Estatística Computacional — Módulo 1

Notas de apoio (Cap. 1)

Manuel Cabral Morais

Secção de Estatística e Aplicações

Instituto Superior Técnico

Lisboa, Setembro–Outubro de 2003

Contents

Chapter 1

Optimização

Toda a gente tem a **tendência natural para optimizar**...
As companhias aéreas programam as suas equipas de bordo, aeronaves e vôos por forma a minimizar custos de operação.
Os investidores procuram criar *portfolios* que evitem riscos excessivos mas que simultaneamente *garantam* bons lucros.
Na indústria pretende-se eficiência máxima no planeamento e operação dos processos de produção.

A natureza também optimiza...
Os sistemas físicos tendem a encontrar-se em estados de energia mínima.
As moléculas num sistema químico isolado reagem umas com as outras até que a energia potencial total dos seus electrões seja mínima.
Os raios de luz seguem percursos que minimizem a duração desses mesmos percursos.

A optimização é fundamental em Estatística...

- Em **estimação pontual**, é usual seleccionar um estimador de um (ou mais) parâmetros desconhecido(s) que satisfaça determinado critério de optimalidade (variância mínima, máxima verosimilhança, mínimos quadrados, etc.).

- Em **inferência estatística** os testes de hipóteses são delineados de modo que sejam óptimos de acordo com certo critério.

  Por exemplo, à custa da aplicação do lema de Neyman–Pearson obtêm-se testes de hipóteses com a particularidade de minimizarem a probabilidade de cometer erro de 2a. espécie uma vez fixa a probabilidade de cometer erro de 1a. espécie.

- No campo da **metodologia das superfícies de resposta**, procuram-se condições de operação óptimas das váriáveis explicativas que produzam respostas máximas (mínimas) em determinada região de interesse.

  Por exemplo, no estudo de reacções químicas é importante determinar a temperatura de reacção e o tempo de reacção que maximizam a produção de certo produto.

- Ao lidarmos com **experiências multiresposta** (i.e., há não só várias variáveis explicativas como váriáveis de resposta) a optimização diz respeito a diversas funções de resposta.

  Basta pensar que se pode pretender maximizar a quantidade fabricada de certo produto e simultaneamente reduzir o custo de produção.

- Em **análise multivariada** é frequente lidar-se com um grande número de observações/variáveis. Por forma a facilitar a análise de dados é conveniente reduzir tais números sem que isso signifique grande perda de informação.

  É neste contexto que surgem as técnicas de redução de dados como é o caso das componentes principais.

O recurso a uma técnica de **optimização** pressupõe a identificação de:

- uma **função objectivo** (que represente, por exemplo, o lucro, o tempo, a energia potencial, a variância, o erro quadrático médio, verosimilhança, etc...);
- variáveis/incógnitas/**parâmetros** que influenciam a função objectivo; e
- eventuais **restrições** a que estão sujeitos os parâmetros.[1]

Posto isto é suposto obter os valores dos parâmetros que maximizam (minimizem) a função objectivo.

Caso se esteja a lidar com restrições, efectua-se o que é usualmente designado de **optimização restringida** (ou restrita).
Caso contrário, a **optimização** diz-se **irrestrita**.

Textos de apoio: Khuri (1993, pp. 326–328) e Nocedal e Wright (1999, pp. 1–3).

Um problema de optimização usual em Estatística consiste em determinar um conjunto de valores dos parâmetros $\underline{\theta} = (\theta_1, \ldots, \theta_p)$ que maximizam uma função objectivo $f(\underline{\theta}) = f(\theta_1, \ldots, \theta_p)$.
É o caso da maximização de uma função de verosimilhança ou a minimização de uma soma de quadrados.
De notar ainda que a maximização da função f é equivalente à minimização de $-f$. Assim, far-se-á, de um modo geral, referência somente à **minimização de uma função objectivo**.

[1] Por exemplo, uma probabilidade deve pertencer a $[0, 1]$.

À minimização poderão estar não só associados parâmetros sujeitos a restrições, como poderá surgir o problema da existência de mínimos locais. (Esquema — Everitt (1987, Fig. 1.1, p. 3).)

Procedimentos de minimização

1. **Ponto de mínimo** — A obtenção do **ponto de mínimo**, $\underline{\hat{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$, da função objectivo diferenciável $f(\underline{\theta})$ passa pela resolução do seguinte sistema de equações:

$$\left.\frac{\partial f(\theta_1, \ldots, \theta_p)}{\partial \theta_j}\right|_{\underline{\theta}=\underline{\hat{\theta}}} = 0, \;\; j = 1, \ldots, p. \tag{1.1}$$

   A satisfação das equações em **(1.1) é condição necessária mas não suficiente** para que se obtenha um ponto de mínimo (eventualmente local). Para que tal aconteça é fundamental que também se verifique:[2]

   - $\left.\frac{d^2 f(\theta)}{d\theta^2}\right|_{\theta=\hat{\theta}} > 0$, quando $p = 1$;
   - a matriz hessiana —

$$\mathbf{H}(\theta) = \nabla^2 f(\underline{\theta}) = [h_{ij}(\underline{\theta})]_{i,j=1,\ldots,p} \tag{1.2}$$

     onde $h_{ij}(\underline{\theta}) = \frac{\partial^2 f(\underline{\theta})}{\partial \theta_i \partial \theta_j}$ — seja definida positiva quando avaliada em $\underline{\hat{\theta}}$, quando $p > 1$.

2. **Procedimentos numéricos de minimização** — Caso não seja possível obter uma solução algébrica para o sistema (1.1), é necessário recorrer a **procedimentos numéricos** de minimização. Tratar-se-ão de **procedimentos iterativos** na medida em que fornecem soluções aproximadas sucessivas de $\underline{\hat{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$, com a particularidade de a solução aproximada na iteração $i + 1$

[2] Não só continuidade das segundas derivadas numa vizinhança do ponto de mínimo como...

ser, de uma forma geral, *melhor* que as soluções aproximadas obtidas nas iteração anteriores. I.e., ao considerar-se que $\underline{\theta}^{(i)}$ representa a aproximação do ponto de mínimo na iteração i ($i = 0, 1, 2, \ldots$), a solução aproximada inicial $\underline{\theta}^{(0)}$ e as soluções aproximadas sucessivas satisfarão, de um modo geral:

$$f\left(\underline{\theta}^{(0)}\right) \geq f\left(\underline{\theta}^{(1)}\right) \geq f\left(\underline{\theta}^{(2)}\right) \geq \ldots$$

- **Métodos de pesquisa directa** — As soluções aproximadas dependem somente dos valores de f obtidos durante o processo iterativo (logo não é necessário que f seja diferenciável).
- **Métodos do gradiente** — Estes métodos iterativos requerem o cálculo de (primeiras ou segundas) derivadas (parciais).

O procedimento numérico de minimização é dado por concluído de acordo com um **critério de convergência**, que poderá traduzir-se em condições tais como,

$$\left|f\left(\underline{\theta}^{(i+1)}\right) - f\left(\underline{\theta}^{(i)}\right)\right| < \epsilon \tag{1.3}$$

$$\left\|\underline{\theta}^{(i+1)} - \underline{\theta}^{(i)}\right\| < \eta, \tag{1.4}$$

onde quer ϵ, quer η são pré-especificados. É preferível requerer que a condição de convergência seja verificada para algumas iterações antes de dar por terminado o procedimento de minimização. ϵ deverá ter em conta a magnitude dos valores de f ou então deverá usar-se um critério de convergência que faça uso de erros relativos ao invés de erros absolutos.

3. **Minimização restringida** — Caso os parâmetros estejam sujeitos a restrições poderá tentar-se transformá-los de modo a que se tornem parâmetros irrestritos. Por exemplo, ao considerar-se

$\theta = P(\text{sucesso})$ numa prova de *Bernoulli* tem-se $\theta \in [0,1]$; ao usar a transformação logística, $\phi = \ln\left(\frac{\theta}{1-\theta}\right)$, tem-se $\phi \in I\!R$.

Caso não seja possível este tipo de transformação é necessário recorrer, por exemplo, a multiplicadores de Lagrange ou a técnicas de minimização restringida adequadas ao problema em mãos.

Exercício 1.1 — Obtenha a estimativa de máxima verosimilhança do parâmetro desconhecido θ, tendo por base uma amostra de dimensão n, $\underline{x} = (x_1, \ldots, x_n)$, proveniente da população:

(a) $X \sim Poisson(\theta)$;

(b) $X \sim Poisson - truncada(\theta)$ que exclua o valor 0.

Tenha o cuidado de averiguar que se tratam efectivamente de pontos de máximo, neste e nos exercícios que se seguem. ■

Exercício 1.2 — Obtenha as estimativas de mínimos quadrados dos dois parâmetros do modelo de regressão linear simples $Y_i = \theta_1 + \theta_2 x_i + \epsilon_i$. (Everitt (1987, pp. 5–7).) ■

Exercício 1.3 — Considere o modelo de regressão logística simples cuja variável resposta $Y_i \sim Bernoulli(p_i)$ onde

$$\begin{aligned} E(Y_i|x_i) &= p_i \\ &= \frac{\exp(\theta_1 + \theta_2 x_i)}{1 + \exp(\theta_1 + \theta_1 x_i)}. \end{aligned} \tag{1.5}$$

Deduza as estimativas de máxima verosimilhança dos parâmetros do modelo. (Everitt (1987, pp. 52–56).) ■

Texto de apoio: Everitt (1987, pp. 1–10).

1.1 Métodos de pesquisa unidimensional

Têm-se verificado consideráveis avanços nas técnicas de minimização nas últimas quatro décadas e este facto tem tido, naturalmente, grande impacto em vários ramos da Estatística.
Esta secção concentra-se em métodos de minimização quando se lida exclusivamente com um parâmetro. Destacar-se-ão dois **métodos de pesquisa directa** (por mera curiosidade) —

- Fibonacci
- Interpolação quadrática

— e três **métodos do gradiente** —

- gradiente (*steepest descent* ou do declive máximo)
- Newton
- Quasi-Newton

Os **métodos de pesquisa directa** de um ponto de mínimo de uma função de um só parâmetro (unidimensional) dividem-se, grosso modo, em **duas categorias**:

- aqueles em que é especificado o intervalo a que pertence o ponto de mínimo (e.g. método de Fibonacci); e
- aqueles em que é adiantada à partida uma solução aproximada (e.g. método de interpolação quadrática).

Na primeira das categorias assumir-se-á que o intervalo é conhecido e contém o ponto de mínimo e que a função é unimodal nesse mesmo intervalo como se poderá ver na sub-secção seguinte.

