} - \mathbf{\Sigma}_r)^H \mathbf{U}^H \mathbf{U} (\mathbf{\Sigma} - \mathbf{\Sigma}_r) \mathbf{V}^H) \\ &= \sum_{i=r+1}^p \sigma_i^2 |v_{i,i}|^2 \end{aligned} \quad (5.20)$$

onde $|v_{i,i}|^2 = 1$. Podemos provar que o erro $(\mathbf{H} - \mathbf{H}_r)$ é ortogonal a \mathbf{H}_r da seguinte forma

$$\begin{aligned} (\mathbf{H} - \mathbf{H}_r)^H \mathbf{H}_r &= \mathbf{V}(\mathbf{\Sigma} - \mathbf{\Sigma}_r)^H \mathbf{\Sigma}_r \mathbf{V}^H \\ &= \mathbf{0} \end{aligned} \quad (5.21)$$

assim se o erro é ortogonal a \mathbf{H}_r podemos então dizer que $\mathbf{H}_r = \mathbf{U} \mathbf{\Sigma}_r \mathbf{V}^H$ é a melhor aproximação no sentido dos mínimos quadrados de ordem r de \mathbf{H} . Neste caso não se trata de mais do que uma generalização do teorema da projecção ortogonal da matriz \mathbf{H} no sub-espço de dimensão r gerado pelas colunas de \mathbf{H}_r (ver 3.6, fig. 3.1 e respectiva discussão).

5.6 Projecções, rotações e pseudoinversos

Projecções, rotações e pseudoinversos são as três mais usadas transformações matriciais. Começemos por considerar duas matrizes \mathbf{H} e \mathbf{A} de ordem completa e de dimensões respectivas $n \times p$ e $(n-p) \times p$, tais que $\mathbf{A}^T \mathbf{H} = \mathbf{0}$. Diz-se neste caso que estas duas matrizes formam uma decomposição ortogonal de \mathfrak{R}^n e portanto que $[\mathbf{H} \mathbf{A}]$ é uma matriz $n \times n$ de ordem completa $= n$. Para facilitar e para introduzir desde já uma terminologia próxima do processamento do sinal, diremos que o sub-espço

gerado por \mathbf{H} , que notaremos $\langle \mathbf{H} \rangle$, será o *espaço do sinal* e que o sub-espaço gerado por \mathbf{A} , $\langle \mathbf{A} \rangle$, será o seu *sub-espaço ortogonal*

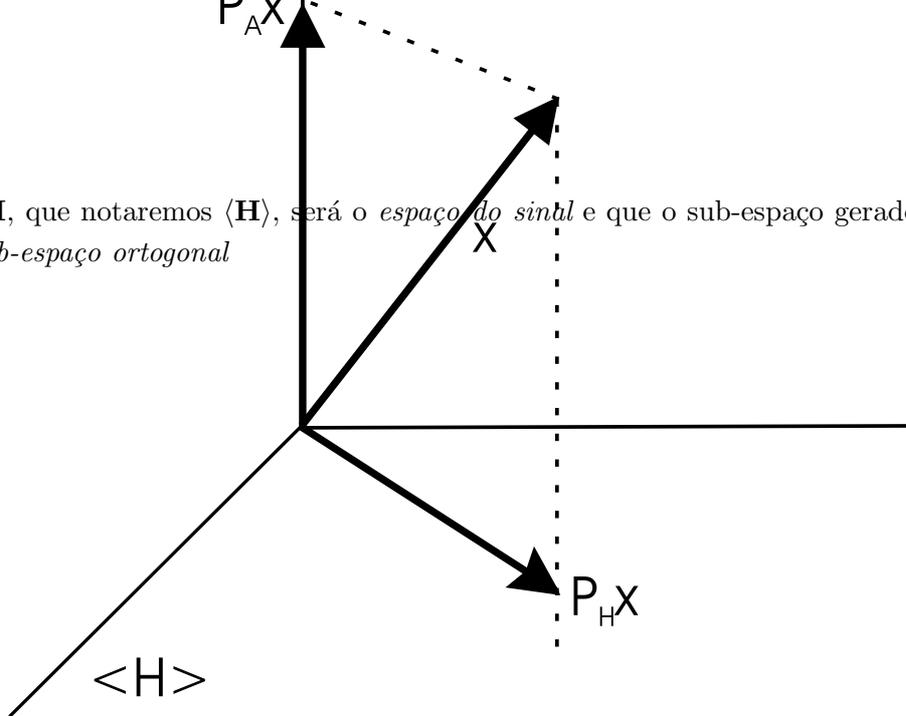


Figura 5.1: sub-espaços e projecções.

Projeções

Considere agora o problema da decomposição de um vector \mathbf{x} de \mathfrak{R}^n segundo duas componentes ortogonais uma em $\langle \mathbf{H} \rangle$ e outra em $\langle \mathbf{A} \rangle$. A forma correcta é

$$\mathbf{x} = \mathbf{P}_H \mathbf{x} + \mathbf{P}_A \mathbf{x} \quad (5.22)$$

onde \mathbf{P}_H e \mathbf{P}_A são as matrizes de projecção respectivamente em $\langle \mathbf{H} \rangle$ e em $\langle \mathbf{A} \rangle$. Estas podem-se encontrar facilmente notando que $\mathbf{P}_H \mathbf{x}$ e $\mathbf{x} - \mathbf{P}_H \mathbf{x}$ são ortogonais e que $\mathbf{P}_A = \mathbf{I} - \mathbf{P}_H$ (figura 5.1). Assim

$$\mathbf{P}_H = \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \quad (5.23)$$

e

$$\mathbf{P}_A = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \quad (5.24)$$

observando obviamente todas as propriedades das matrizes de projecção em geral, i.e., que $\mathbf{P}_{(\cdot)}^T = \mathbf{P}_{(\cdot)}$, $\mathbf{P}_A + \mathbf{P}_H = \mathbf{I}$ e que $\mathbf{P}_H \mathbf{P}_A = \mathbf{0}$. Se usarmos as SVD's de \mathbf{A} e \mathbf{H} obtemos as formas simplificadas

$$\mathbf{P}_H = \mathbf{U}_H \mathbf{U}_H^T \quad \mathbf{P}_A = \mathbf{U}_A \mathbf{U}_A^T \quad (5.25)$$

Rotações

A rotação de um vector \mathbf{x} em $\langle \mathbf{H} \rangle$ ou de forma equivalente em torno de $\langle \mathbf{A} \rangle$ pode definir-se sob a forma $\mathbf{Q}_H \mathbf{x}$ onde \mathbf{Q}_H é uma matriz de rotação construída a partir de \mathbf{P}_H e \mathbf{P}_A e uma matriz ortogonal \mathbf{Q} ,

$$\mathbf{Q}_H = \mathbf{U}_H \mathbf{Q} \mathbf{U}_H^T + \mathbf{P}_A \quad (5.26)$$

Pseudoinversos

O pseudoinverso de \mathbf{H} (também chamado inverso de Pen-Rose) escreve-se,

$$\mathbf{H}^\# = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \quad (5.27)$$

e que não é mais do que o inverso da SVD de \mathbf{H} ,

$$\mathbf{H}^\# = \mathbf{V}_H \mathbf{\Lambda}_H^{-1} \mathbf{U}_H^T \quad (5.28)$$

e tem as seguintes propriedades fundamentais

$$\mathbf{H}^\# \mathbf{H} = \mathbf{I} \quad (5.29a)$$

$$\mathbf{H} \mathbf{H}^\# = \mathbf{P}_H \quad (5.29b)$$

$$\mathbf{H} \mathbf{H}^\# \mathbf{H} = \mathbf{H} \quad (5.29c)$$

$$\mathbf{H}^\# \mathbf{H} \mathbf{H}^\# = \mathbf{H}^\# \quad (5.29d)$$

Os problemas e as soluções

Mais à frente nesta disciplina vamos provar a utilidade das definições acima, na solução dos seguintes problemas:

1. Problemas de mínimos quadrados: demonstraremos que a solução do seguinte problema de mínimos quadrados

$$\min_{\theta} (\mathbf{y} - \mathbf{H}\theta)^T (\mathbf{y} - \mathbf{H}\theta) \quad (5.30)$$

é dada por

$$\hat{\theta} = \mathbf{H}^\# \mathbf{y} \quad (5.31)$$

5. Aproximação entre vectores: a projecção \mathbf{P}_H soluciona o seguinte problema de aproximação (como aliás já vimos),

$$\min_{\mathbf{x}: \mathbf{x} = \mathbf{H}\mathbf{y}} (\mathbf{y} - \mathbf{x})^T (\mathbf{y} - \mathbf{x}) \quad (5.32)$$

onde neste caso

$$\hat{\mathbf{x}} = \mathbf{P}_H \mathbf{y} \quad (5.33)$$

5.7 Formas quadráticas

Por definição chama-se forma quadrática em \mathbf{Q} o escalar

$$f = \mathbf{x}^T \mathbf{Q} \mathbf{x}; \quad \mathbf{x} \in \mathfrak{R}^n, \quad \mathbf{Q} \in \mathfrak{R}^{n \times n} \quad (5.34)$$

A matriz \mathbf{Q} é definida não negativa se $f \geq 0$ para $\mathbf{x} \neq 0$ e definida positiva se $f > 0$. Formas quadráticas aparecem frequentemente na prática da estimação e detecção pois são equivalentes

a uma energia. Neste sentido é frequente a necessidade de minimizar (ou maximizar) formas normalizadas do tipo do quociente de Rayleigh,

$$R(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \quad (5.35)$$

em relação a \mathbf{x} . É relativamente fácil de provar que $R(\mathbf{x})$ é minimizado pelo primeiro vector próprio de \mathbf{A} , $\mathbf{x} = \mathbf{e}_1$, assim

$$R(\mathbf{e}_1) = \frac{\mathbf{e}_1^T \mathbf{A} \mathbf{e}_1}{\mathbf{e}_1^T \mathbf{e}_1} = \frac{\mathbf{e}_1^T \mathbf{E} \mathbf{\Lambda} \mathbf{E}^T \mathbf{e}_1}{\mathbf{e}_1^T \mathbf{e}_1} \quad (5.36)$$

e como os vectores próprios são ortogonais entre si temos que

$$R(\mathbf{e}_1) = \frac{\mathbf{e}_1^T \lambda_1 \mathbf{e}_1}{\mathbf{e}_1^T \mathbf{e}_1} = \lambda_1 \quad (5.37)$$

que é o menor dos valores próprios de \mathbf{A} . Existem muitas outras propriedades que se aplicam às formas quadráticas e que podem ser consultadas em Scharf [1], pp.51-54 e Strang [2] pp.347-352.