1.1.1 Método de Fibonacci

Comece-se por considerar que o ponto de mínimo da função $f(\theta)$, $\hat{\theta}$, pertence ao intervalo (θ_1, θ_2) e que se seleccionam dois pontos, θ_3 e θ_4, tais que $\theta_1 < \theta_3 < \theta_4 < \theta_2$.

Uma vez que se assume que a **função objectivo** é **unimodal** no intervalo (θ_1, θ_2) pode concluir-se que o ponto de mínimo se encontra no intervalo:

- (θ_3, θ_2), caso $f(\theta_3) \geq f(\theta_4)$;
- (θ_1, θ_4), caso $f(\theta_3) \leq f(\theta_4)$.

(Esquema — Everitt (1987, Fig. 2.1, p. 12).)

A **redução** progressiva da **amplitude do intervalo de pesquisa** passa pela avaliação da função noutros pontos escolhidos no último dos intervalos considerado.

A questão fundamental é, sem sombra de dúvida, obter esses pontos. É claro que a escolha dos subsequentes pontos (i.e., o próximo intervalo de pesquisa) não deve ser feita sem qualquer critério mas sim feita tendo em conta os valores da função obtidos anteriormente.

Caso se especifique à partida que a função só pode ser avaliada n (pares de) vezes, o procedimento de pesquisa mais eficiente é conhecido por método de Fibonacci.

Este procedimento faz uso de uma sequência de números inteiros designados por números de Fibonacci F_i que têm a particularidade de ser definidos pelas seguintes equações:

$$\begin{aligned} &F_0 = F_1 = 1 \\ &F_i = F_{i-1} + F_{i-2},\ i \geq 2. \end{aligned} \tag{1.6}$$

Método de Fibonacci — Os primeiros oito números de Fibonacci são iguais a 1, 1, 2, 3, 5, 8, 13, 21 e o procedimento de pesquisa pode resumir-se do seguinte modo:

1. Considerar o intervalo inicial de pesquisa $(\theta_1^{(1)}, \theta_2^{(1)})$ e $I_0 = \theta_2^{(1)} - \theta_1^{(1)}$ a respectiva amplitude.
2. Obter $I_1 = \frac{F_{n-1}}{F_n} \times I_0$.
3. Determinar o intervalo $(\theta_3^{(1)}, \theta_4^{(1)})$ cujos extremos são iguais a

$$\theta_3^{(1)} = \theta_1^{(1)} - I_1 = \theta_1^{(1)} + \frac{F_{n-2}}{F_n} \times I_0 \tag{1.7}$$

$$\theta_4^{(1)} = \theta_2^{(1)} + I_1 = \theta_1^{(1)} + \frac{F_{n-1}}{F_n} \times I_0 \tag{1.8}$$

4. Tirar partido da unimodalidade e restringir a pesquisa ao intervalo
   - $(\theta_1^{(2)}, \theta_2^{(2)}) = (\theta_3^{(1)}, \theta_2^{(1)})$, caso $f(\theta_3^{(1)}) \geq f(\theta_4^{(1)})$
   - $(\theta_1^{(2)}, \theta_2^{(2)}) = (\theta_1^{(1)}, \theta_4^{(1)})$, caso $f(\theta_3^{(1)}) \leq f(\theta_4^{(1)})$
5. Voltar ao passo 2 substituindo I_0 por $\theta_2^{(2)} - \theta_1^{(2)}$. ■

Exercício 1.4 — Na Tabela 2.1 de Everitt (1987, p. 13), parcialmente transcrita abaixo, encontram-se os 21 primeiros números de Fibonacci e a redução da amplitude do intervalo de pesquisa inicial (I_0) que advém do facto de decidirmos pela avaliação de n (pares de) valores da função f, redução definida por I_n/I_0.

Table 1.1: Redução da amplitude do intervalo de pesquisa.

n	0	1	2	3	4	5	6	7	8	9	10
F_n	1	1	2	3	5	8	13	21	34	55	89
I_n/I_0	1.0	1.0	0.5	1/3	0.2	0.1250	0.07692	0.04762	0.02941	0.01818	0.01124

Obtenha os valores da Tabela 1.1 e trace o gráfico de I_n/I_0. ■

Exercício 1.5 — Considere a seguinte tabela de frequências que dizem respeito à variável X com distribuição *Poisson Truncada* (que exclui o zero) e obtenha a estimativa de máxima verosimilhança sabendo à partida que a solução se encontra no intervalo $(0.7, 1.1)$.

Table 1.2: Frequências absolutas.

Valor (x_i)	1	2	3	4	5	6
Freq. Abs. de x_i (n_i)	1486	694	195	37	10	1

Recorra ao método de Fibonacci com $n = 20$ bem como a um programa na linguagem de programação que lhe for mais familiar. ■

Texto de apoio: Everitt (1987, pp. 11–14).

1.1.2 Interpolação quadrática

A aplicação do método de interpolação quadrática pressupõe grosso modo:

1. Considerar um valor inicial aproximado de $\hat{\theta}$, θ^*, bem como o *tamanho* dos passos (l).
2. Obter o valor da função $f(\theta)$ em três pontos relacionados com θ^* e l
   (a) $\theta_1 = \theta^*$
   (b) $\theta_2 = \theta^* + l$
   (c) $\theta_3 = \theta^* - l$, se $f(\theta_1) < f(\theta_2)$, $\theta_3 = \theta^* + 2l$, se $f(\theta_1) > f(\theta_2)$.
3. Obter uma função quadrática $f^*(\theta) = A\theta^2 + B\theta + C$ que interpole f (interpolação baseada nos três valores conhecidos de f).
4. Obter o próximo valor aproximado do ponto de mínimo para a iteração seguinte à custa dos coeficientes A e B: $\theta^* = \theta_{min} = -\frac{B}{2A}$.

Exercício 1.6 — Retome o Exercício 1.5 considerando à partida que o valor inicial aproximado de $\hat{\theta}$ é igual a $\theta^* = 0.8$. ■

Texto de apoio: Everitt (1987, pp. 14–16).

1.1.3 Método do gradiente (steepest descent)

O método de optimização descrito nesta subsecção, à semelhança dos descritos nas duas subsecções que se seguem, requerem o cálculo de valores de derivadas da função f bem como de valores dessa mesma função.[3]

O método do gradiente (*steepest descent*) encontra uma justificação numa importante **propriedade do vector gradiente**,

$$\nabla f(\underline{\theta}) = \left(\frac{\partial f(\underline{\theta})}{\partial \theta_1}, \ldots, \frac{\partial f(\underline{\theta})}{\partial \theta_p} \right). \tag{1.9}$$

Caso se desloque numa direcção definida por $\nabla f(\underline{\theta})$ a partir de um ponto arbitrário $(\underline{\theta}, f(\underline{\theta}))$, a função crescerá à taxa mais elevada.

Analogamente, se se deslocar numa direcção definida por $-\nabla f(\underline{\theta})$, a função decrescerá à taxa mais elevada.

Caso se considere um só parâmetro ($p = 1$) rapidamente se perceberá esta argumentação. (Esquema...)

O método do gradiente (*steepest descent*) procura explorar esta propriedade da direcção do vector gradiente. Assim, ao considerar-se que na $i-$ésima iteração se lida com a solução aproximada $\underline{\theta}^{(i)}$, a iteração seguinte fará uso da solução aproximada:

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \lambda^{(i)} \times \nabla f(\underline{\theta}^{(i)}), \tag{1.10}$$

[3] As observações feitas já de seguida dizem respeito a uma função real f com domínio $\mathbb{R}^p$ e serão particularizadas para o caso em que $p = 1$ nesta e nas duas subsecções seguintes.

onde a constante $\lambda^{(i)}$ é o ponto de mínimo da função

$$\phi(\lambda) = f\left[\underline{\theta}^{(i)} - \lambda \nabla f\left(\underline{\theta}^{(i)}\right)\right]. \tag{1.11}$$

A obtenção de $\lambda^{(i)}$ passa pela resolução (eventualmente numérica) da equação

$$\left.\frac{d\phi(\lambda)}{d\lambda}\right|_{\lambda=\lambda^{(i)}} = 0, \tag{1.12}$$

como sugere Alves (2000, p. 200).

De notar ainda que a resolução de (1.12) pode envolver demasiados cálculos, pelo que é usual aproximar a função $\phi(\lambda)$ por um polinómio de segundo grau usando para o efeito uma interpolação em três pontos próximos de 0. Para mais detalhes acerca desta interpolação consulte-se Alves (2000, pp. 201-202).

Exercício 1.7 — Após ter definido o método do gradiente (*steepest descent*) para o caso de uma função real de argumento real, retome o Exercício 1.5 e considere mais uma vez que o valor inicial aproximado de $\hat{\theta}$ é igual a $\theta^{(0)} = 0.8$. ■

O método do gradiente (*steepest descent*) poderá parecer à partida o melhor método para minimizar uma função. No entanto, a propriedade do gradiente que está por trás da sua definição é somente local e não uma propriedade global. Para além disso são frequentes as mudanças de direcção. Todos estes inconvenientes tornam o método ineficiente e a sua aplicação pouco recomendada.

Textos de apoio: Alves (2000, pp. 200–201), Everitt (1987, pp. 21–23) e Nocedal e Wright (1999, pp. 21–22).

1.1.4 Método de Newton-Raphson

O método de Newton (também conhecido por método de Newton-Raphson) faz uso de uma direcção de pesquisa derivada de uma **expansão de Taylor de segunda ordem**. Senão veja-se o caso de uma função real com um parâmetro $f(\theta)$.
Considere-se neste caso que a solução aproximada inicial é $\theta^{(0)}$ e efectue-se a referida expansão de $f(\theta)$ em torno do ponto $\theta^{(0)}$:

$$\begin{aligned} f(\theta) &\simeq f^*(\theta) \\ &= f\left(\theta^{(0)}\right) + \left(\theta - \theta^{(0)}\right) \times f'\left(\theta^{(0)}\right) \\ &\quad + \frac{1}{2}\left(\theta - \theta^{(0)}\right)^2 \times f''\left(\theta^{(0)}\right). \end{aligned} \tag{1.13}$$

Ao resolver-se a equação $\frac{df^*(\theta)}{d\theta} = 0$ obtém-se

$$\theta = \theta^{(0)} - \frac{f'\left(\theta^{(0)}\right)}{f''\left(\theta^{(0)}\right)}. \tag{1.14}$$

Deste modo o passo crucial na iteração do método de Newton-Raphson é a seguinte atribuição:

$$\theta^{(i+1)} = \theta^{(i)} - \frac{f'\left(\theta^{(i)}\right)}{f''\left(\theta^{(i)}\right)}. \tag{1.15}$$

Exercício 1.8 — Defina o método de Newton-Raphson para uma função real com p parâmetros, $f(\underline{\theta})$. ■

Exercício 1.9 — Considere de novo a v.a. Poisson truncada de parâmetro θ que exclui o zero.