6. Mínimos quadrados (MQ)

Depois dos capítulos introdutórios anteriores vamos agora abordar a teoria da estimação através de uma aproximação puramente determinística. Óbviamente os sinais com que trabalharemos serão sempre aleatórios e, em parte, desconhecidos (senão não seria necessário estimar nada!...) mas as suas propriedades podem ser conhecidas ou desconhecidas á priori. Neste capítulo partiremos do princípio de que as suas propriedades estatísticas são desconhecidas antes de dar início a qualquer tentativa de estimação. O critério adoptado será o dos mínimos quadrados e a ideia é de ajustar um modelo nos dados de tal forma que os erros entre o modelo e os dados sejam minimizados precisamente no sentido dos mínimos quadrados.

Em geral, o modelo utilizado poderá ser do tipo linear ou não linear. Aqui, trataremos em profundidade apenas o caso linear e o caso linear com condicionamentos lineares ou não lineares. O caso quadrático será abordado superficialmente.

6.1 O modelo linear

Segundo o modelo linear, o vector de observações \mathbf{y} de dimensão N , é obtido como a soma de um vector de sinal \mathbf{x} mais um vector de ruído \mathbf{n}

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \quad (6.1)$$

mais, faz-se a hipótese de que

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} \quad (6.2)$$

onde $\boldsymbol{\theta}$ é um vector de parâmetros, desconhecido, de dimensão p em geral $< N$. A matriz \mathbf{H} , $N \times p$, pode ser conhecida, só parcialmente conhecida ou totalmente desconhecida. A interpretação deste modelo é de que \mathbf{y} representa o sinal observado, \mathbf{x} o sinal procurado e \mathbf{n} o ruído de tipo aditivo. De alguma forma \mathbf{n} representa os erros de medida de \mathbf{x} . A forma particular de $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}$ consiste a particularidade do modelo linear e é esta estrutura que nos permite dizer que

$$\mathbf{x} = \sum_{n=1}^p \theta_n \mathbf{h}_n \quad (6.3)$$

onde os \mathbf{h}_n são os vectores coluna de \mathbf{H} e os θ_n as componentes do vector de parâmetros $\boldsymbol{\theta}$. (6.3) significa efectivamente que o sinal \mathbf{x} é uma combinação linear das colunas de \mathbf{H} (chamados frequentemente “modos”). Os pesos desta combinação linear são precisamente os parâmetros que desejamos determinar. Uma forma alternativa, interessante em alguns problemas, consiste em escrever (6.3) em função, não das colunas, mas das linhas da \mathbf{H} que notaremos \mathbf{c} ,

$$x_n = \mathbf{c}_n^T \boldsymbol{\theta} \quad (6.4)$$

No estudo de modelos lineares devemos distinguir três casos: o caso sobredeterminado, o caso determinado e o caso subdeterminado, segundo os valores relativos de p e N , respectivamente, $p < N$, $p = N$ e $p > N$.

No caso sobredeterminado, $p < N$, temos mais equações que incógnitas e um ajuste exacto do modelo aos dados é impossível. Este é o caso mais encontrado na prática ao qual se adapta a teoria de ajuste dos mínimos quadrados.

No caso determinado, $p = N$, pode haver um ajuste exacto entre o modelo e as observações desde que estas sejam independentes, i.e., desde que as colunas de \mathbf{H} sejam linearmente independentes.

No caso subdeterminado, $p > N$, há mais parâmetros do que observações tornando impossível determinar $\boldsymbol{\theta}$ de forma única. Neste caso é necessário impor condições adicionais sob os parâmetros para estes poderem ser determinados a partir de \mathbf{y} .

A solução do problema exposto acima em torno a (6.1) e (6.2) pode ser solucionado usando os mínimos quadrados independentemente da natureza estatística de $\boldsymbol{\theta}$ ou \mathbf{n} . No entanto, quando $\mathbf{n} : N(\mathbf{0}, \mathbf{R})$ é uma realização de um vector aleatório gaussiano, então dizemos que (6.1) é um modelo estatístico linear com erros normais ou gaussianos, ou mais simplesmente, um modelo de erros normais. De (6.1) deduzimos facilmente que $\mathbf{y} : N(\mathbf{H}\boldsymbol{\theta}, \mathbf{R})$ e a sua densidade é

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{N/2} \det(\mathbf{R})^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\boldsymbol{\theta}) \right\} \quad (6.5)$$

6.2 Soluções dos MQ

Considere de novo o modelo (6.1) e (6.2). Para uma dada estimativa $\boldsymbol{\theta}$ do vector de parâmetros o quadrado dos erros entre as observações \mathbf{y} e o modelo $\mathbf{H}\boldsymbol{\theta}$ escreve-se

$$\begin{aligned} e^2 &= \text{tr}[(\mathbf{y} - \mathbf{H}\boldsymbol{\theta})(\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T] \\ &= (\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{H}\boldsymbol{\theta}) \\ &= \mathbf{n}^T \mathbf{n} \end{aligned} \quad (6.6)$$

muito simplesmente, e sem nenhuma hipótese estatística, podemos derivar a expressão (6.6) em relação ao vector $\boldsymbol{\theta}$ obtendo

$$\frac{\partial}{\partial \boldsymbol{\theta}} e^2 = 2\mathbf{H}^T (\mathbf{y} - \mathbf{H}\boldsymbol{\theta}) \quad (6.7)$$

que quando equacionado a zero nos permite obter o valor de $\boldsymbol{\theta}$ que minimiza e^2 , i.e., o estimador dos mínimos quadrados

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{y} \quad (6.8)$$

Esta solução existe, e é única, se o inverso de $\mathbf{G} = \mathbf{H}^T \mathbf{H}$ existe, que não é mais do que a matriz de Gram já definida em (5.6). Se a matriz de Gram é não singular, a solução de $\hat{\boldsymbol{\theta}}$ é única. Se é singular temos muitas soluções possíveis. Como já vimos atrás \mathbf{G} será singular se e só se os vectores coluna de \mathbf{H} forem linearmente dependentes. Equivalentemente, \mathbf{G} será não singular se as colunas de \mathbf{H} forem linearmente independentes.

Projectções

Consideremos que \mathbf{H} é não singular. Então usando (6.8) podemos escrever o sinal estimado como

$$\hat{\mathbf{x}} = \mathbf{H}\hat{\boldsymbol{\theta}} \quad (6.9)$$

que podemos também escrever sob a forma

$$\hat{\mathbf{x}} = \mathbf{P}_H \mathbf{y} \quad (6.10)$$

por substituição de (6.8) em (6.9) e definindo $\mathbf{P}_H = \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$ como uma matriz de projecção ortogonal. A diferença entre as observações \mathbf{y} e o modelo estimado $\hat{\mathbf{x}}$ é o erro de ajuste que se escreve

$$\hat{\mathbf{n}} = \mathbf{y} - \hat{\mathbf{x}} = (\mathbf{I} - \mathbf{P}_H) \mathbf{y} = \mathbf{P}_A \mathbf{y} \quad (6.11)$$

onde $\mathbf{P}_A = \mathbf{I} - \mathbf{P}_H$. Estas duas matrizes \mathbf{P}_H e \mathbf{P}_A são matrizes de projecção que gozam de todas as propriedades já descritas no capítulo 5.6.

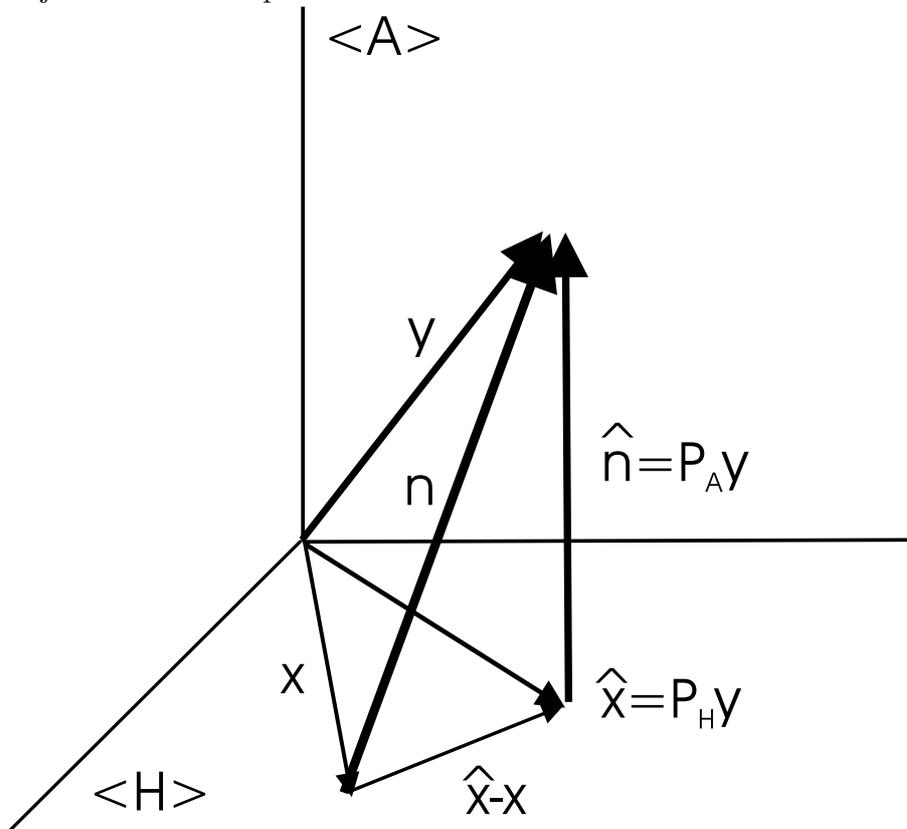


Figura 6.1: decomposição ortogonal de \mathbf{y} e \mathbf{n} .

Sub-espaço do sinal e seu ortogonal

A figura 6.1 mostra os projectores \mathbf{P}_H e \mathbf{P}_A . No plano horizontal da figura 6.1 estão representados todos os vectores que se podem escrever como combinação linear das colunas da matriz \mathbf{H} . Este é chamado sub-espaço do sinal que notaremos $\langle \mathbf{H} \rangle$. No eixo vertical da figura estão representados todos os vectores \mathbf{u} que são ortogonais às colunas de \mathbf{H} . Este é chamado sub-espaço ortogonal $\langle \mathbf{A} \rangle$.

Quando p colunas de \mathbf{H} são linearmente independentes, então, tendo em conta que o espaço total é de dimensão N , só poderemos encontrar $N - p$ vectores linearmente independentes e ortogonais às p colunas de \mathbf{H} . Organizando estes $N - p$ vectores como as colunas de uma matriz

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{N-p}] \quad (6.12)$$

podemos dizer que cada coluna de \mathbf{A} é ortogonal às colunas de \mathbf{H} , assim

$$\mathbf{a}_i^T \mathbf{h}_j = 0 \quad (6.13a)$$

$$\mathbf{A}^T \mathbf{H} = \mathbf{0} \quad (6.13b)$$

No sub-espaço ortogonal, cada vector \mathbf{u} é uma combinação linear dos \mathbf{a}_i , tal que

$$\mathbf{u} = \mathbf{A}\boldsymbol{\phi} \quad (6.14)$$

Visto tudo isto, e o que foi dito anteriormente sobre projecções ortogonais, podemos afirmar que a representação

$$\mathbf{I} = \mathbf{P}_H + \mathbf{P}_A = \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T + \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \quad (6.15)$$

separa o espaço Euclidiano de dimensão N em dois sub-espaços ortogonais o primeiro gerado pelas colunas de \mathbf{H} e o segundo gerado pelas colunas de \mathbf{A} . \mathbf{P}_H projecta no primeiro sub-espaço e \mathbf{P}_A projecta no segundo sub-espaço. Qualquer vector arbitrário \mathbf{y} pode ser representado pelas suas projecções num e noutro sub-espaço tal que

$$\begin{aligned} \mathbf{y} &= \mathbf{P}_H \mathbf{y} + \mathbf{P}_A \mathbf{y} \\ &= \hat{\mathbf{x}} + \hat{\mathbf{n}} \end{aligned} \quad (6.16)$$

onde $\hat{\mathbf{x}}$ e $\hat{\mathbf{n}}$ são o sinal e o ruído estimados que fazem parte do sub-espaço gerado por \mathbf{H} e \mathbf{A} respectivamente. Por esta razão, $\langle \mathbf{H} \rangle$ é chamado o sub-espaço do sinal e $\langle \mathbf{A} \rangle$ é chamado o sub-espaço do ruído.

A propriedade mais útil nesta decomposição das observações \mathbf{y} é o facto de que o ruído estimado $\hat{\mathbf{n}}$ é ortogonal a qualquer sinal pertencente a $\langle \mathbf{H} \rangle$. Esta propriedade escreve-se

$$\hat{\mathbf{n}}^T \hat{\mathbf{x}} = \mathbf{y}^T \mathbf{P}_A \mathbf{P}_H \mathbf{y} = 0. \quad (6.17)$$

O modulo ao quadrado do erro de ajustamento, ou seja o quadrado da diferença entre o sinal estimado e a observação é

$$\hat{\mathbf{n}}^T \hat{\mathbf{n}} = \mathbf{y}^T (\mathbf{I} - \mathbf{P}_H) \mathbf{y} = \mathbf{y}^T \mathbf{y} - \hat{\mathbf{x}}^T \hat{\mathbf{x}} \quad (6.18)$$

o que não é mais do que a forma vectorial do teorema de Pitágoras,

$$\mathbf{y}^T \mathbf{y} = \hat{\mathbf{x}}^T \hat{\mathbf{x}} + \hat{\mathbf{n}}^T \hat{\mathbf{n}} \quad (6.19)$$

A importante conclusão é de que *não existe outro valor de $\hat{\mathbf{x}}$ pertencente ao sub-espaço do sinal que permita obter um valor menor de $\hat{\mathbf{n}}^T \hat{\mathbf{n}}$* . Por outras palavras o estimador (6.8) e o sinal estimado

respectivo (6.9) são aqueles que permitem o melhor ajuste às observações, i.e., o menor erro de ajuste.

6.3 Decomposição em valores singulares e problemas de MQ

Acabámos de determinar o estimador dos MQ de $\boldsymbol{\theta}$ dado por (6.8), o sinal respectivo dado por (6.9) e o ruído de estimação determinado por (6.11). Todas estas formas, como já vimos, dependem explicitamente da estrutura da matriz de Gram $\mathbf{G} = \mathbf{H}^T \mathbf{H}$ e portanto na de \mathbf{H} ou seja a forma do modelo do sinal ele mesmo. Em geral porém, as colunas de \mathbf{H} são raramente ortogonais e portanto \mathbf{G} pode ter valores próprios muito pequenos o que a torna em muitos casos quase singular. O mesmo pode ser dito de \mathbf{A} . O problema é saber se poderemos determinar outras representações para \mathbf{H} e \mathbf{A} que permitiriam gerar representações ortogonais para $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{x}}$ e $\hat{\mathbf{n}}$. Vamos começar por formar a matriz \mathbf{S} tal que

$$\mathbf{S} = [\mathbf{H}|\mathbf{A}] \quad (6.20)$$

que é uma matriz de dimensão $N \times N$ formada pelas p colunas de \mathbf{H} e as $N - p$ colunas de \mathbf{A} . A matriz de Gram de \mathbf{S} escreve-se

$$\mathbf{G} = \mathbf{S}^T \mathbf{S} = \begin{bmatrix} \mathbf{H}^T \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^T \mathbf{A} \end{bmatrix} \quad (6.21)$$

Na procura de uma dupla representação ortogonal para (6.21) e para $\mathbf{I} = \mathbf{S}(\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T = \mathbf{P}_H + \mathbf{P}_A$ encontramos a decomposição em valores singulares já abordada em 5.5. Podemos então escrever \mathbf{S} sob a seguinte forma

$$\mathbf{S} = [\mathbf{H}|\mathbf{A}] = [\mathbf{U}_H \boldsymbol{\Sigma}_H \mathbf{V}_H^T | \mathbf{U}_A \boldsymbol{\Sigma}_A \mathbf{V}_A^T] \quad (6.22)$$

onde \mathbf{U}_H , \mathbf{U}_A , \mathbf{V}_H e \mathbf{V}_A são matrizes ortogonais e $\boldsymbol{\Sigma}_H$ e $\boldsymbol{\Sigma}_A$ são matrizes diagonais definidas por

$$\mathbf{U}_H = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p] \quad (6.23a)$$

$$\mathbf{U}_A = [\mathbf{u}_{p+1}, \mathbf{u}_{p+2}, \dots, \mathbf{u}_N] \quad (6.23b)$$

$$\mathbf{V}_H = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p] \quad (6.23c)$$

$$\mathbf{V}_A = [\mathbf{v}_{p+1}, \mathbf{v}_{p+2}, \dots, \mathbf{v}_N] \quad (6.23d)$$

$$\boldsymbol{\Sigma}_H = \text{diag}[\sigma_1, \dots, \sigma_p] \quad (6.23e)$$

$$\boldsymbol{\Sigma}_A = \text{diag}[\sigma_{p+1}, \dots, \sigma_N] \quad (6.23f)$$

Usando a decomposição em valores singulares podemos então escrever as matrizes de Gram $\mathbf{H}^T \mathbf{H}$, $\mathbf{A}^T \mathbf{A}$ como

$$\mathbf{H}^T \mathbf{H} = \mathbf{V}_H \boldsymbol{\Sigma}_H^2 \mathbf{V}_H^T = \sum_{i=1}^p \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T \quad (6.24a)$$

$$\mathbf{A}^T \mathbf{A} = \mathbf{V}_A \boldsymbol{\Sigma}_A^2 \mathbf{V}_A^T = \sum_{i=p+1}^N \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T \quad (6.24b)$$

Da mesma forma, a decomposição da matriz unidade segundo as matrizes de projecção é

$$\mathbf{I} = \mathbf{P}_H + \mathbf{P}_A = \mathbf{U}_H \mathbf{U}_H^T + \mathbf{U}_A \mathbf{U}_A^T \quad (6.25)$$

que se pode também escrever

$$\mathbf{P}_H = \sum_{i=1}^p \mathbf{u}_i \mathbf{u}_i^T \quad \mathbf{P}_A = \sum_{i=p+1}^N \mathbf{u}_i \mathbf{u}_i^T \quad (6.26)$$

Como já vimos no caso geral, a decomposição em valores próprios dada para \mathbf{H} e \mathbf{A} por (6.22) podem simultaneamente diagonalizar não só \mathbf{S} mas também \mathbf{I} sob a forma das suas matrizes de projecção ortogonais. Assim podemos escrever

$$\mathbf{U}_H^T \mathbf{H} \mathbf{V}_H = \boldsymbol{\Sigma}_H \quad (6.27a)$$

$$\mathbf{U}_A^T \mathbf{A} \mathbf{V}_A = \boldsymbol{\Sigma}_A \quad (6.27b)$$

$$\mathbf{V}_H^T \mathbf{G}_H \mathbf{V}_H = \boldsymbol{\Sigma}_H^2 \quad (6.27c)$$

$$\mathbf{V}_A^T \mathbf{G}_A \mathbf{V}_A = \boldsymbol{\Sigma}_A^2 \quad (6.27d)$$

$$\mathbf{U}^T \mathbf{P}_H \mathbf{U} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (6.27e)$$

$$\mathbf{U}^T \mathbf{P}_A \mathbf{U} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad (6.27f)$$

onde $\mathbf{U} = [\mathbf{U}_H | \mathbf{U}_A]$. A decomposição em valores singulares permite abordar o problema da singularidade de \mathbf{H} mantendo representações ortogonais para \mathbf{P}_H e \mathbf{P}_A .