(a) Prove que $\frac{1}{n}\sum_{r=2}^{+\infty} r\, n_r$, onde n_r representa a frequência absoluta do valor r, é uma estimativa centrada de θ, ao contrário do que acontece com a média da amostra.

(b) Use a estimativa em (a) como solução aproximada inicial na aplicação do método de Newton-Raphson ao conjunto de dados do Exercício 1.5.

(c) Repita (b) considerando como solução aproximada inicial a média da amostra.

(d) Compare os valores da primeira derivada da log-verosimilhança ao fim 6 iterações obtidas nas alíneas (b) e (c). ■

O método de Newton-Raphson **converge rapidamente** quando a **solução aproximada inicial** se encontra **perto do ponto de mínimo** já que, de um modo geral, $f^*(\theta)$ constitui uma boa aproximação de $f(\theta)$ na vizinhança de θ. O mesmo **não acontece** quando a **solução aproximada inicial** se encontra **distante do ponto de mínimo**.

As **desvantagens** do método quando $p > 1$ passam pela necessidade do **cálculo e inversão da matrix hessiana** em cada iteração e pela eventualidade de a matriz $\mathbf{H}\left(\underline{\theta}^{(i)}\right) = \nabla^2 f\left(\underline{\theta}^{(i)}\right)$ ser definida negativa caso $\underline{\theta}^{(i)}$ diste muito do ponto de mínimo.

Por este e outros motivos que se prendem, por exemplo, com a convergência do método de Newton-Raphson foram propostas na literatura variantes deste método que serão descritas na secção seguinte.

Textos de apoio: Everitt (1987, pp. 23–24) e Nocedal e Wright (1999, pp. 22–24).

1.2 Principais variantes do método de Newton-Raphson

As variantes do método de Newton-Raphson pretendem acelerar o processo de convergência daquele método, fazê-lo depender cada vez menos da solução aproximada inicial ou aligeirar os cálculos em cada iteração.

1.2.1 Fisher's scoring method

O *Fisher's scoring method* pode ser entendido como uma **variante estatística** do método de Newton-Raphson.
Este método resulta da **substituição da segunda derivada** $f''(\theta)$ (ou da matriz hessiana, caso $p > 1$) pelo seu valor esperado. Assim

$$\theta^{(i+1)} = \theta^{(i)} - \frac{f'\left(\theta^{(i)}\right)}{E\left[f''\left(\theta^{(i)}\right)\right]}. \tag{1.16}$$

De notar que, ao pretender-se maximizar uma função como a log-verosimilhança, i.e.

$$f(\theta) = \ln L(\theta|\underline{x}), \tag{1.17}$$

o valor esperado a que se refere a equação (1.16) não passa de

$$E\left[\frac{d^2 \ln L(\theta|\underline{X})}{d\theta^2}\right] \tag{1.18}$$

calculado no ponto $\theta^{(i)}$. Entenda-se $\ln L(\theta|\underline{X})$ como a v.a. que se obtém após se ter substituído x_i por X_i $(i = 1, \ldots, n)$ na expressão geral da função log-verosimilhança.
Este método de optimização tem-se revelado mais eficiente que o método de Newton-Raphson como se poderá constatar no exercício que se segue.

Exercício 1.10 — Após ter obtido a iteração do *Fisher's scoring method* para o Exercício 1.5 implemente o referido método e certifique-se que a solução aproximada é 0.89249, quando se considera $\theta^{(0)} = 2.0$. Compare o número de iterações necessárias para que este método e o de Newton-Raphson assegurem soluções aproximadas tais que $f'\left(\theta^{(i)}\right) \leq 10^{-6}$. ■

Exercício 1.11 — O número de partículas α emitidas por uma fonte radioactiva em 2612 unidades de tempo ($1/8min.$) estão condensadas na seguinte tabela de frequências:

Part. emitidas (i)	0	1	2	3	4	5	6	7	8	9	10	> 10
Freq. Abs de i	57	203	383	525	532	408	273	139	49	27	10	6

Assuma que os números de partículas emitidas por unidade de tempo são v.a.'s i.i.d. $Poisson(\theta)$ para responder às questões seguintes.

(a) Obtenha a função de verosimilhança, $L(\theta|\underline{n}) = L(\theta|n_0, \ldots, n_t, n_c)$, bem como a função de log-verosimilhança e verifique que

$$\frac{d \ln L(\theta|\underline{n})}{d\theta} = \frac{y}{\theta} - (n - n_c) + \frac{n_c p_t}{p_c} \tag{1.19}$$

$$\frac{d^2 \ln L(\theta|\underline{n})}{d\theta^2} = -\frac{y}{\theta^2} + \frac{n_c}{p_c}\left(p_{t-1} - p_t - \frac{p_t^2}{p_c}\right) \tag{1.20}$$

onde: n_i e n_c representam as frequências absolutas do valor i ($i = 0, \ldots, t$) e de valores maiores que t (respectivamente); $p_i = P(X = i|\theta), i = 0, \ldots, t$; $p_c = P(X > t|\theta)$; $n = \sum_{i=0}^{t} n_i + n_c$; e $y = \sum_{i=0}^{t} i\, n_i$.

(b) Elabore um programa que permita obter a estimativa de máxima verosimilhança de θ de acordo com o *Fisher's scoring method.*

(c) Execute o programa considerando $t = 10, 8, 6$ e comente os resultados obtidos. ■

É sabido que os problemas de optimização mais interessantes dizem respeito à situação em que lidamos com mais do que um parâmetro. A obtenção da estimativa de máxima verosimilhança do vector de p parâmetros $\underline{\theta} = (\theta_1, \ldots, \theta_p)$ é disso um exemplo.
Começe-se por relembrar que $\nabla f\left(\underline{\theta}^{(i)}\right)$ e $\mathbf{H}\left(\underline{\theta}^{(i)}\right)$ representam o vector gradiente e a matriz hessiana de $f(\underline{\theta})$ calculados no ponto $\underline{\theta}^{(i)}$, respectivamente, e fazer notar que as iterações do método de Newton-Raphson e do *Fisher's scoring method* podem encontrar-se na Tabela 1.3.

Table 1.3: Iterações de alguns métodos de pesquisa.

Método	Iteração
Newton-Raphson	$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \mathbf{H}^{-1}\left(\underline{\theta}^{(i)}\right) \nabla f\left(\underline{\theta}^{(i)}\right)$
Fisher's scoring	$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \left\{E\left[\mathbf{H}\left(\underline{\theta}^{(i)}\right)\right]\right\}^{-1} \nabla f\left(\underline{\theta}^{(i)}\right)$

Exercício 1.12 — Considere uma população bi-paramétrica à sua escolha para a qual não exista solução algébrica para a estimativa de máxima verosimilhança do vector de parâmetros $\underline{\theta} = (\theta_1, \theta_2)$ e determine o passo da iteração do *Fisher's scoring method*. ■

Outro aspecto ainda não abordado diz respeito ao **critério de convergência** de qualquer destes e de outros métodos de optimização. De um modo geral, prossegue-se a pesquisa enquanto a norma do vector gradiente não for suficientemente pequena, i.e., considera-se que o procedimento iterativo convergiu assim que

$$||\nabla f\left(\underline{\theta}^{(i)}\right)||^2 = \nabla f\left(\underline{\theta}^{(i)}\right)^\top \nabla f\left(\underline{\theta}^{(i)}\right) \leq \epsilon, \tag{1.21}$$

onde a constante ϵ diz respeito à precisão desejada.

Texto de apoio: Everitt (1987, pp. 23–24, 31–32).

1.2.2 Método de Newton-Raphson modificado

O método de Newton-Raphson modificado não consiste numa substituição da matriz hessiana mas sim na introdução do tamanho do passo em cada iteração. Deste modo passa a ter-se a iteração

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \lambda^{(i)} \, \mathbf{H}^{-1} \left(\underline{\theta}^{(i)} \right) \nabla f \left(\underline{\theta}^{(i)} \right), \tag{1.22}$$

onde $\lambda^{(i)}$ minimiza a função

$$\phi(\lambda) = f \left[\underline{\theta}^{(i)} - \lambda \, \mathbf{H}^{-1} \left(\underline{\theta}^{(i)} \right) \nabla f \left(\underline{\theta}^{(i)} \right) \right]. \tag{1.23}$$

O ponto de mínimo da função definida pela Equação (1.23), $\lambda^{(i)}$, é obtido recorrendo a qualquer dos métodos de pesquisa unidimensional descritos na secção anterior, caso não haja solução algébrica para a equação $\left. \frac{d\phi(\lambda)}{d\lambda} \right|_{\lambda=\lambda^{(i)}} = 0$.

A inclusão de $\lambda^{(i)}$ pretende essencialmente acelerar a convergência do método de Newton-Raphson. Contudo continua a pressupor o cálculo e a inversão da matriz hessiana em cada iteração — agravado, a nosso ver, por um problema adicional de optimização para obter $\lambda^{(i)}$.

Não surpreende pois que este método de optimização numérica seja preterido a favor de outros que contemplam a substituição da inversa da matriz hessiana por uma matriz simétrica definida positiva, actualizada em cada iteração, matriz que converge eventualmente para $\mathbf{H}^{-1}$. Estes últimos métodos são usualmente designados de métodos Quasi-Newton e serão descritos nas subsecções seguintes.

Exercício 1.13 — Resolva o Exercício 1.11 recorrendo ao método de Newton-Raphson modificado. ■

Texto de apoio: Nocedal e Wright (1999, pp. 141–142).

1.2.3 Método de Davidon-Fletcher-Powell

Em meados dos anos 50, o físico W.C. Davidon viu frustradas as suas tentativas de resolução de um complexo problema de optimização.[4] Em 1959, Davidon teve a brilhante ideia de acelerar o processo de convergência propondo um algoritmo — o primeiro algoritmo quasi-newtoniano — que começa como o método do gradiente (*steepest descent*) e posteriormente se resume ao método de Newton-Raphson em que a inversa da matriz hessiana é substituída e continuamente actualizada por uma matriz definida positiva.

Em 1963, Fletcher e Powell demostraram que o algoritmo de optimização proposto por Davidon era mais rápido e mais fiável que os métodos então existentes. Este avanço dramático transformou os problemas de optimização não linear de um dia para outro.[5]

A iteração contempla dois passos — a actualização da solução aproximada e da matriz que aproxima a inversa da matriz hessiana $\mathbf{H}^{-1}$ calculada em $\underline{\theta}^{(i)}$, $\mathbf{C}^{(i)}$:

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \lambda^{(i)}\ \mathbf{C}^{(i)}\ \nabla f\left(\underline{\theta}^{(i)}\right) \tag{1.24}$$

$$\mathbf{C}^{(i+1)} = \mathbf{C}^{(i)} + \mathbf{A}^{(i)} + \mathbf{B}^{(i)} \tag{1.25}$$

onde

$$\mathbf{A}^{(i)} = \frac{1}{\underline{z}^{\top}\,\underline{u}}\ \underline{z}\,\underline{z}^{\top} \tag{1.26}$$

$$\mathbf{B}^{(i)} = -\frac{1}{\underline{u}^{\top}\,\mathbf{C}^{(i)}\,\underline{u}}\ \mathbf{C}^{(i)}\underline{u}\,\underline{u}^{\top}\mathbf{C}^{(i)} \tag{1.27}$$

e

$$\underline{z} = -\lambda^{(i)}\,\mathbf{C}^{(i)}\,\nabla f\left(\underline{\theta}^{(i)}\right) \tag{1.28}$$

[4] O método a que Davidon recorreu denomina-se *coordinate descent method* e está descrito em Nocedal e Wright (1999, pp. 53–55) e não convergia.

[5] É curioso notar que o artigo em que Davidon propôs o método de optimização não foi aceite quando submetido para publicação e veio a ser publicado somente em 1991.

$$\underline{u} \;=\; \nabla f\left(\underline{\theta}^{(i+1)}\right) - \nabla f\left(\underline{\theta}^{(i)}\right). \tag{1.29}$$

De notar que é usual tomar $\mathbf{C}^{(0)} = \mathbf{I}$, e neste caso o primeiro passo da método de Davidon-Fletcher-Powell corresponde efectivamente ao método do gradiente (*steepest descent*). Como vem a ser hábito $\lambda^{(i)}$ é obtido por pesquisa unidimensional e corresponderá ao ponto de mínimo de

$$\phi(\lambda) = f\left[\underline{\theta}^{(i)} - \lambda \mathbf{C}^{(i)} \nabla f\left(\underline{\theta}^{(i)}\right)\right]. \tag{1.30}$$

A justificação para este procedimento encontra-se no trabalho de 1963 de Fletcher e Poweel. Essencialmente, a matriz $\mathbf{A}^{(i)}$ assegura que a sequência de matrizes $\mathbf{C}^{(i)}$ que aproximam a inversa da matriz hessiana convergem para essa mesma inversa. Por seu lado, a matriz $\mathbf{B}^{(i)}$ assegura que $\mathbf{C}^{(i)}$ é definida positiva.

Exercício 1.14 — A função de Rosenbrock é definida por

$$f(\theta_1, \theta_2) = 100(\theta_1^2 - \theta_2)^2 + (1 - \theta_1)^2 \tag{1.31}$$

e possui ponto de mínimo conhecido e igual a $(\theta_1, \theta_2) = (1, 1)$. Rosenbrock propôs o uso desta função para testar algoritmos de minimização, em 1960.

(a) Recorra ao método de Davidon-Fletcher-Powell para obter uma solução aproximada para o ponto de mínimo, considerando para o efeito a solução aproximada inicial $\underline{\theta}^{(0)} = (0, 0)$. (Everitt (1987, pp. 25–26).)

(b) Compare os resultados obtidos em (a) com os obtidos pelo método do gradiente (*steepest descent*) e de Newton-Raphson. ■

Textos de apoio: Everitt (1987, pp. 24–25), Khuri (1993, pp. 330–331) e Nocedal e Wright (1999, pp. 193–201).

1.2.4 Método do gradiente conjugado (Fletcher-Reeves)

A escolha da direcção de pesquisa em cada iteração é um passo extraordinariamente delicado e dele depende qualquer procedimento de minimização.

Para descrever o método de optimização que se segue é conveniente relembrar em que circunstâncias os vectores $\underline{p}$ e $\underline{q}$ se dizem vectores conjugados em relação à matriz definida positiva $\mathbf{G}$ —

$$\underline{p}^{\top}\mathbf{G}\underline{q} = 0 \tag{1.32}$$

— e já agora notar que a justificação para a utilização deste procedimento se encontra descrita em detalhe em Nocedal e Wright (1999, pp. 102–108) e passa pelo facto de o problema da resolução do sistema linear de equações $\mathbf{A}\underline{\theta} = \underline{b}$ ser equivalente ao problema de minimização da função quadrática

$$f(\underline{\theta}) = \frac{1}{2}\,\underline{\theta}^{\top}\,\mathbf{A}\,\underline{\theta} - \underline{b}^{\top}\,\underline{\theta}, \tag{1.33}$$

cujo gradiente é igual a

$$\nabla f(\underline{\theta}) = \mathbf{A}\underline{\theta} - \underline{b}. \tag{1.34}$$

O problema da minimização da função quadrática $f(\underline{\theta})$ definida pela Equação (1.33) resolve-se recorrendo a aquilo que Nocedal e Wright (1999, pp. 102) denominam de método das direcções conjugadas. Este método faz uso de um conjunto de p vectores não nulos e conjugados em relação à matriz $\mathbf{A}$, $\{\underline{q}^{(0)}, \ldots, \underline{q}^{(p-1)}\}$. A sua iteração $i+1$ é dada por:

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \lambda^{(i)}\,\underline{q}^{(i)} \tag{1.35}$$

onde $\lambda^{(i)}$ minimiza a função

$$\phi(\lambda) = f\left(\underline{\theta}^{(i)} - \lambda\underline{q}^{(i)}\right), \tag{1.36}$$

i.e.,

$$\lambda^{(i)} = \frac{\left(\mathbf{A}\underline{\theta}^{(i)} - \underline{b}\right)^{\top} \underline{q}^{(i)}}{\underline{q}^{(i)\top}\mathbf{A}\underline{q}^{(i)}}. \tag{1.37}$$

O primeiro método do gradiente conjugado não linear surge nos anos 60 e deve-se a Fletcher e Reeves. Ao considerar-se uma função regular $f(\underline{\theta})$ este método possui direcção de pesquisa e iteração dadas, de acordo com Alves (2000, pp. 206-207), por:

$$\underline{d}^{(i)} = -\nabla f\left(\underline{\theta}^{(i)}\right) + \frac{\nabla f\left(\underline{\theta}^{(i)}\right)^{\top} \nabla f\left(\underline{\theta}^{(i)}\right)}{\nabla f\left(\underline{\theta}^{(i-1)}\right)^{\top} \nabla f\left(\underline{\theta}^{(i-1)}\right)} \underline{d}^{(i-1)} \tag{1.38}$$

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \lambda^{(i)}\, \underline{d}^{(i)} \tag{1.39}$$

onde $\lambda^{(i)}$ é, mais uma vez, o ponto de mínimo da função $\phi(\lambda) = f\left(\underline{\theta}^{(i)} - \lambda \underline{d}^{(i)}\right)$. Para além disso, faz sentido considerar a primeira direcção de pesquisa igual à direcção do método do gradiente (*steepest descent*), ou seja, $\underline{d}^{(0)} = -\nabla f\left(\underline{\theta}^{(0)}\right)$.
Fletcher e Reeves sugerem que de p em p iterações se adopte a direcção do método do gradiente (*steepest descent*), $-\nabla f\left(\underline{\theta}^{(i)}\right)$.

A direcções $\underline{d}^{(i)}$ têm a particularidade de ser conjugadas e, caso se estivesse a minimizar uma função quadrática com p parâmetros, o procedimento de minimização convergiria quando muito em p passos.

Exercício 1.15 — Repita o Exercício 1.14 aplicando desta feita o método do gradiente conjugado (Fletcher-Reeves) à minimização da função de Rosenbrock. (Everitt (1987, pp. 25–26).)
Compare os resultados obtidos pelos diversos métodos. ■

Textos de apoio: Alves(2000, pp. 204-207), Everitt (1987, pp. 25–27) e Nocedal e Wright (1999, pp. 102–112).

1.2.5 Método de Broyden-Fletcher-Goldfarb-Shanno

O método de Broyden-Fletcher-Goldfarb-Shanno é provavelmente o método Quasi-Newton mais popular. Tem a particularidade de aproximar a matriz hessiana $\mathbf{H}\left(\underline{\theta}^{(i)}\right)$ pela matriz $\mathbf{B}^{(i)}$ que, na iteração $i+1$, é definida por

$$\mathbf{B}^{(i+1)} = \mathbf{B}^{(i)} - \frac{\mathbf{B}^{(i)}\,\underline{\phi}^{(i)}\underline{\phi}^{(i)\top}\,\mathbf{B}^{(i)}}{\underline{\phi}^{(i)\top}\,\mathbf{B}^{(i)}\,\underline{\phi}^{(i)}} + \frac{\underline{u}^{(i)}\,\underline{u}^{(i)\top}}{\underline{\phi}^{(i)\top}\,\underline{u}^{(i)}} \tag{1.40}$$

onde

$$\underline{\phi}^{(i)} = \underline{\theta}^{(i+1)} - \underline{\theta}^{(i)} \tag{1.41}$$

$$\underline{u}^{(i+1)} = \nabla f\left(\underline{\theta}^{(i+1)}\right) - \nabla f\left(\underline{\theta}^{(i)}\right). \tag{1.42}$$

A direcção de pesquisa é neste caso dada por

$$\underline{d}^{(i)} = -[\mathbf{B}^{(i)}]^{-1}\,\nabla f\left(\underline{\theta}^{(i)}\right), \tag{1.43}$$

a iteração por

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \lambda^{(i)}\,\underline{d}^{(i)}, \tag{1.44}$$

e $\mathbf{B}^{(0)} = \mathbf{H}\left(\underline{\theta}^{(0)}\right)$.

Na Tabela 1.9 poderá encontrar uma lista das rotinas da *NAG* (*The Numerical Algorithms Group Ltd.*) com alguns dos métodos de optimização até agora descritos.

O endereço http://www.nag.co.uk/numeric/numerical_libraries.asp poderá ser de alguma utilidade...

Tanto quanto se pode averiguar o *package Mathematica* possui uma rotina denominada *FindMinimum* que faz uso do método do gradiente (*steepest descent*) para a pesquisa numérica de pontos de mínimo.