6.4 Algoritmos resolventes dos MQ

Este capítulo interessa-se pelos algoritmos resolventes dos seguinte sistemas de equações:

$$\mathbf{H}^T \mathbf{H} \hat{\boldsymbol{\theta}} = \mathbf{H}^T \mathbf{y} \quad (6.28a)$$

e

$$\det(\mathbf{H}^T \mathbf{H}) \neq 0. \quad (6.28b)$$

6.4.1 Algoritmo de factorização de Cholesky

O algoritmo de factorização de Cholesky permite factorizar uma matriz simétrica sob a forma do produto de duas matrizes triangulares inferiores. Assim a matriz de Gram $\mathbf{G} = \mathbf{H}^T \mathbf{H}$ pode-se escrever

$$\mathbf{H}^T \mathbf{H} = \mathbf{L}_H \mathbf{L}_H^T \quad (6.29)$$

com

$$\mathbf{L}_H = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ l_{p1} & l_{p2} & \dots & l_{pp} \end{bmatrix} = [\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_p] \quad (6.30)$$

Com a factorização (6.29) podemos escrever (6.28a)

$$\mathbf{L}_H \mathbf{L}_H^T \hat{\boldsymbol{\theta}} = \mathbf{H}^T \mathbf{y} \quad (6.31)$$

que pode ser resolvido em dois passos sucessivos:

$$(1.) \quad \mathbf{L}_H \mathbf{w} = \mathbf{H}^T \mathbf{y}$$

$$(2.) \quad \mathbf{L}_H^T \hat{\boldsymbol{\theta}} = \mathbf{w}$$

No primeiro passo resolve-se a equação (1.) em relação a \mathbf{w} começando pela primeira linha e depois as linhas sucessivas por substituição. Uma vez que temos \mathbf{w} podemos resolver o passo (2.) mas desta vez começando pela última linha e substituindo no sentido ascendente. O algoritmo de factorização de Cholesky pode ser encontrado em qualquer biblioteca de cálculo numérico.

6.4.2 Algoritmo QR

Aqui utiliza-se o facto de poder escrever a matriz \mathbf{H} sob a forma de uma matrix ortogonal \mathbf{U} e uma matrix triangular inferior \mathbf{L} ,

$$\mathbf{H} = \mathbf{U}_H \mathbf{L}_H^T \quad (6.32)$$

o que consiste simplesmente no facto de escrever a coluna n de \mathbf{H} como uma combinação linear de vectores ortogonais

$$\mathbf{h}_n = \sum_{i=1}^n l_{ni} \mathbf{u}_i \quad (6.33)$$

Com (6.32) podemos escrever (6.28) sob a forma simplificada

$$\mathbf{L}_H \mathbf{L}_H^T \hat{\boldsymbol{\theta}} = \mathbf{L}_H \mathbf{U}_H^T \mathbf{y} \quad (6.34)$$

implicando

$$\mathbf{L}_H^T \hat{\boldsymbol{\theta}} = \mathbf{U}_H^T \mathbf{y} \quad (6.35)$$

e que se resolve muito simplesmente dada a estrutura triangular de \mathbf{L}_H começando pela última linha para o vector de parâmetros $\hat{\boldsymbol{\theta}}$ e processando por substituição no sentido ascendente.

6.4.3 Decomposição em valores singulares

Como já vimos em bastante detalhe anteriormente a decomposição em valores singulares da matriz \mathbf{H} sob a forma

$$\mathbf{H} = \mathbf{U}_H \boldsymbol{\Sigma}_H \mathbf{V}_H^T \quad (6.36)$$

leva a uma simplificação da equação (6.28)

$$\hat{\boldsymbol{\theta}} = \mathbf{V}_H \boldsymbol{\Sigma}_H^{-1} \mathbf{U}_H^T \mathbf{y} = \sum_{i=1} \mathbf{v}_i \sigma_i^{-1} \mathbf{u}_i^T \mathbf{y} \quad (6.37)$$

Esta solução mostra claramente a interpretação geométrica feita em torno á figura 6.1., em particular, que o vector de parâmetros é obtido primeiramente por uma correlação entre as observações \mathbf{y} e os vectors \mathbf{u}_i que geram o espaço do sinal $\langle \mathbf{H} \rangle$ em seguida estas correlações são pesadas pelo inverso dos valores singulares, σ_i^{-1} , e finalmente as correlações já pesadas são usadas para construir o vector de parâmetros a partir dos vectores que geram os espaço de parâmetros, i.e., os \mathbf{v}_i .

6.5 Desempenho dos MQ

O desempenho dos MQ pode ser analisado considerando de novo o modelo linear estatístico com ruídos normais $\mathbf{n} : N[\mathbf{0}, \sigma^2 \mathbf{I}]$. Já provamos que se as observações são $\mathbf{y} : N[\mathbf{H}\boldsymbol{\theta}, \sigma^2 \mathbf{I}]$ então segundo (6.10) o sinal estimado é

$$\hat{\mathbf{x}} : N[\mathbf{H}\boldsymbol{\theta}, \sigma^2 \mathbf{P}_H] \quad (6.38)$$

e portanto trata-se de um estimador não enviesado

$$E[\hat{\mathbf{x}}] = \mathbf{H}\boldsymbol{\theta} = \mathbf{x}$$

e de matriz de covariância de erro

$$E[(\hat{\mathbf{x}} - \mathbf{H}\boldsymbol{\theta})(\hat{\mathbf{x}} - \mathbf{H}\boldsymbol{\theta})^T] = \sigma^2 \mathbf{P}_H = \sigma^2 \mathbf{U}_H \mathbf{U}_H^T$$

A variância do estimador é neste caso

$$\begin{aligned} \xi^2 &= E[(\hat{\mathbf{x}} - \mathbf{H}\boldsymbol{\theta})^T (\hat{\mathbf{x}} - \mathbf{H}\boldsymbol{\theta})] \\ &= \text{tr}\{E[(\hat{\mathbf{x}} - \mathbf{H}\boldsymbol{\theta})(\hat{\mathbf{x}} - \mathbf{H}\boldsymbol{\theta})^T]\} \\ &= \sigma^2 \text{tr}(\mathbf{P}_H) \\ &= \sigma^2 \text{tr}(\mathbf{U}_H \mathbf{U}_H^T) \\ &= \sigma^2 \text{tr}(\mathbf{U}_H^T \mathbf{U}_H) = p\sigma^2 \end{aligned}$$

onde p é a dimensão do sub-espaço do sinal.

O estimador do vector de parâmetros $\hat{\boldsymbol{\theta}}$ também se encontra distribuído segundo uma lei normal $N[\boldsymbol{\theta}, \sigma^2 (\mathbf{H}^T \mathbf{H})^{-1}]$. Trata-se portanto de um estimador sem viés e cuja matriz de covariância de erro se escreve

$$\begin{aligned} E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] &= \sigma^2 (\mathbf{H}^T \mathbf{H})^{-1} \\ &= \sigma^2 \mathbf{V}_H \boldsymbol{\Sigma}_H^{-2} \mathbf{V}_H^T. \end{aligned}$$

Note que esta última matriz é o inverso da matriz de Gram e que depende do inverso dos valores próprios desta matriz: *valores próprios próximos de zero dão grandes erros de estimação.*

O desempenho não se mede unicamente através das qualidades estatísticas dos estimadores do vector de parâmetros e do sinal. Na prática o nosso modelo é muitas vezes desconhecido, ora a necessidade de conhecer \mathbf{H} faz-nos fazer hipóteses de trabalho conducentes a uma determinada estrutura do sinal que pode ser mais ou menos verificada na prática. Tendo em conta que o desempenho do nosso método dos MQ está tão dependente das possibilidades de ajuste do modelo, seria muito interessante poder dispor de uma forma de julgar a qualidade do ajuste do modelo hipotetizado às observações. Esta qualidade do ajuste pode ser obtida através do ruído de estimação (que é por isso muitas vezes chamado ruído de modelo) e que é

$$\hat{\mathbf{n}} = \mathbf{y} - \hat{\mathbf{x}} : N[\mathbf{0}, \sigma^2 \mathbf{P}_A] \quad (6.39)$$

A partir da nossa discussão sobre a distribuição de formas quadráticas de vectores normais multivariados (capítulo 4.7) podemos dizer que

$$\hat{\mathbf{n}}^T \hat{\mathbf{n}} : \sigma^2 \chi_{N-p}^2 \quad (6.40)$$

de onde podemos deduzir que quando ajustamos o modelo estimado $\hat{\mathbf{x}}$ às observações \mathbf{y} , o vector de erro, $\hat{\mathbf{n}}$ é normal, e o seu módulo quadrado segue uma lei do $\sigma^2 \chi_{N-p}^2$ com $N-p$ graus de liberdade. O teste dos erros com uma lei de distribuição do $\sigma^2 \chi_{N-p}^2$ dá-nos uma ideia muito precisa sobre a qualidade do ajuste do nosso modelo.

Outro meio de julgar o desempenho do nosso método de estimação é de calcular o ganho em relação sinal/ruído. No modelo inicial temos as nossas observações $\mathbf{y} : N[\mathbf{x}, \sigma^2 \mathbf{I}]$. Vamos definir o rácio sinal/ruído como sendo o rácio entre o quadrado da média e a variância, assim para o sinal de entrada SNR_{in}

$$\text{SNR}_{\text{in}} = \frac{\mathbf{x}^T \mathbf{x}}{\text{tr}[\sigma^2 \mathbf{I}]} = \frac{\mathbf{x}^T \mathbf{x}}{N \sigma^2} \quad (6.41)$$

No modelo de saída temos o sinal estimado $\hat{\mathbf{x}} : N[\mathbf{x}, \sigma^2 \mathbf{P}_H]$ e portanto podemos definir de modo análogo o rácio de sinal/ruído de saída SNR_{out} como

$$\text{SNR}_{\text{out}} = \frac{\mathbf{x}^T \mathbf{x}}{\text{tr}[\sigma^2 \mathbf{P}_H]} = \frac{\mathbf{x}^T \mathbf{x}}{p \sigma^2} \quad (6.42)$$

e portanto o rácio entre de (6.42) sobre (6.41) dá-nos o ganho em sinal/ruído entre entrada e saída que é

$$G = \frac{N}{p} \quad (6.43)$$

Trata-se aqui de um resultado fundamental que nos diz que o ganho em SNR dos MQ é tanto maior quanto maior é a relação entre a dimensão das observações em relação á dimensão do modelo. Maximizar este quociente é a tarefa do modelizador.