Table 1.4: Algumas rotinas de minimização da NAG.

Rotina	Método
E04DGF	Gradiente conjugado (requer que se forneça valores da função e do vector gradiente)
E04JAF	Quasi-Newton (requer que se forneça somente valores da função)
E04KAF	Quasi-Newton — Broyden-Fletcher-Goldfarb-Shanno (requer que se forneça valores da função e do vector gradiente)
E04KCF	Newton-Raphson modificado (requer que se forneça valores da função e do vector gradiente)
E04LAF	Newton-Raphson modificado (requer que se forneça valores da função, do gradiente e da matriz hessiana)

Exercício 1.16 — Faça um levantamento das rotinas/métodos de minimização dos *packages BMDP*, *Maple*, *Matlab*, *Mathematica*, *NAG*, *R*, *SAS*, *SPSS*, *Statistica* e outros *packages* com que esteja familiarizado. ■

Exercício 1.17 — Os habitantes de determinada população sofrem de uma doença congénita que afecta a visão e cujos efeitos se tornam mais evidentes com a idade. Foram recolhidas amostras de 50 pessoas de 5 grupos etários, tendo-se registado o número de pessoas cegas na Tabela 1.5.
Considere desta feita o modelo de regressão logística com variável resposta $Y_i \sim Binomial(n_i, p_i)$, onde

$$\begin{aligned} E(Y_i|x_i) &= n \times p_i \\ &= n \times \frac{\exp(\theta_1 + \theta_2 x_i)}{1 + \exp(\theta_1 + \theta_2 x_i)}, \end{aligned} \tag{1.45}$$

— ou, equivalentemente,

$$\ln\left(\frac{p_i}{1+p_i}\right) = \theta_1 + \theta_2 x_i \tag{1.46}$$

Table 1.5: Frequência de pessoas cegas por grupo etário.

Idade (x_i)	20	35	45	55	70
Total de pessoas (n_i)	50	50	50	50	50
No. pessoas cegas (y_i)	6	17	26	37	44

— e determine as estimativas de máxima verosimilhança dos parâmetros do modelo utilizando um método de optimização à sua escolha. ■

Exercício 1.18 — Em determinada experiência planeada para simular uma operação de produção, solicita-se a um funcionário que desempenhe uma tarefa rotineira repetidamente durante um período de tempo fixo. A experiência é efectuada com uma máquina que opera às velocidades 1, 2, 3, 4 e 5.
Registou-se o número de erros cometidos pelo funcionário em 25 períodos de tempo iguais, 5 para cada velocidade, tendo-se obtido a seguinte tabela de dados:

Table 1.6: No. de erros cometidos para cada velocidade.

Velocidade (x_i)	1	2	3	4	5
No. erros (y_i)	2	7	25	47	121

Assuma que o número de erros cometidos Y_i é uma v.a. com distribuição de $Poisson(m_i)$, quando a velocidade a que opera a máquina é igual a x_i, onde

$$\ln(m_i) = [E(Y_i|x_i)] = \theta_1 + \theta_2 x_i. \tag{1.47}$$

Obtenha as estimativas de máxima verosimilhança dos parâmetros deste modelo log-linear utilizando o *Fisher's scoring method.* ■

Texto de apoio: Nocedal e Wright (1999, pp. 193–201).

1.2.6 Aplicações a modelos lineares generalizados

Os modelos de regressão linear, logística e log-linear descritos em exemplos anteriores são casos particulares do que usualmente se denomina de modelos lineares generalizados.
Um **modelo linear generalizado** tem a particularidade de possuir:

- uma **componente aleatória** — $Y_1, \ldots, Y_n$ são v.a.s independentes (respostas) tais que

$$E[Y_i|x_i(1), \ldots, x_i(p)] = m_i, \tag{1.48}$$

  onde $x_i(1), \ldots, x_i(p)$ são os valores das p variáveis explicativas associadas à $i-$ésima resposta ($i = 1, \ldots, n$);

- uma **componente sistemática** — o preditor linear

$$\eta_i = \theta_1 x_i(1) + \ldots + \theta_p x_i(p); \tag{1.49}$$

- e uma **função de ligação** entre as componentes aleatória e sistemática

$$\lambda(m_i) = \eta_i. \tag{1.50}$$

Na Tabela 1.7 podem encontrar-se as funções de ligação dos modelos de regressão linear, logística e log-linear.
Qualquer destas funções de ligação define-se para valores reais logo a maximização da função de log-verosimilhança em ordem aos θ_j's é um problema de optimização irrestrita.

Table 1.7: Algumas funções de ligação.

Modelo	Função de ligação $\lambda(m_i)$
Regressão linear	m_i
Regressão logística	$\ln\left(\frac{m_i}{n_i - m_i}\right)$ onde $m_i = n_i p_i$
Regressão log-linear	$\ln(m_i)$

O *software* estatístico **GLIM** assume que a f.p. da v.a. resposta Y (ou f.d.p. se a v.a. resposta for contínua) possui a seguinte forma genérica:

$$f_Y(y) = \exp\left\{\frac{y\beta - b(\beta)}{a(\phi)} + c(y,\phi)\right\}. \tag{1.51}$$

Exercício 1.19 — Após ter completado a tabela seguinte, onde $a(\phi) = \phi$,

Modelo	β	$b(\beta)$	ϕ	$c(y,\phi)$
$normal(\mu, \sigma^2)$	μ	$\frac{1}{2}\beta^2$	σ^2	$-\frac{1}{2} \times \left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]$
$Poisson(\mu)$	$\ln(\mu)$		1	
$binomial(N, p)$		$N\ln(1 + e^{\beta})$		
$gama(\alpha, \frac{\alpha}{\mu})$				$(\alpha - 1)\ln(y) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)]$

demonstre que, para uma v.a. genérica Y, se tem $E(Y) = b'(\beta)$ e $V(Y) = b''(\beta)a(\phi)$. Verifique ainda estes resultados para as quatro distribuições aqui consideradas. ■

É altura de se falar da **estimação de máxima verosimilhança** do vector de parâmetros $\underline{\theta} = (\theta_1, \ldots, \theta_p)$ do preditor linear $\eta_i = \theta_1 x_i(1) + \ldots + \theta_p x_i(p)$.

Para tal é necessário obter a função de log-verosimilhança no contexto de um modelo linear generalizado, com base no vector dos valores

observados das respostas $\underline{y} = (y_1, \ldots, y_n)$ e na matriz $n \times p$ com os valores das variáveis explicativas $\mathbf{X} = [x_i(j)]_{i=1,\ldots,n;j=1,\ldots,p}$:

$$\ln L(\underline{\theta}|\underline{y}, \mathbf{X}) = \sum_{i=1}^{n} \left[\frac{y_i\beta_i - b_i(\beta_i)}{a(\phi)} + c_i(y_i, \phi) \right] \tag{1.52}$$

Apesar de a Equação (1.52) estar escrita à custa dos β_i's convém realçar que a maximização se fará em relação ao vector de parâmetros $\underline{\theta}$ de que depende o preditor linear.

Assim — ao considerar-se que

$$\mathcal{L}_i = \frac{y_i\beta_i - b_i(\beta_i)}{a(\phi)} + c_i(y_i, \phi), \tag{1.53}$$

ao relembrar-se que $m_i = E(Y_i) = b'(\beta_i)$ e $\frac{dm_i}{d\beta_i} = b''(\beta_i)$, e ao tomar-se

$$w_i = \frac{1}{V(Y_i)} \left(\frac{dm_i}{d\eta_i} \right)^2, \tag{1.54}$$

— conclui-se que, para $r = 1, \ldots, p$,

$$\begin{aligned} \frac{\partial \ln L(\underline{\theta}|\underline{y}, \mathbf{X})}{\partial \theta_r} &= \sum_{i=1}^{n} \frac{\partial \mathcal{L}_i}{\partial \beta_i} \frac{d\beta_i}{dm_i} \frac{dm_i}{d\eta_i} \frac{d\eta_i}{d\theta_r} \\ &= \sum_{i=1}^{n} w_i(y_i - m_i) \frac{d\lambda(m_i)}{dm_i} x_i(r) \end{aligned} \tag{1.55}$$

Exercício 1.20 — Prove o resultado (1.55) e calcule

$$\frac{\partial^2 \ln L(\underline{\theta}|\underline{y}, \mathbf{X})}{\partial \theta_r \partial \theta_s} \tag{1.56}$$

por forma a verificar que a **matriz de informação de Fisher** satisfaz

$$\begin{aligned} \mathbf{I}(\underline{\beta}) &= \left[-E\left(\frac{\partial^2 \ln L(\underline{\theta}|\underline{y}, \mathbf{X})}{\partial \theta_r \partial \theta_s} \right) \right]_{r,s=1,\ldots,p} \\ &= \sum_{i=1}^{n} w_i x_i(r) x_i(s). \end{aligned} \tag{1.57}$$

Defina por fim a iteração do *Fisher's scoring method*, por sinal usada no GLIM. ■

Exercício 1.21 — Retome o Exercício 1.17 referente a um modelo de regressão logística cuja variável resposta é o número de pessoas afectadas pela cegueira para o grupo etário i com idade x_i que verifica $Y_i \sim Binomial(n_i, p_i)$, onde

$$\begin{aligned} E(Y_i|x_i) &= n \times p_i \\ &= n \times \frac{\exp(\theta_1 + \theta_2 x_i)}{1 + \exp(\theta_1 + \theta_2 x_i)}. \end{aligned} \tag{1.58}$$

(a) Tirando partido do vector gradiente e matriz hessiana derivados nesta subsecção, expresse a iteração do método de Newton-Raphson para a obtenção das estimativas de máxima verosimilhança à custa de uma equação do tipo

$$\mathbf{A}^{(i)}\underline{\theta}^{(i)} = \underline{b}^{(i)} \tag{1.59}$$

onde $\mathbf{A}^{(i)}$ é uma matriz 2×2 e $\underline{b}^{(i)}$ um vector 2×1 que dependem exclusivamente de $\theta_1^{(i)}$ e $\theta_2^{(i)}$.

(b) Verifique que as entradas de $\mathbf{A}^{(i)}$ e $\underline{b}^{(i)}$ se obtêm à custa das que figuram na Equação (1.57). ■

Texto de apoio: Everitt (1987, pp. 56–58).

1.3 Alguns algoritmos para o problema de mínimos quadrados não lineares

As técnicas de optimização descritas nas secções anteriores requerem um número considerável de operações, pelo que não é de estranhar que a sua aplicação rotineira no domínio da Estatística só tenha sido possível com o advento de computadores muito rápidos.
Esta secção concentrar-se-á em alguns algoritmos para o problema de mínimos quadrados não lineares. Estes resultam por vezes de melhorias de métodos de optimização já existentes, melhorias estas possíveis dada a estrutura específica do problema. Estes algoritmos aliados à velocidade dos computadores dos dias de *hoje* permitem resolver problemas de **estimação pelo método dos mínimos quadrados** no contexto da **regressão não linear**, algo impensável há, por exemplo, três décadas atrás.

No Exercício 1.2 foram obtidas estimativas dos parâmetros do modelo de regressão linear simples — $E(Y_i|x_i) = \theta_1 + \theta_2 x_i$ — recorrendo para o efeito ao método dos míninos quadrados. Estas estimativas obtêm-se por minimização da soma de quadrados

$$h(\underline{\theta}) = h(\theta_1, \theta_2) = \sum_{i=1}^{n} [y_i - (\theta_1 + \theta_2 x_i)]^2 \tag{1.60}$$

em ordem a (θ_1, θ_2). Ter-se-á obtido as duas bem conhecidas expressões para as estimativas dos mínimos quadrados:

$$\hat{\theta}_1 = \overline{y} - \hat{\theta}_2 \, \overline{x} \tag{1.61}$$

$$\hat{\theta}_2 = \frac{\sum_{i=1}^{n} x_i y_i - n \, \overline{x} \, \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \, \overline{x}^2}. \tag{1.62}$$

No entanto, alguns dos problemas aplicados de maior interesse envolvem uma v.a. resposta Y categórica/discreta em vez de contínua ou então que se relaciona de modo não linear com a(s) variável(is) explicativa(s).

Basta pensar nos modelos de regressão logística e log-linear e no seguinte modelo descrito por Everitt (1987, p. 43) que diz respeito à concentração de iões de ligados (B) e livres (F) em equilíbrio num receptor[6] é dada pela equação não linear

$$B = \frac{\theta_1 F}{\theta_2 + F} \tag{1.63}$$

onde θ_1 e θ_2 se denominam afinidade e capacidade do sistema receptor, respectivamente.[7] É claro que este modelo poderia ser transformado num modelo de regressão linear simples ao considerar-se desta feita as variáveis de $1/B$ e $1/F$ e o modelo

$$\frac{1}{B} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\frac{1}{F}. \tag{1.64}$$

e obter as estimativas de mínimos quadrados para a ordenada na origem $\frac{1}{\theta_1}$ e o declive $\frac{\theta_2}{\theta_1}$. No entanto, este tipo de *linearização* do modelo original raramente é possível, pelo que o utilizador se vê confrontado com um problema de minimização numérica a resolver por recurso a métodos como o do gradiente (*steepest descent*), o de Newton-Raphson, o *Fisher's scoring method* e o método do gradiente conjugado, entre outros.

[6] Tradução livre de *concentration of bound (B) and free (F) ligands at equilibrium in a receptor assay.*

[7] Na versão *estatística* do modelo deve figurar um erro aleatório no segundo membro da respectiva equação.

De um modo geral lida-se com

- Y_i, a $i-$ésima **resposta aleatória** $(i = 1, \ldots, n)$,
- $x_i(1), \ldots, x_i(p)$, os valores das p **variáveis explicativas** associadas à $i-$ésima resposta $(i = 1, \ldots, n)$ e
- $\underline{\theta} = (\theta_1, \ldots, \theta_p)$, o **vector** de p **parâmetros**.

Partir-se-á do princípio que o valor esperado de Y_i se relaciona funcionalmente com os valores $x_i(1), \ldots, x_i(p)$ através do preditor

$$\begin{aligned} E[Y_i|x_i(1), \ldots, x_i(p)] &= m[x_i(1), \ldots, x_i(p); \underline{\theta}] \\ &= m_i. \end{aligned} \tag{1.65}$$

Eis alguns exemplos triviais de preditores:

- linear múltiplo — $E(Y_i|\underline{x}_i) = \theta_1 x_i(1) + \ldots + \theta_p x_i(p)$
- não lineares (simples) — $E(Y_i|x_i) = \theta_1 + \theta_2 e^{-\theta_3 x_i}$, $E(Y_i|x_i) = \frac{\exp(\theta_1+\theta_2 x_i)}{1+\exp(\theta_1+\theta_2 x_i)}$, etc.

Em qualquer dos casos pretende minimizar-se a seguinte **soma de quadrados**:

$$\begin{aligned} h(\underline{\theta}) &= \sum_{i=1}^{n} h_i^2(\underline{\theta}) \\ &= \sum_{i=1}^{n} \left\{ y_i - m[x_i(1), \ldots, x_i(p); \underline{\theta}] \right\}^2 \\ &= \sum_{i=1}^{n} (y_i - m_i)^2 \end{aligned} \tag{1.66}$$

em ordem a $\underline{\theta}$.[8]

[8] Caso $V(Y_i)$ não seja constante e dependa de i é usual minimizar a seguinte soma de quadrados pesados: $\sum_{i=1}^{n} w_i \left(y_i - m_i\right)^2$ onde os w_i's são pesos escolhidos adequadamente.

Saliente-se que o vector gradiente, $\underline{g}(\underline{\theta}) = \nabla h(\underline{\theta})$, e a matriz hessiana de $h(\underline{\theta})$, $\mathbf{H}(\underline{\theta}) = \nabla^2 h(\underline{\theta})$, possuem entradas j $(j = 1, \ldots, p)$ e (j, k) $(j, k = 1, \ldots, p)$ iguais a

$$\frac{\partial h(\underline{\theta})}{\partial \theta_j} = 2 \sum_{i=1}^{n} h_i(\underline{\theta}) \frac{\partial h_i(\underline{\theta})}{\partial \theta_j} \tag{1.67}$$

$$\frac{\partial^2 h(\underline{\theta})}{\partial \theta_j \partial \theta_k} = 2 \left\{ \sum_{i=1}^{n} \frac{\partial h_i(\underline{\theta})}{\partial \theta_j} \frac{\partial h_i(\underline{\theta})}{\partial \theta_k} + \sum_{i=1}^{n} h_i(\underline{\theta}) \frac{\partial^2 h_i(\underline{\theta})}{\partial \theta_j \partial \theta_k} \right\} \tag{1.68}$$

respectivamente. Então, ao considerar-se o vector $(n \times 1)$

$$\underline{H}(\underline{\theta}) = [h_i(\underline{\theta})]_{i=1,\ldots,n}\,, \tag{1.69}$$

a matriz jacobiana $(n \times p)$

$$\mathbf{J}(\underline{\theta}) = \left[\frac{\partial h_i(\underline{\theta})}{\partial \theta_j} \right]_{i=1,\ldots,n; j=1,\ldots,p} \tag{1.70}$$

e ainda as matrizes auxiliares $(p \times p)$

$$\nabla^2 h_i(\underline{\theta}) = \left[\frac{\partial^2 h_i(\underline{\theta})}{\partial \theta_j \partial \theta_k} \right]_{j,k=1,\ldots,p} \tag{1.71}$$

$$\mathbf{Q}(\underline{\theta}) = \sum_{i=1}^{n} h_i(\underline{\theta}) \nabla^2 h_i(\underline{\theta}), \tag{1.72}$$

pode concluir-se que o **vector gradiente** e a **matriz hessiana** da **soma de quadrados** $h(\underline{\theta})$ são dados — matricialmente — por

$$\underline{g}(\underline{\theta}) = 2\mathbf{J}(\underline{\theta})^\top \underline{H}(\underline{\theta}) \tag{1.73}$$

$$\mathbf{H}(\underline{\theta}) = 2 \left[\mathbf{J}(\underline{\theta})^\top \mathbf{J}(\underline{\theta}) + \mathbf{Q}(\underline{\theta}) \right], \tag{1.74}$$

respectivamente.

Está-se, por fim, em condições de escrever as iterações de diversos métodos de optimização devidamente adaptados à minimização numérica de somas de quadrados.

1.3.1 Métodos de Newton-Raphson, Gauss-Newton e Newton-Raphson modificado

Ao dispor do vector gradiente e da matriz hessiana de $h(\underline{\theta})$ a iteração **método de Newton-Raphson** é dada por

$$\begin{aligned}\underline{\theta}^{(i+1)} &= \underline{\theta}^{(i)} - \left[\mathbf{H}\left(\underline{\theta}^{(i)}\right)\right]^{-1} \underline{g}\left(\underline{\theta}^{(i)}\right) \\ &= \underline{\theta}^{(i)} - \left[\mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top} \mathbf{J}\left(\underline{\theta}^{(i)}\right) + \mathbf{Q}\left(\underline{\theta}^{(i)}\right)\right]^{-1} \\ &\qquad \times \mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top} \underline{H}\left(\underline{\theta}^{(i)}\right). \qquad (1.75)\end{aligned}$$

Uma **alteração possível** ao método de Newton-Raphson por forma a aligeirar o trabalho numérico passa por considerar desprezável a matriz $\mathbf{Q}\left(\underline{\theta}^{(i)}\right)$ em $\mathbf{H}\left(\underline{\theta}^{(i)}\right)$ o que é perfeitamente razoável já que o factor $h_i\left(\underline{\theta}^{(i)}\right)$ se torna cada vez mais pequeno à medida que $\underline{\theta}^{(i)} \rightarrow \underline{\hat{\theta}}$, tal como acontece aquando da aplicação do método dos mínimos quadrados ao modelo de regressão linear. De notar que ao efectuar esta modificação deixou de ser necessário calcular segundas derivadas e a iteração passa a:

$$\theta^{(i+1)} = \theta^{(i)} - \left[\mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top} \mathbf{J}\left(\underline{\theta}^{(i)}\right)\right]^{-1} \mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top} \underline{H}\left(\underline{\theta}^{(i)}\right). \quad (1.76)$$

Este algoritmo de minimização será denominado de **método de Gauss-Newton** como o fazem Nocedal e Wright (1999, pp. 259–264) (ou de **método de Newton-Raphson linearizado**).

Exercício 1.22 — Prove que a iteração em (1.76) corresponde à da aplicação do método dos mínimos quadrados após uma *linearização* do preditor m_i à custa de uma expansão de Taylor de primeira ordem. ■

Exercício 1.23 — Considerou-se que a evolução do número de vendas y de um *software* de sistema ao longo de 9 meses desde o seu lançamento ($t = 0, 1, \ldots, 9$) é bem modelada pela seguinte equação diferencial

$$\frac{dy}{dt} + \theta_1 y = \theta_2 + \theta_3 t, \tag{1.77}$$

cuja solução é

$$y = k_0 + k_1 t + k_2 e^{-\theta_1 t}, \tag{1.78}$$

onde $k_0 = \frac{\theta_1\theta_2 - \theta_3}{\theta_1^2}$, $k_1 = \frac{\theta_3}{\theta_1}$, $_2 = \frac{\theta_3 + \theta_1\theta_2(\theta_1 - 1)}{\theta_1^2}$.