6.6 Determinação e redução da ordem do modelo linear

Acabámos de ver que o ruído de estimação $\hat{\mathbf{n}}$ é $N[\mathbf{0}, \sigma^2 \mathbf{P}_A]$ e o erro quadrado médio é $\sigma^2(N-p)$. A ideia por detrás da redução da ordem é a de obter um estimador de ordem reduzida, $\hat{\mathbf{x}}_r$, em vez do

estimador de ordem completa $\hat{\mathbf{x}}$ tal que, o erro quadrado médio do ruído de estimação, seja inferior ao de $\hat{\mathbf{x}}$. Consideremos o seguinte estimador de ordem reduzida

$$\hat{\mathbf{x}}_r = \mathbf{P}_r \mathbf{y} \quad (6.44)$$

onde \mathbf{P}_r é uma versão de ordem reduzida de \mathbf{P}_H que se obtém através da eliminação de $p - r$ vectores ortogonais de \mathbf{U}_H . A selecção dos vectores a eliminar vai ser um dos problemas deste capítulo. Assim se

$$\mathbf{P}_H = \mathbf{U}_H \mathbf{U}_H^T; \quad \mathbf{U}_H = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p] \quad (6.45a)$$

então

$$\mathbf{P}_r = \mathbf{U}_r \mathbf{U}_r^T; \quad \mathbf{U}_r = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r] \quad (6.45b)$$

O estimador $\hat{\mathbf{x}}_r$ está distribuído segundo $N[\mathbf{x}_r, \sigma^2 \mathbf{P}_r]$, onde \mathbf{x}_r é a projecção de \mathbf{x} no sub-espaço gerado por \mathbf{U}_r , i.e.,

$$\mathbf{x}_r = \mathbf{P}_r \mathbf{x} \quad (6.46)$$

o erro de estimação é neste caso $\mathbf{x} - \hat{\mathbf{x}}_r$ e encontra-se distribuído segundo $N[\mathbf{b}_r, \sigma^2 \mathbf{P}_r]$, onde \mathbf{b}_r é o viés (bias) do estimador $\hat{\mathbf{x}}_r$ e é igual a*

$$\mathbf{b}_r = E[\mathbf{x} - \hat{\mathbf{x}}_r] = (\mathbf{P}_H - \mathbf{P}_r) \mathbf{x} \quad (6.47)$$

e o erro quadrado médio entre \mathbf{x} e $\hat{\mathbf{x}}_r$ escreve-se

$$\text{mse}(r) = E[(\mathbf{x} - \hat{\mathbf{x}}_r)^T (\mathbf{x} - \hat{\mathbf{x}}_r)] = \mathbf{b}_r^T \mathbf{b}_r + \sigma^2 \text{tr}[\mathbf{P}_r] = \mathbf{x}^T (\mathbf{P}_H - \mathbf{P}_r) \mathbf{x} + r \sigma^2 \quad (6.48)$$

aqui temos que o primeiro termo é o quadrado do viés do estimador e o segundo termo é a sua variância. Podemos ainda escrever

$$\text{mse}(r) = \sum_{i=r+1}^p |\mathbf{u}_{(i)}^T \mathbf{x}|^2 + r \sigma^2 \quad (6.49)$$

onde os $\mathbf{u}_{(i)}$ são vectores próprios da matriz de ordem completa \mathbf{U}_H (de ordenação a determinar) e que foram eliminados para formar a matriz de ordem reduzida \mathbf{U}_r . O estimador de ordem reduzida só será efectivo se $\text{mse}(r) \leq p \sigma^2$ onde $p \sigma^2$ é o erro quadrado médio do estimador de ordem completa assim, usando (6.49), podemos escrever a condição necessária para um ganho através da redução de ordem que é

$$\sum_{i=r+1}^p |\mathbf{u}_{(i)}^T \mathbf{x}|^2 \leq (p - r) \sigma^2 \quad (6.50)$$

Isto significa que a escolha óptima para o valor da ordem é

$$r^* = \arg \min_r [\mathbf{x}^T (\mathbf{P}_H - \mathbf{P}_r) \mathbf{x} + r \sigma^2] \quad (6.51)$$

* porque $\mathbf{x} = \mathbf{P}_H \mathbf{x}$.

o que demonstra que para obter um estimador não enviesado basta corrigir a estatística $\hat{\mathbf{b}}_r^T \hat{\mathbf{b}}_r$ com o termo $-\sigma^2(p-r)$ para obter um estimador não enviesado. Podemos agora propor um estimador do erro quadrado médio do erro entre \mathbf{x} e $\hat{\mathbf{x}}_r$ que pode ser

$$\widehat{\text{mse}}(r) = \hat{\mathbf{b}}_r^T \hat{\mathbf{b}}_r - \sigma^2(p-r) + r\sigma^2 = \hat{\mathbf{b}}_r^T \hat{\mathbf{b}}_r + (2r-p)\sigma^2 \quad (6.56)$$

e podemos, a partir de (6.55), dizer que (6.56) é um estimador não enviesado de $\text{mse}(r)$, e portanto inserir (6.56) em (6.51)

$$r^* = \arg \min_r [\widehat{\text{mse}}(r)] \quad (6.57)$$

onde

$$\begin{aligned} \widehat{\text{mse}}(r) &= \hat{\mathbf{b}}_r^T \hat{\mathbf{b}}_r + (2r-p)\sigma^2 \\ &= \sum_{r+1}^p |\mathbf{u}_{(i)}^T \mathbf{y}|^2 + (2r-p)\sigma^2 \end{aligned} \quad (6.58)$$

Esta última forma permite dizer que a ordenação dos vectores próprios de \mathbf{H} deve ser tal que

$$|\mathbf{u}_{(1)}^T \mathbf{y}|^2 \geq \dots \geq |\mathbf{u}_{(r)}^T \mathbf{y}|^2 \geq \dots \geq |\mathbf{u}_{(p)}^T \mathbf{y}|^2 \quad (6.59)$$

e os primeiros $\mathbf{u}_{(i)}$; $i = 1, \dots, r$ devem ser usados para construir a matriz de projecção de ordem r , \mathbf{P}_r . O valor óptimo r^* de r é o valor de r que minimiza $\widehat{\text{mse}}(r)$.

6.7 MQ recursivos

O método dos MQ recursivos responde à necessidade de obter uma estimativa dos MQ que evolui no tempo (ou no espaço) à medida que novas observações vão sendo adquiridas e também à necessidade que esta evolução se faça de modo eficiente. Consideremos o modelo linear

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n} \quad (6.60)$$

neste modelo vamos fixar a dimensão de $\boldsymbol{\theta}$ a p e vamos permitir que o número de observações N varie. Neste caso vamos chamar-lhe t e escrever

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{t-1} \\ y_t \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{t-1} \\ \mathbf{c}_t^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_{t-1} \\ n_t \end{bmatrix} \quad (6.61)$$

$$\mathbf{y}_t = \mathbf{H}_t \boldsymbol{\theta} + \mathbf{n}_t$$

que separadamente para $t-1$ e t pode ser

$$\mathbf{y}_{t-1} = \mathbf{H}_{t-1} \boldsymbol{\theta} + \mathbf{n}_{t-1} \quad (6.62)$$

$$y_t = \mathbf{c}_t^T \boldsymbol{\theta} + n_t \quad (6.63)$$

onde \mathbf{c}_t é o vector linha de ordem t da matriz \mathbf{H} . No instante t podemos calcular o estimador dos MQ

$$\mathbf{P}_t^{-1} \hat{\boldsymbol{\theta}}_t = \mathbf{H}_t^T \mathbf{y} \quad (6.64)$$

onde

$$\begin{aligned} \mathbf{P}_t^{-1} &= \mathbf{H}_t^T \mathbf{H}_t = (\mathbf{H}_{t-1}^T \mathbf{H}_{t-1} + \mathbf{c}_t \mathbf{c}_t^T) \\ &= \mathbf{P}_{t-1}^{-1} + \mathbf{c}_t \mathbf{c}_t^T \end{aligned} \quad (6.65a)$$

e

$$\mathbf{H}_t^T \mathbf{y}_t = \mathbf{H}_{t-1}^T \mathbf{y}_{t-1} + \mathbf{c}_t y_t \quad (6.65b)$$

Utilizando (6.65) em (6.64) e depois de alguma algebra (que deixamos como exercicio!) chega-se á equação recursiva

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} + \mathbf{k}_t [y_t - \mathbf{c}_t^T \hat{\boldsymbol{\theta}}_{t-1}] \quad (6.66)$$

com $\mathbf{k}_t = \gamma_t \mathbf{P}_{t-1}^{-1} \mathbf{c}_t$ e $\gamma_t^{-1} = 1 + \mathbf{c}_t^T \mathbf{P}_{t-1}^{-1} \mathbf{c}_t$. Da mesma maneira a matriz de Gram pode ser obtida recursivamente

$$\mathbf{P}_t^{-1} = \mathbf{P}_{t-1}^{-1} + \mathbf{c}_t \mathbf{c}_t^T \quad (6.67)$$

6.8 MQ ponderados

Os MQ ponderados (weighted least squares=WLS) são uma generalização dos MQ. Enquanto os MQ minimizam (6.6) os MQ ponderados minimizam

$$\min_{\boldsymbol{\theta}} [(\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{W} (\mathbf{y} - \mathbf{H}\boldsymbol{\theta})] \quad (6.68)$$

onde \mathbf{W} é uma matriz de ponderação não singular e simétrica. Demonstra-se que o estimador dos MQ ponderados correspondente se escreve

$$\hat{\boldsymbol{\theta}}_{\text{WLS}} = (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W} \mathbf{y} \quad (6.69)$$

neste caso a matriz $\mathbf{H}^T \mathbf{W} \mathbf{H}$ é invertível se \mathbf{H} for de ordem completa p , e \mathbf{W} for também de ordem completa N . No caso do modelo linear $\mathbf{y} : N[\mathbf{H}\boldsymbol{\theta}, \mathbf{R}]$ e que $\mathbf{W} = \mathbf{R}^{-1}$, onde \mathbf{R} é a matriz de covariância das observações, por substituição em (6.69) temos que

$$\hat{\boldsymbol{\theta}}_{\text{WLS}} : N[\boldsymbol{\theta}, (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1}] \quad (6.70)$$

Prova-se que este estimador é neste caso idêntico ao estimador do máximo de verossimilhança no mesmo modelo.