(a) Averigue graficamente a razoabilidade do modelo de regressão não linear sugerido acima ao conjunto de dados da Tabela 1.8.

Table 1.8: Evolução de vendas de *software*.

Tempo (t_i)	Vendas (y_i)
0	1990
1	2025
2	2440
3	2515
4	2800
5	3060
6	3085
7	3225
8	3220
9	3240

(b) Obtenha as estimativas de mínimos quadrados dos parâmetros $(\theta_1, \theta_2, \theta_3)$ recorrendo para o efeito ao método de Gauss-Newton (ou Newton-Raphson *linearizado*) com estimativa inicial:

$$\underline{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}) = (0.6, 1900, 48).$$

■

A **NAG** dispõe de algumas rotinas que permitem obter estimativas de mínimos quadrados.

Table 1.9: Algumas rotinas para o método dos mínimos quadrados (NAG).

Rotina	Método
E04HFF	Newton-Raphson (requer que se forneça o gradiente e a matriz hessiana)
E04GCF	Quasi-Newton (requer que se forneça o vector gradiente; a matriz hessiana é aproximada de modo comparável ao método BGFS)
E04GEF	Newton-Raphson modificado (requer que se forneça o vector gradiente; as segundas derivadas são aproximadas por diferenção finitas)
E04KCF	Newton-Raphson modificado (as primeiras e segundas derivadas são aproximadas)

Estas rotinas baseam-se, por exemplo, no **método de Newton-Raphson modificado** que faz uso da constante $\lambda^{(i)}$ e de uma *linearização* da matriz hessiana (sob certas condições) e possui iteração dada por:

- caso a redução da soma de quadrados na última iteração tiver sido grande,

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \lambda^{(i)} \times \left[\mathbf{H}\left(\underline{\theta}^{(i)}\right)\right]^{-1} \times \mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top} \underline{H}\left(\underline{\theta}^{(i)}\right), \quad (1.79)$$

- caso contrário,

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \lambda^{(i)} \left[\mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top} \mathbf{J}\left(\underline{\theta}^{(i)}\right)\right]^{-1} \times \mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top} \underline{H}\left(\underline{\theta}^{(i)}\right), \quad (1.80)$$

onde $\lambda^{(i)}$ se obtém por minimização unidimensional de

$$\phi(\lambda) = h\left[\underline{\theta}^{(i)} - \lambda \times \left[\mathbf{H}\left(\underline{\theta}^{(i)}\right)\right]^{-1} \underline{g}\left(\underline{\theta}^{(i)}\right)\right]. \tag{1.81}$$

Exercício 1.24 — Considere o seguinte conjunto de dados que se reportam à concentração de iões de ligados (B) e livres (F) em equilíbrio num receptor.

Table 1.10: Concentração de iões de livres (F) e ligados (B) em equilíbrio num receptor.

Livres (f_i)	Ligados (b_i)
84.6	12.1
83.9	12.5
148.2	17.2
147.8	16.7
463.9	28.3
463.8	26.9
964.1	37.6
967.6	35.8
1925.0	38.5
1900.0	39.9

(a) Elabore um gráfico que permite averiguar a razoabilidade do modelo linear

$$\frac{1}{B} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\frac{1}{F}. \tag{1.82}$$

Estime os parâmetros θ_1 e θ_2 pelo método dos mínimos quadrados.

(b) Obtenha agora as estimativas dos mínimos quadrados dos parâmetros do modelo não linear

$$B = \frac{\theta_1 F}{\theta_2 + F} \tag{1.83}$$

recorrendo para o efeito aos métodos de Newton-Raphson e de Newton-Raphson modificado com $\underline{\theta}^{(0)} = (10.00, 50.00)$ (Everitt (1987, pp. 45–47)).

Compare os resultados obtidos pelos dois métodos de minimização numérica.

(c) Obtenha estimativas da variância dos estimadores dos parâmetros obtidos pelo método de Newton-Raphson. ■

Exercício 1.25 — Determinada reacção química pode ser descrita pelo modelo de regressão não linear

$$Y = \frac{\theta_1 \theta_3 X_1}{1 + \theta_1 X_1 + \theta_2 X_2} + \epsilon \tag{1.84}$$

onde: Y representa a taxa de reacção; X_1 e X_2 são pressões parciais do reagente e do produto (respectivamente); θ_1 e θ_2 são constantes de absorção em equilíbrio para o reagente e o produto (respectivamente); e θ_3 é a constante relativa à taxa de reacção efectiva. Os dados referentes a 13 reacções químicas encontram-se na Tabela 1.11.

(a) Obtenha as estimativas de mínimos quadrados dos parâmetros $(\theta_1, \theta_2, \theta_3)$ recorrendo ao método de Newton-Raphson modificado com estimativa inicial

$$\underline{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}) = (3.0, 12.0, 0.7). \tag{1.85}$$

(b) Produza os gráficos que entender razoável para averiguar o ajustamento do modelo ao conjunto de dados. ■

1.3.2 Método de Levenberg-Marquardt

A motivação deste método passa pela constatação do seguinte facto: longe do ponto de mínimo a matriz hessiana poderá não ser definida

Table 1.11: Pressões parciais do reagente e do produto e taxa de reacção.

Reacção i	Pressão reagente $x_i(1)$	Pressão produto $x_i(2)$	Taxa reacção y_i
1	1.0	1.0	0.126
2	2.0	1.0	0.219
3	1.0	2.0	0.076
4	2.0	2.0	0.126
5	0.1	0.0	0.186
6	3.0	0.0	0.606
7	0.2	0.0	0.268
8	3.0	0.0	0.614
9	0.3	0.0	0.318
10	3.0	0.8	0.298
11	3.0	0.0	0.509
12	0.2	0.0	0.247
13	3.0	0.8	0.319

positiva. Uma possibilidade será perturbar a matriz hessiana por forma a que seja, ao longo das iterações, definida positiva. É deste modo que surge uma variante do método de Newton-Raphson, o **método de Levenberg-Marquardt** cuja iteração é:

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \left[\mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top}\mathbf{J}\left(\underline{\theta}^{(i)}\right) + \epsilon^{(i)}\,\mathbf{I}_p\right]^{-1} \times \mathbf{J}\left(\underline{\theta}^{(i)}\right)^{\top}\underline{H}\left(\underline{\theta}^{(i)}\right), \qquad (1.86)$$

onde $\mathbf{I}_p$ é a matriz identidade $(p \times p)$, e $\epsilon^{(i)} > 0$.

A constante $\epsilon^{(i)}$ deve ser escolhida de tal modo que a matriz, que substitui a hessiana, seja definida positiva. Recomenda-se que se comece por considerar um valor modesto para $\epsilon^{(0)}$ (e.g 0.001). Nas iterações seguintes a escolha deve ser feita com algum critério:

- se $h\left(\underline{\theta}^{(i+1)}\right) \geq h\left(\underline{\theta}^{(i)}\right)$ repete-se a iteração considerando um

valor superior para $\epsilon^{(i)}$ (e.g. 10 vezes superior ao tomado anterior nessa mesma iteração);

- se $h\left(\underline{\theta}^{(i+1)}\right) < h\left(\underline{\theta}^{(i)}\right)$ deve passar-se à iteração seguinte considerando $\epsilon^{(i+1)} < \epsilon^{(i)}$ (e.g. $\epsilon^{(i+1)} = \epsilon^{(i)}/10$).

Exercício 1.26 — Os tempos dos vencedores das finais olímpicas dos 100, 200, 400 e 800 metros de 1900 a 1976 encontram-se na seguinte tabela:

Table 1.12: Tempos (em s) de finais olímpicas dos 100, 200, 400 e 800 metros.

Ano	$100m$	$200m$	$400m$	$800m$
1900	10.80	22.20	49.40	121.40
1904	11.00	21.60	49.20	116.00
1908	10.80	22.60	50.00	112.80
1912	10.80	21.70	44.20	111.90
1920	10.80	22.00	49.60	113.40
1924	10.60	21.60	47.60	112.40
1928	10.80	21.80	47.80	111.80
1932	10.30	21.20	46.20	109.80
1936	10.30	20.70	46.50	112.90
1948	10.30	21.10	46.20	109.20
1952	10.40	20.70	45.90	109.20
1956	10.50	20.60	46.70	107.70
1960	10.20	20.50	44.90	106.30
1964	10.00	20.30	45.10	105.10
1968	9.90	19.80	43.80	104.30
1972	10.14	20.00	44.66	105.90
1976	10.06	20.23	44.26	103.50

Em 1982, Chatterjee e Chatterjee sugeriram o seguinte modelo de regressão não linear sugerido para quaisquer dos tempos:

$$t_i = \theta_1 + \theta_2 e^{-\theta_3 \times i},\ \theta_2, \theta_3 > 0. \tag{1.87}$$

Obtenha as estimativas de mínimos quadrados dos três parâmetros do modelo considerando as estimativas iniciais da Tabela 5.4 da p. 47 de Everitt (1987). Comente os resultados. (Everitt (1987, pp. 46–47).) ■

Textos de apoio (secção): Everitt (1987, pp. 42–48) e Nocedal e Wright (1999, pp. 250–273).

1.4 Introdução à optimização restringida

Até ao momento só foram considerados problemas de optimização (numérica) em que os parâmetros da função a optimizar não estão sujeitos a qualquer tipo de restrição. É altura de considerar o caso da **optimização restringida**.

Não se tratará de um exercício estéril mas sim de ir ao encontro de uma **necessidade premente em Estatística**.

Basta pensar no problema da estimação de **máxima verosimilhança** num contexto **Multinomial**. Com efeito, o vector de frequências absolutas (aleatórias) de p classes possíveis para um conjunto de n observações é um vector aleatório

$$\underline{N} \sim Multinomial_{p-1}(n, \underline{\theta})$$

que depende de um vector de probabilidades com dimensão p, $\underline{\theta} = (\theta_1, \ldots, \theta_{p-1}, \theta_p)$, sujeito à **restrição**

$$\sum_{i=1}^{p} \theta_i = 1.$$

É necessário estimar somente $k-1$ parâmetros já que $\theta_p = 1 - \sum_{i=1}^{p-1} \theta_i$ (daí o índice $p-1$ na notação) mas tendo sempre bem presente a restrição.