6.9 MQ condicionados e condicionamento linear

Os MQ condicionados são outra variante dos MQ que respondem ao problema de estimar um vector de parâmetros cuja escolha não é livre. Neste caso é necessário exprimir esta falta de liberdade por uma equação de condicionamento como por exemplo

$$\mathbf{C}^T \boldsymbol{\theta} = \mathbf{c} \quad (6.71)$$

onde \mathbf{C} e \mathbf{c} são uma matriz e um vector conhecidos que impõem r condicionamentos lineares em $\boldsymbol{\theta}$. Trata-se então da minimização do problema clássico (6.6) sujeito a (6.71). Para resolver este problema forma-se o Lagrangiano

$$J = \frac{1}{2}(\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{H}\boldsymbol{\theta}) - (\mathbf{C}^T\boldsymbol{\theta} - \mathbf{c})^T\boldsymbol{\lambda} \quad (6.72)$$

Expressando o gradiente de J em relação ao vector de parâmetros e equacionando a zero obtém-se

$$\hat{\boldsymbol{\theta}}_{\text{CLS}} = \hat{\boldsymbol{\theta}}_{\text{LS}} + (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{C}\boldsymbol{\lambda} \quad (6.73)$$

onde $\hat{\boldsymbol{\theta}}_{\text{CLS}}$ é a solução dos MQ condicionada e $\hat{\boldsymbol{\theta}}_{\text{LS}}$ é a solução não condicionada dada por (6.8). Para determinar o vector de coeficientes de Lagrange $\boldsymbol{\lambda}$ deve-se substituir (6.73) em (6.71) e resolver para $\boldsymbol{\lambda}$ obtendo

$$\boldsymbol{\lambda} = [\mathbf{C}^T(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{C}]^{-1}(\mathbf{c} - \mathbf{C}^T\boldsymbol{\theta}_{\text{LS}}) \quad (6.74)$$

A solução completa para $\boldsymbol{\theta}_{\text{CLS}}$ obtém-se substituindo (6.74) em (6.73). No caso em que $\mathbf{c} = \mathbf{0}$ diz-se que as condicionantes são homogéneas e neste caso (6.74) toma uma forma relativamente simples o mesmo acontecendo para o estimador respectivo, i.e., no caso homogéneo

$$\hat{\boldsymbol{\theta}}_{\text{CLS}} = \mathbf{P}\hat{\boldsymbol{\theta}}_{\text{LS}}$$

$$\mathbf{P} = \mathbf{I} - (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{C}[\mathbf{C}^T(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{C}]^{-1}\mathbf{C}^T \quad (6.75)$$

Existe um caso particular dos MQ condicionados muito frequente na prática que é quando se pretende determinar o vector de parâmetros $\boldsymbol{\theta}$ tal que

$$\min_{\boldsymbol{\theta}}[\boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}] \quad \text{condicionado por} \quad \mathbf{C}^T\boldsymbol{\theta} = \mathbf{c} \quad (6.76)$$

Este caso é conhecido como sendo o caso de minimização quadrática sob condicionantes lineares e obtém-se formalmente do caso geral dos MQ condicionados fazendo $\mathbf{y} = \mathbf{0}$. Escrevendo $\mathbf{R} = \mathbf{H}^T\mathbf{H}$ a solução é dada também pelo método dos Lagrangianos

$$J = \boldsymbol{\theta}^T\mathbf{R}\boldsymbol{\theta} - (\mathbf{C}^T\boldsymbol{\theta} - \mathbf{c})^T\boldsymbol{\lambda} \quad (6.77)$$

onde, utilizando o mesmo método que anteriormente, se obtém a solução

$$\hat{\boldsymbol{\theta}}_0 = \mathbf{R}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C})^{-1}\mathbf{c} \quad (6.78)$$

A interpretação geométrica deste resultado permite-nos dizer que $\boldsymbol{\theta}_0$ é a soma de uma componente projectada no sub-espaco gerado por \mathbf{C} e outra no sub-espaco ortogonal: $\hat{\boldsymbol{\theta}}_0 = \mathbf{P}_C\hat{\boldsymbol{\theta}}_0 + (\mathbf{I} - \mathbf{P}_C)\hat{\boldsymbol{\theta}}_0$, sabendo que

$$\mathbf{P}_C = \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T$$

é uma matriz de projecção formando $\boldsymbol{\theta}_c = \mathbf{P}_C\boldsymbol{\theta}_0$ que não é mais do que a solução de norma mínima ilustrada na figura 6.3.

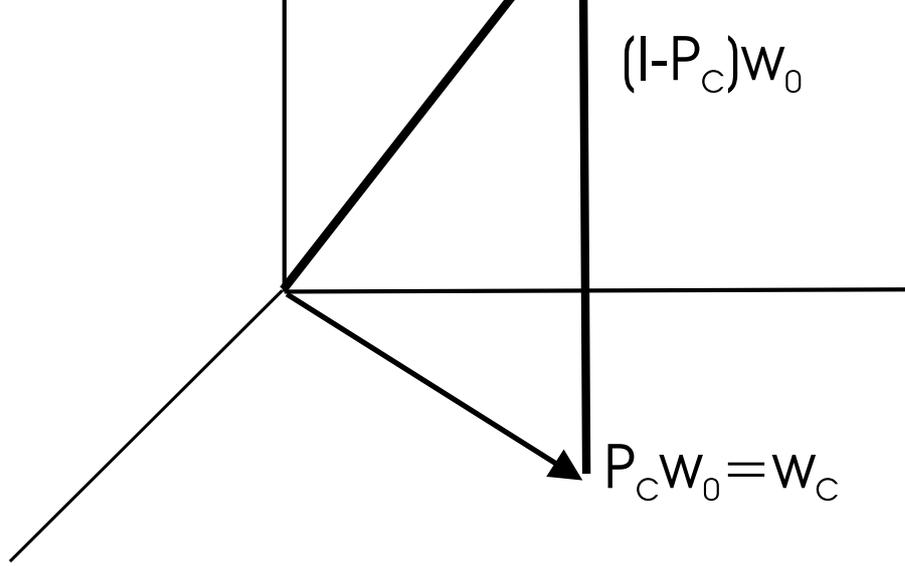


Figura 6.3.: decomposição de θ_0 .

6.10 MQ totais

Este caso é mais uma variante do caso geral que, apesar de não acontecer frequentemente na prática, é bastante realístico (e difícil !...). Como vimos nos capítulos anteriores sobre MQ a solução do problema para o sinal \mathbf{x} no caso geral é dada pela projecção das observações \mathbf{y} no sub-espaço do sinal, i.e., $\hat{\mathbf{x}} = \mathbf{P}_H \mathbf{y}$. Isto é bastante restrictivo porque força $\hat{\mathbf{x}}$ a pertencer a $\langle \mathbf{H} \rangle$ o que implica que deveremos ter uma grande confiança no nosso modelo e conhecê-lo perfeitamente. Vamos supor agora que existem erros na matriz do modelo \mathbf{H} . Será possível afastar-se do modelo \mathbf{H} ? E como devemos fazê-lo? A resposta a estas perguntas encontra-se na teoria dos Mínimos Quadrados Totais (MQT, ou Total Least Squares=TLS).

Na teoria do MQT o modelo á posteriori é dado por

$$\begin{aligned} \mathbf{y} &= \mathbf{P}_s \mathbf{y} + (\mathbf{I} - \mathbf{P}_s) \mathbf{y} \\ &= \hat{\mathbf{x}}_{\text{TLS}} + \hat{\mathbf{n}}_{\text{TLS}} \end{aligned} \quad (6.79)$$

onde o projector \mathbf{P}_s é escolhido de forma a minimizar a soma dos quadrados dos elementos de $\hat{\mathbf{n}}_{\text{TLS}}$ mais a soma dos quadrados dos elementos $\Delta_H = (\mathbf{I} - \mathbf{P}_s) \mathbf{H}$, daí o termo de MQ totais. Neste modelo $\mathbf{P}_s \mathbf{H}$ é o modelo corrigido onde deveremos encontrar o sinal estimado $\hat{\mathbf{x}}_{\text{TLS}}$. Não vamos entrar em muito mais detalhe neste método mas a diferença essencial com o caso dos MQ geral situa-se no facto de que o projector utilizado, \mathbf{P}_s , não se encontra ajustado ao sub-espaço $\langle \mathbf{H} \rangle$ mas sim ao sub-espaço $\langle \hat{\mathbf{H}} \rangle$, que é uma espécie de compromisso entre o sub-espaço do sinal definido por \mathbf{H} e as observações \mathbf{y} .

Este método minimiza

$$\xi^2 = \text{tr} \begin{bmatrix} \mathbf{H}^T \\ \mathbf{y}^T \end{bmatrix} [\mathbf{I} - \mathbf{P}_s] [\mathbf{H} | \mathbf{y}] \quad (6.80)$$

que se obtém calculando a decomposição em valores singulares de $[\mathbf{H} | \mathbf{y}]$ e construindo \mathbf{P}_s com todos os vectores singulares menos aquele correspondendo ao mais pequeno valor singular, i.e., se

$$[\mathbf{H} | \mathbf{y}] = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \quad (6.81)$$

então $\mathbf{U} = [\mathbf{U}_s | \mathbf{u}]$ e o projector \mathbf{P}_s escreve-se

$$\mathbf{P}_s = \mathbf{U}_s \mathbf{U}_s^T \quad (6.82)$$

daí podemos escrever a expressão do estimador

$$\boldsymbol{\theta}_{\text{TLS}} = (\mathbf{H}^T \mathbf{P}_s \mathbf{H})^{-1} \mathbf{H}^T \mathbf{P}_s \mathbf{y} \quad (6.83)$$

que é idêntica à solução do problema dos MQ ponderados com $\mathbf{W} = \mathbf{P}_s$.

6.11 Problemas inversos e MQ sub-determinados

O caso sub-determinado acontece quando $N < p$ e temos mais parâmetros do que observações. Neste caso costuma-se dizer que se “inverte as observações” e o que torna o problema particularmente difícil (e interessante) é que não existe uma solução única. A escolha de uma solução de entre a infinidade de soluções possíveis necessita a definição de condições adicionais no vector de parâmetros $\boldsymbol{\theta}$.

Comecemos por identificar o tipo de soluções do sistema de equações puramente determinístico

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} \quad (6.84)$$

com a dimensão de $\boldsymbol{\theta}$ igual a $p > N$ dimensão de \mathbf{y} . A decomposição em valores singulares de \mathbf{H} permite caracterizar a solução qualquer que seja a matriz \mathbf{H} . Assim se $\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ a solução da equação (6.84) é dada pelo pseudoinverso $\mathbf{H}^\#$ de \mathbf{H} tal que

$$\begin{aligned} \boldsymbol{\theta}_0 &= \mathbf{H}^\# \mathbf{y} = \mathbf{V} \boldsymbol{\Sigma}^{-1} \mathbf{U}^T \mathbf{y} \\ &= \sum_{i=1}^N \mathbf{v}_i \frac{1}{\sigma_i} \mathbf{u}_i^T \mathbf{y} \end{aligned} \quad (6.85)$$

Esta solução é também a solução com norma mínima. Esta solução pertence obviamente ao sub-espaco gerado pelos vectores próprios de \mathbf{V} , i.e., $\langle \mathbf{V} \rangle$. Podemos além disso abranger toda a dimensão do problema notando que qualquer solução do tipo $\boldsymbol{\theta} + \boldsymbol{\theta}_1$ tal que $\boldsymbol{\theta}_1$ pertença ao sub-espaco $\langle \mathbf{V}^\perp \rangle$ ortogonal a $\langle \mathbf{V} \rangle$ é também solução do problema visto que

$$\mathbf{H}(\boldsymbol{\theta} + \boldsymbol{\theta}_1) = \mathbf{y} \quad (6.86)$$

Esta questão encontra-se ilustrada na figura 6.4.