O problema acabado de descrever resolve-se recorrendo ao bem conhecido método dos **Multiplicadores de Lagrange** que não é adequado à resolução de alguns problemas de optimização restringida. Aliás, não existe um método geral que permita obter o ponto de mínimo de uma função sujeita a uma restrição e a optimização restringida é de longe mais sofisticada que a optimização irrestrita.

A formulação geral destes problemas pode ser feita do seguinte modo. Trata-se da obtenção do **ponto de mínimo**

$$\arg\min_{\underline{\theta}\in\mathbb{R}^p} f(\underline{\theta}) \tag{1.88}$$

sujeito ao **conjunto de restrições**

$$\begin{cases} c_i(\underline{\theta}) = 0,\ i \in \mathcal{E} \\ c_i(\underline{\theta}) \geq 0,\ i \in \mathcal{I}. \end{cases} \tag{1.89}$$

Tal como anteriormente $f(\underline{\theta})$ é a função objectivo, ao passo que as restrições se dividem em dois grupos:

- $\#\mathcal{E}$ restrições envolvendo igualdades; e
- $\#\mathcal{I}$ restrições envolvendo desigualdades.

Ao considerar-se o conjunto

$$\Omega = \{\underline{\theta} \in \mathbb{R}^p : c_i(\underline{\theta}) = 0,\ i \in \mathcal{E}; c_i(\underline{\theta}) \geq 0,\ i \in \mathcal{I}\} \tag{1.90}$$

o problema de optimização reduz-se simplemente à obtenção de

$$\arg\min_{\underline{\theta}\in\Omega} f(\underline{\theta}). \tag{1.91}$$

Podia julgar-se à partida que o facto de se acrescentar restrições ao conjunto inicial de valores possíveis dos parâmetros $\mathbb{R}^p$ melhoraria o problema da destrinça entre ponto de mínimo global e pontos de mínimo locais uma vez que tal conjunto se reduz a Ω. No entanto, as restrições podem agravar o problema de optimização senão veja-se o exemplo da obtenção de

$$\arg\min_{\underline{\theta}\in\Omega} ||\underline{\theta})||^2. \tag{1.92}$$

onde

$$\Omega = \{\underline{\theta} \in \mathbb{R}^p : ||\underline{\theta}||^2 \geq 1\}. \tag{1.93}$$

Ora, sem a restrição ter-se-ia um único ponto de mínimo global $\underline{\theta} = \underline{0}$ ao passo que com a restrição passa a ter-se uma infinidade de soluções — todo e qualquer vector de $\mathbb{R}^p$ de norma unitária.

1.4.1 Método dos multiplicadores de Lagrange

Comece-se por considerar que a função objectivo $f(\underline{\theta})$ está sujeita a m ($m < p$) **restrições envolvendo somente igualdades** ($c_i(\underline{\theta}) = 0, i \in \mathcal{E}$), i.e., $\#\mathcal{E} = m$, $\#\mathcal{I} = 0$.
Então o problema de optimização passa por considerar m constantes e uma nova função objectivo e pela resolução de uma equação envolvendo o gradiente desta nova função com $p + m$ parâmetros:

- $\lambda_i, i = 1, \ldots, m$, também denominados de **multiplicadores de Lagrange** (um para cada restrição);
- $g(\underline{\theta}; \lambda_1, \ldots, \lambda_m) = f(\underline{\theta}) - \sum_{i=1}^{m} \lambda_i c_i(\underline{\theta})$, a **função lagrangeana**;
- $\nabla g(\underline{\theta}; \lambda_1, \ldots, \lambda_m) = \underline{0}$.

De facto se a função objectivo $f(\underline{\theta})$ admitir um extremo local, quando sujeita a m restrições, então existirão m constantes reais $\lambda_1, \ldots, \lambda_m$ tais que

$$\nabla f(\underline{\theta}) = \sum_{i=1}^{m} \lambda_i \nabla c_i(\underline{\theta}) = \underline{0} \tag{1.94}$$

em todos os pontos de extremo local. Este método é válido quando o número de restrições m for inferior ao número de parâmetros p e se nem todos os jacobianos das funções $c_i(\theta)$, com respeito a m dos parâmetros $(\theta_1, \ldots, \theta_p)$, forem nulos no ponto de extremo.
De um modo geral é a Equação (1.94) — que se traduz nas $p + m$ equações seguintes

$$\begin{cases} \frac{\partial}{\partial \theta_j} g(\underline{\theta}; \underline{\lambda})|_{\underline{\theta}=\underline{\hat{\theta}}; \underline{\lambda}=\underline{\hat{\lambda}}} = 0, \; j = 1, \ldots, p \\ \frac{\partial}{\partial \lambda_i} g(\underline{\theta}; \underline{\lambda}), |_{\underline{\theta}=\underline{\hat{\theta}}; \underline{\lambda}=\underline{\hat{\lambda}}} = 0, \; i \in \mathcal{E}, \end{cases} \tag{1.95}$$

— que necessita de resolução numérica para a obtenção do ponto de mínimo.

Exercício 1.27 — Prove que a expressão geral das estimativas de máxima verosimilhança de $\underline{\theta} = (\theta_1, \ldots, \theta_{p-1}, \theta_p)$, com base nos valores observados de um vector de frequências aleatórias

$$\underline{N} \sim Multinomial_{p-1}(n, \underline{\theta}), \tag{1.96}$$

i.e.,

$$\begin{aligned} P(\underline{N} = \underline{n}) &= P(N_1 = n_1, \ldots, N_{p-1} = n_{p-1}, N_p = n_p) \\ &= \frac{n!}{\prod_{i=1}^{p} n_i!} \prod_{i=1}^{p} \theta_i^{n_i}, \end{aligned} \tag{1.97}$$

é $\hat{\underline{\theta}} = (\frac{n_1}{n}, \ldots, \frac{n_{p-1}}{n}, \frac{n_p}{n})$. $\hat{\underline{\theta}}$ é pois o vector das frequências relativas (observadas). ■

Deve acrescentar-se que o exercício anterior não requer o recurso de nenhuma técnica de optimização numérica.

No entanto, caso as probabilidades θ_i dependessem de, e.g. dois parâmetros μ e σ^2, ser-se-ia imediatamente tentado a obter as frequências relativas observadas e a invocar a propriedade da invariância dos estimadores de máxima verosimilhança para obter $\hat{\mu}$ e $\hat{\sigma^2}$, bastando para isso resolver duas equações. Contudo não deve proceder-se deste modo mas sim considerar a função objectivo $f(\mu, \sigma^2)$ e só depois aplicar um procedimento de optimização numérica.

Exercício 1.28 — Com o objectivo de estudar o tempo até falha de certo equipamento electrónico (em milhares de horas), X, foram recolhidas e ordenadas 50 observações na Tabela 1.13.

Dada a natureza dos dados e alguns estudos prévios, suspeita-se que as observações sejam provenientes de um modelo *Pareto* com parâmetros α e β, i.e.,

$$F_X(x) = 1 - \frac{\alpha^\beta}{x^\beta},\ x \geq \alpha. \tag{1.98}$$

Table 1.13: Tempos até falha de equipamento electrónico.

2.001	2.007	2.017	2.026	2.036	2.075	2.077	2.082	2.101	2.137
2.156	2.161	2.181	2.196	2.214	2.227	2.320	2.367	2.424	2.443
2.444	2.449	2.478	2.520	2.579	2.581	2.598	2.637	2.691	2.715
2.720	2.825	2.863	2.867	3.016	3.176	3.360	3.413	3.567	3.721
3.727	3.769	3.803	4.329	4.420	4.795	6.009	6.281	6.784	8.305

(a) Prove que a estimativa de máxima verosimilhança de (α, β) é $(2.001, 2.822)$.

(b) Obtenha as frequências observadas absolutas das classes $[2.001, 2.1656]$, $(2.1656, 2.3981]$, $(2.3981, 2.7686]$, $(2.7686, 3.5394]$ e $(3.5394, +\infty)$ e prove que estas classes são equiprováveis sob a hipótese $X \sim Pareto(2.001, 2.822)$.

(c) Admita agora que não dispunha da amostra ordenada mas somente das frequências absolutas obtidas em (b). Reavalie a estimativa de máxima verosimilhança de (α, β).

(d) Obtenha a estimativa de máxima verosimilhança de (α, β) sujeita à restrição:

$$P[X \in (3.5394, +\infty) | X \sim Pareto(\alpha, \beta)] = 0.2. \tag{1.99}$$

■

Nas Secções 8.3, 8.9 e 8.10 de Khuri (1993) podem encontrar-se mais exemplos da aplicação do método de multiplicadores de Lagrange a Estatística na minimização de funções objectivo sujeitas a restrições envolvendo exclusivamente igualdades, no âmbito, nomeadamente, da metodologia de superfícies de resposta, da determinação de estimativas centradas com norma quadrática mínima e da obtenção de intervalos de Scheffé.

Em Nocedal e Wright (1999, pp. 321–327) pode encontrar-se a descrição do método dos multiplicadores de Lagrange aplicado a situações em que $\#\mathcal{I} > 0$, i.e., em que há pelo menos uma restrição envolvendo uma desigualdade.
Nos Capítulos 5 e 6 de Gill *et al.* (1981) é dado um tratamento completo às situações em que a minimização está sujeita a restrições lineares e não lineares, respectivamente, pelo que merecem uma leitura mais cuidada.

Termina-se este capítulo citando Robert Fletcher que descreve a optimização como uma "fascinante mistura de teoria e cálculo, heurísticas e rigor" (Nocedal e Wright (1999, p. x)) e, acrescente-se, de uma importância crucial em Estatística.

Texto de apoio: Nocedal e Wright (1999, pp. 314–357).

1.5 Referências

Alves, C.J.S. (2000). *Fundamentos de Análise Numérica (I) — Teoria e Exercícios*. Secção de Folhas — Instituto Superior Técnico.

Everitt, E.S. (1987). *Introduction to Optimization Methods and their Application in Statistics*. Chapman and Hall, Ltd. (QA278–279/1. EVE.36891)

Gill, P.E., Murray, W. e Wright, M.H. (1981). *Practical Optimization*. Academic Press, Inc. (06-13.4880.30580)

Nocedal, J. e Wright, S.J. (1999). *Numerical Optimization*. Springer-Verlag, Inc. (QA297.5.NOC.50578)

Khuri, A.I. (1993). *Advanced Calculus with Applications in Statistics*. John Wiley & Sons, Inc.