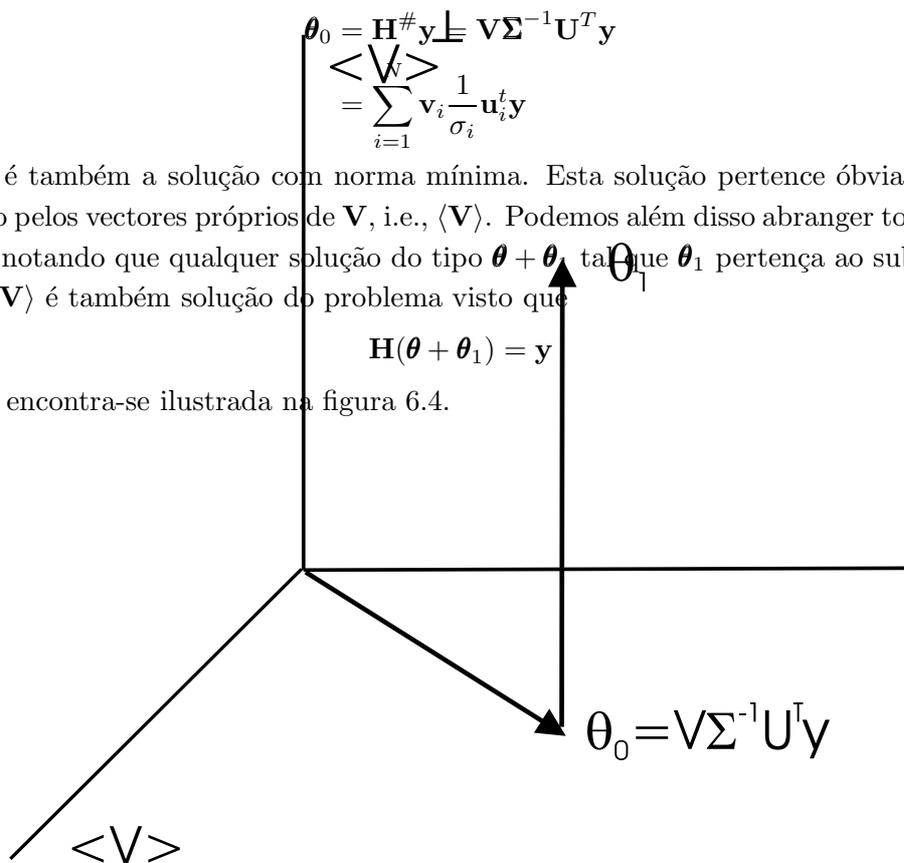


Figura 6.4.: soluções de $\mathbf{y} = \mathbf{H}\boldsymbol{\theta}$.

Vários critérios podem ser utilizados para escolher uma solução entre as soluções possíveis. Em geral estes critérios são específicos ao problema considerado e não poderão ser todos descritos aqui. Optaremos por descrever apenas alguns dos métodos mais usuais.

6.11.1 Solução de norma mínima

Esta solução encontra-se através da minimização de $(1/2)\boldsymbol{\theta}^T\boldsymbol{\theta}$ sob a condicionante $\mathbf{y} = \mathbf{H}\boldsymbol{\theta}$. A solução encontra-se pelo método dos Lagrangianos e é equivalente á solução do caso geral dado por (6.85) e é portanto nada mais nada menos do que a solução do pseudoinverso.

6.11.2 Solução de ordem reduzida

Inspecionando (6.85) pode-se notar que uma pequena mudança no vector de observações pode originar uma grande alteração na solução sobretudo devido à ponderação introduzida pelo inverso dos valores singulares $1/\sigma_i$ quando estes tomam valores próximos de zero. Trata-se aqui de excluir os valores singulares de menor valor definindo uma solução

$$\hat{\boldsymbol{\theta}}_r = \mathbf{V}\boldsymbol{\Sigma}_r^{-1}\mathbf{U}^T\mathbf{y} \quad (6.87)$$

onde $\boldsymbol{\Sigma}_r^{-1}$ é uma aproximação de ordem reduzida $r < N$ de $\boldsymbol{\Sigma}^{-1}$ onde os $N - r$ menores valores singulares foram substituidos por zero.

6.11.3 Solução do máximo de entropia

O princípio de máxima entropia fornece uma forma de condicionar $\boldsymbol{\theta}_i$ a ser não negativo maximizando uma quantidade que é de grande utilidade em muitos ramos de ciência. A idea é

$$\max_{\boldsymbol{\theta}} \left(- \sum_{i=1}^p \theta_i \log \theta_i \right) \quad \text{condicionado a} \quad \mathbf{y} = \mathbf{H}\boldsymbol{\theta} \quad (6.88)$$

A solução deste problema encontra-se como sempre através do método dos Lagrangianos formando

$$L = - \sum_{i=1}^p \frac{\theta_i}{M} \log \frac{\theta_i}{M} - (\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T \boldsymbol{\mu} \quad (6.89)$$

onde $\sum_{i=1}^p \theta_i = M$. Equacionando a derivada a zero obtem-se

$$\theta_i = \exp \left\{ -1 + \sum_{n=1}^N \lambda_n c_{ni} \right\} \quad (6.90)$$

onde a solução é parametrizada por $N - p$ parâmetros $\lambda_n; n = 1, \dots, N$. Este parâmetros são determinados, como de costume, através do sistema de equações de condicionamento $\mathbf{y} = \mathbf{H}\boldsymbol{\theta}$.

6.12 Problemas de identificação

Até agora o problema considerado tem sido o de determinar o vector de parâmetros $\boldsymbol{\theta}$ num modelo linear conhecido, $\mathbf{y} = \mathbf{H}\boldsymbol{\theta}$. Em todos os casos \mathbf{H} era conhecido. Em muitos casos práticos porém, \mathbf{H} é completamente desconhecido ou só parcialmente conhecido o que leva á necessidade de simultâneamente estimar $\boldsymbol{\theta}$ e \mathbf{H} . Este problema é chamado problema de identificação do modelo. Começemos por substituir $\boldsymbol{\theta}$ pela sua estimada dos MQ, $\hat{\boldsymbol{\theta}}$, tal que

$$\begin{aligned}\mathbf{y} &= \mathbf{H}\hat{\boldsymbol{\theta}} + \hat{\mathbf{n}} \\ &= \mathbf{P}_H\mathbf{y} + \mathbf{P}_A\mathbf{y}\end{aligned}\quad (6.91)$$

onde erro quadrático do ruído de ajustamento é

$$e^2 = \hat{\mathbf{n}}^T\hat{\mathbf{n}} = \mathbf{y}^T\mathbf{P}_A\mathbf{y} = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_H)\mathbf{y}\quad (6.92)$$

Se a matriz \mathbf{H} for desconhecida esta deverá ser estimada minimizando também e^2 ,

$$\hat{\mathbf{H}} = \arg \max_{\mathbf{H}}(\mathbf{y}^T\mathbf{P}_H\mathbf{y})\quad (6.93)$$

o que na realidade é um problema de maximização da projecção de $\hat{\mathbf{n}}$ em $\langle\mathbf{H}\rangle$. De modo equivalente podemos estimar \mathbf{A} tal que

$$\hat{\mathbf{A}} = \arg \min_{\mathbf{A}}(\mathbf{y}^T\mathbf{P}_A\mathbf{y})\quad (6.94)$$

onde $\mathbf{P}_A = \mathbf{I} - \mathbf{P}_H$. Podemos ver fácilmente que este problema está mal definido (ill-posed) porque qualquer matriz $\hat{\mathbf{H}} = [\mathbf{y}, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_p]$ onde os vectores \mathbf{h}_i são arbitrários é solução do problema pois com este $\hat{\mathbf{H}}$, temos que $\mathbf{P}_{\hat{\mathbf{H}}}\mathbf{y} = \mathbf{y}$ e o erro e^2 é zero e portanto uma solução óptima do problema. Na realidade o que aconteceu foi que nós demos demasiada liberdade na escolha de \mathbf{H} e para responder a casos práticos concretos em geral, \mathbf{H} tem que ser determinado tal que verifique (6.93), mas condicionado a pertencer a uma família de matrizes definida por uma forma paramétrica. Ou seja, não se trata de determinar uma matriz qualquer, mas sim uma matriz que tenha uma estrutura determinada bem definida, mas com parâmetros desconhecidos que esses, sim deverão ser identificados.

6.12.1 Identificação do modelo ARMA

Talvez este seja o caso mais conhecido em teoria do sinal: o modelo auto-regressivo de média móvel (auto-regressive moving average = ARMA). A função de transferência de tal sistema é dada por

$$H(z) = \frac{b_0 + b_1z^{-1} + \dots + b_{p-1}z^{-(p-1)}}{1 + a_1z^{-1} + \dots + a_pz^{-p}}\quad (6.95)$$

o que significa que a resposta impulsiva unitária se escreve

$$h_t = \begin{cases} 0, & t < 0 \\ -\sum_{n=1}^p a_n h_{t-n} + b_t, & 0 \leq t < p \\ -\sum_{n=1}^p a_n h_{t-n}, & t \geq p \end{cases}\quad (6.96)$$

Existem outros tipos de estruturas possíveis e normalmente utilizadas para definir a matriz \mathbf{H} do modelo linear. Algumas destas encontram-se em Scharf [1] pp.405 e seguintes.

6.13 Exemplos

6.13.1 Modelo de regressão polinomial

Considere um vector \mathbf{y} contendo as amostras de uma curva $y(t)$ nos pontos $t = T, 2T, \dots, NT$.

$$\mathbf{y} = [y_1, y_2, \dots, y_N]^T \quad (6.104)$$

com $y_k = y(kT)$. Gostaríamos de ajustar nesta curva um modelo polinomial do tipo

$$x(t) = \sum_{n=1}^p \theta_n t^{n-1} \quad (6.105)$$

onde $x_k = x(kT) = \sum_{n=1}^p \theta_n (kT)^{n-1}$. O problema dos MQ associado é o de determinar $\boldsymbol{\theta}$ tal que

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} \quad (6.106)$$

se ajuste ás observações \mathbf{y} em (6.104), com

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_p] \quad (6.107a)$$

$$\mathbf{h}_n = [T^{n-1}, (2T)^{n-1}, \dots, (NT)^{n-1}] \quad (6.107b)$$

6.13.2 Modelo de exponenciais complexas

Considere uma estrutura ressonante que suporte modos complexos da forma

$$h_n(t) = e^{j\omega_n t} \quad (6.108)$$

onde ω_n é a frequência angular do modo de ordem n . Aos instantes de amostragem $t = kT$ estes modos tomam a forma

$$h_{nk} = h_n(kT) = z_n^k \quad (6.109)$$

com $z_n = \exp(j\omega_n T)$. Um modelo para as observações \mathbf{y} pode ser formado a partir da combinação linear de p modos, assim

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \quad (6.110)$$

onde $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}$ com

$$\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_p] \quad (6.111a)$$

$$\mathbf{h}_n = [z_n^0, \dots, z_n^{N-1}]^T \quad (6.111b)$$

e mais uma vez o estimador dos MQ de $\boldsymbol{\theta}$ é o que consegue um melhor ajuste (no sentido dos MQ) entre \mathbf{x} e \mathbf{y} .

6.13.3 Representação de Fourier discreta

Um modelo, amplamente usado em teoria do sinal, e algo semelhante ao de 6.13.2., é o da representação de Fourier discreta. Neste caso admitimos que um sinal temporal se escreve sob a forma

$$x(n) = \sum_{m=0}^{N-1} \theta_m e^{j2\pi f_m n \Delta t} \quad (6.112)$$

onde os θ_m são os coeficientes de Fourier, $f_m = m\Delta f$ são as componentes de Fourier discretas equiespaçadas em $[-N/2, N/2 - 1]$ com $\Delta f = f_s/N$ e $\Delta t = 1/f_s$ é o intervalo de amostragem. (6.112) pode ser posto sob a forma familiar

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + \mathbf{n} \\ &= \mathbf{H}\boldsymbol{\theta} + \mathbf{n} \end{aligned} \quad (6.113)$$

com $\mathbf{y} = [y_0, y_1, \dots, y_{N-1}]^T$ e

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \exp(\lambda_0) & \exp(\lambda_1) & \dots & \exp(\lambda_{N-1}) \\ \vdots & \vdots & \dots & \vdots \\ \exp(\lambda_0[N-1]) & \exp(\lambda_1[N-1]) & \dots & \exp(\lambda_{N-1}[N-1]) \end{bmatrix} \quad (6.114a)$$

$$\boldsymbol{\theta} = [\theta_0, \theta_1, \dots, \theta_{N-1}]^T \quad (6.114b)$$

e $\lambda_i = j2\pi f_i \Delta t$. A solução dos MQ deste problema é portanto (conhecendo os f_i !)

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \mathbf{y} \quad (6.115)$$

e dada a estrutura particular de \mathbf{H} neste caso onde as suas colunas são mutuamente ortogonais temos que

$$(\mathbf{H}^H \mathbf{H})^{-1} = \frac{1}{N} \mathbf{I} \quad (6.116)$$

e por substituição em (6.115) temos que

$$\hat{\boldsymbol{\theta}} = \frac{1}{N} \mathbf{H}^H \mathbf{y} \quad (6.117)$$

que tendo em conta a forma de \mathbf{H} se escreve

$$\theta_m = \frac{1}{N} \sum_{n=0}^{N-1} y_n e^{-j2\pi f_m n \Delta t}, \quad \text{para } m = 0, \dots, N-1 \quad (6.118)$$

que não é mais do que a expressão da TFD se substituirmos $f_m = m/N$. Acabámos portanto de provar que a TFD não é mais do que a solução dos MQ entre um modelo constituído pela soma de N sinusoides complexas a frequências equiespaçadas e as N observações do sinal temporal $y(n), n = 0, \dots, N-1$.

Referências

- [1] L.L. Scharf, *Statistical Signal Processing*, Addison-Wesley, 1991.
- [2] G. Strang, *Linear Algebra and its Applications*, Academic, New York, 1976.

Mestrado em Engenharia de Sistemas e Computação

Exame de Processamento Estatístico do Sinal

Problema 1: Seja \mathbf{y} um vector aleatório real e de média nula com componentes $y_i; i = 1, \dots, N$ e x uma variável aleatória escalar também de média nula. Suponha que a matriz de covariância de \mathbf{y} , \mathbf{R}_y , pode ser factorizada sob a forma $\mathbf{R}_y = \mathbf{A}\mathbf{A}^T$ tal que $\mathbf{B} = \mathbf{A}^{-1}$.

- a) considerando a transformação linear $\mathbf{u} = \mathbf{B}\mathbf{y}$, demonstre que a matriz de covariância de \mathbf{u} é $\mathbf{R}_u = \mathbf{I}$. Qual a propriedade fundamental de \mathbf{u} ?
- b) introduzindo o vector $\mathbf{c} = E[x\mathbf{y}]$ e o vector \mathbf{r} cujas componentes r_i são a correlação entre x e u_i , calcule \mathbf{r} .
- c) admitindo um estimador linear não enviesado $\hat{x} = \mathbf{h}^T \mathbf{u} + a$ de x , demonstre que $a = 0$.
- d) utilizando a mesma forma linear do estimador \hat{x} e com o valor de a de c), demonstre que \mathbf{h} que minimiza $\epsilon = E[(x - \hat{x})^2]$ se escreve

$$\mathbf{h} = \mathbf{r}$$

- e) para \hat{x} calculado em c) e d) determinar o erro quadrático médio ϵ .
- f) voltando ao vector \mathbf{y} , calcule agora \hat{x} em função de \mathbf{y} , \mathbf{c} e \mathbf{B} e o erro de estimação correspondente.
- g) qual o interesse da factorização de \mathbf{R}_y neste caso ?

Problema 2:

Considere o sinal $x(t)$ formado por uma soma de p sinusoides complexas

$$x(t) = \sum_{n=1}^p A_n e^{(\beta_n + j\omega_n)t}$$

onde β_n é o coeficiente de amortecimento, A_n é a amplitude e ω_n é a frequência radial da senoide n .

- a) amostrando $x(t)$ aos instantes $t = kT; k = 0, \dots, N - 1$, demonstre que $x(t)$ se pode escrever sob a forma de um modelo linear $\mathbf{x} = \mathbf{H}\mathbf{a}$ onde

$$\mathbf{a}^t = [A_1, A_2, \dots, A_p]$$

$$\mathbf{x}^t = \{x(0), x(T), \dots, x[(N - 1)T]\}$$

Determine a matriz \mathbf{H} .

b) fazendo $\beta_n = 0 \forall n$ e $\omega_n = \frac{2\pi n}{NT}$, demonstre que \mathbf{H} é uma matriz ortogonal.

c) na mesmas condições que em b), e assumindo uma observação \mathbf{y} com ruído aditivo, branco, de média nula e de variância σ^2 , determine o estimador dos mínimos quadrados $\hat{\mathbf{a}}$ de \mathbf{a} em função de \mathbf{H} , N e \mathbf{y} .

d) calcule o viés b e erro quadrado médio de estimação e^2 do estimador $\hat{\mathbf{a}}$.

e) calcule o ganho em relação sinal/ruído usando $\hat{\mathbf{a}}$.

f) calcule $\hat{\mathbf{x}}$ e $\hat{\mathbf{n}}$ respectivamente sinal estimado e ruído de estimação em função de \mathbf{y} .

Mestrado em Engenharia de Sistemas e Computação

Exame de Processamento Estatístico do Sinal: época de recurso

Problema 1: Seja \mathbf{y}_n uma realização de um processo aleatório vectorial, real de dimensão L tal que

$$\mathbf{y}_n = \mathbf{z} + \boldsymbol{\epsilon}_n$$

onde $\boldsymbol{\epsilon}$ é um vector aleatório de média nula e matriz de covariância igual $\sigma^2\mathbf{I}$ e o vector \mathbf{z} é determinístico.

- calcular a matriz de correlação \mathbf{R}_y de \mathbf{y}_n .
- responder de novo á alinea a), desta vez com $\mathbf{z} = \sum_{i=1}^I \mathbf{h}_i x_i$ onde os $x_i; i = 1, \dots, I$ são um conjunto de variáveis aleatórias independentes e não correladas com $\boldsymbol{\epsilon}$.
- no caso em que $I < L$ qual a ordem da matriz \mathbf{R}_y se os vectores $\mathbf{h}_i; i = 1, \dots, I$ formarem um conjunto de vectores linearmente independentes ?
- qual o espectro dos valores próprios de \mathbf{R}_y ?

Problema 2:

Considere o sinal modulado $x(t)$ definido por

$$x(t) = \sum_{i=1}^{p/2} a_i \cos i\omega t + \sum_{k=1}^{p/2} b_k \sin k\omega t$$

onde os a_i e b_k são coeficientes reais. Depois de transmissão através de um canal ruidoso o sinal observado escreve-se

$$y(t) = x(t) + n(t) \quad t = 0, 1, \dots, N - 1$$

onde $n(t)$ representa o ruído de transmissão e $N > p$.

- construa o modelo linear para este problema.
- encontre o estimador dos MQ $\hat{\boldsymbol{\theta}}$ do vector de parâmetros $\boldsymbol{\theta} = [a_1, \dots, a_{p/2}, b_1, \dots, b_{p/2}]^T$.

c) encontre o estimador dos MQ $\hat{\mathbf{x}}$ de $\mathbf{x} = [x_0, x_1, \dots, x_{N-1}]^T$

d) calcule o elemento (i, j) da matriz de Gram do modelo linear acima. Sob que condição esta matriz é diagonal ?

Supondo que $n(t)$ é um processo gaussiano, branco, de média nula e variância unidade determinar

e) a distribuição de $\hat{\boldsymbol{\theta}}$?

f) a distribuição de $\hat{\mathbf{x}}$?

g) a distribuição de $\mathbf{y} - \hat{\mathbf{x}}$?

h) a distribuição de $N^{-1}(\mathbf{y} - \hat{\mathbf{x}})^T(\mathbf{y} - \hat{\mathbf{x}})$?