Capítulo 7

Aproximación de funciones y ajuste de datos experimentales

En este capítulo trataremos dos problemas íntimamente ligados. El primero es el problema de la aproximación de funciones que lo podemos enunciar como:

Dada una función $f(x)$ definida en $[a,b]$ y una serie de funciones base $\psi_r(x)$ definidas también en $[a,b]$, encontrar los coeficientes a_r de forma que la suma $\sum_{r=0}^{n} a_r \psi_r(x)$ sea lo más próxima posible a $f(x)$ en el intervalo $[a,b]$.

El concepto de proximidad lo definiremos más adelante. El problema de la aproximación es esencial cuando queremos representar una función en serie de otras más sencillas, como potencias o funciones trigonométricas.

El segundo problema surge cuando medimos datos que satisfacen una ley que se comporta como una función. Típicamente medimos un conjunto de N puntos (x_i, y_i), donde la variable independiente x_i se supone exacta y todo el error de medida de cada punto se atribuye a la variable dependiente y_i, que viene afectada de un error experimental σ_i. Suponemos que la ley que satisfacen los datos se puede describir mediante un modelo de la forma $y = f(x)$ que depende de una serie de parámetros a_i. Nos limitaremos al caso particular en que la dependencia de los parámetros es lineal, es decir $f(x) = \sum_{r=0}^{n} a_r \psi_r(x)$ donde $\psi_r(x)$ son funciones base convenientes para describir nuestro modelo teórico de los datos. Podemos enunciar el segundo problema como:

Determinar los valores de los parámetros a_i que hacen que la cantidad

$$\chi^2(a_0, a_1, \ldots a_n) = \sum_{i=1}^{N} \frac{(y_i - \sum_{r=0}^{n} a_r \psi_r(x_i))^2}{\sigma_i^2}$$

sea mínima.

Este es el problema del modelado de datos experimentales. Ambos problemas, aproximación de funciones y modelado de datos, están íntimamente ligados y comparten las mismas técnicas de resolución.

7.1. Proximidad de funciones: Distancias y Normas

En primer lugar, hay que definir el concepto de proximidad de dos funciones en un intervalo. Para ello hay que introducir una distancia entre las dos funciones. Las distancias se suelen definir mediante normas. Si tenemos una norma definida para funciones $\|f(x)\|$, se define la distancia entre dos funciones $f(x)$ y $g(x)$ como $d(f(x),g(x)) = \|f(x) - g(x)\|$. Hay diversas normas utilizadas frecuentemente. La más utilizada es la norma de mínimos cuadrados o L_2 definida como

$$\|f(x) - g(x)\|_2 = \int_a^b (f(x) - g(x))^2 \, dx$$

en un intervalo y como

$$\|f(x) - g(x)\|_2 = \sum_{i=0}^{n} (f(x_i) - g(x_i))^2$$

sobre un conjunto discreto de puntos. En general la norma L_p se define como

$$\|f(x) - g(x)\|_p = \int_a^b |f(x) - g(x)|^p \, dx$$

sobre un intervalo y como

$$\|f(x) - g(x)\|_p = \sum_{i=1}^{N} |f(x_i) - g(x_i)|^p$$

sobre un conjunto discreto de puntos. En aproximación de funciones, aparte de la norma L_2, se utilizan usualmente la norma L_1 y la llamada norma L_∞, definida como

$$\|f(x) - g(x)\|_\infty = \text{máx}\, |f(x) - g(x)|$$

sobre un intervalo o conjunto discreto de puntos. La aproximación de funciones que minimiza la norma L_∞ se conoce como aproximación minimax. Cuando deseamos una aproximación a una función en un intervalo por otra más sencilla, la aproximación minimax es quizás la más razonable, ya que limita el error máximo cometido en un punto arbitrario del intervalo. Sin embargo, cuando tenemos puntos experimentales afectados de un error estadístico, entonces la aproximación de mínimos cuadrados, en la versión de mínimo χ^2, es la única justificada desde el punto de vista estadístico.

7.2. Aproximación de mínimos cuadrados

7.2.1. Normas a partir de productos escalares

Si definimos el producto escalar de dos funciones como

$$< f(x)|g(x) >= \int_a^b f(x)g(x)dx$$

sobre un intervalo y

$$< f(x)|g(x) >= \sum_{i=1}^{N} f(x_i)g(x_i)$$

sobre un conjunto discreto de puntos. La norma L_2 se puede escribir en función del producto escalar como

$$\|f(x)-g(x)\|_2 =< f(x)-g(x)|f(x)-g(x) >$$

tanto sobre un intervalo como un conjunto discreto de puntos.

7.2.2. Las ecuaciones normales de mínimos cuadrados

En general deseamos aproximar una función $f(x)$ por una combinación lineal de un conjunto de $n+1$ funciones base $\psi_r(x)$

$$f(x) = \sum_{r=0}^{n} a_r \psi_r(x)$$

El caso más frecuente es cuando $\psi_r(x) = x^r$, que se denomina aproximación polinómica. Para lleva a cabo la aproximación tenemos que encontrar los coeficientes a_0, $a_1, \ldots, a_n$ que hacen la función

$$E(a_0, a_1, \ldots, a_n) = \left\| f(x) - \sum_{r=0}^{n} a_r \psi_r(x) \right\|$$

mínimo. Tenemos que minimizar E considerada como una función de los parámetros a_r,

$$\begin{aligned} E(a_0, a_1, \ldots, a_n) &= < f(x) - \sum_{r=0}^{n} a_r \psi_r(x) | f(x) - \sum_{r=0}^{n} a_r \psi_r(x) >= \\ & < f(x)|f(x) > -2 \sum_{r=0}^{n} a_r < f(x)|\psi_r(x) > + \sum_{r,s=0}^{n} a_r a_s < \psi_s(x)|\psi_r(x) > \end{aligned}$$

Las condiciones que se deben de cumplir para que exista un mínimo son, en primer lugar, la anulación de las derivadas primeras con respecto de los parámetros, y en segundo lugar que la matriz de derivadas segundas o Hessiano sea definida positiva

$$\begin{aligned} \frac{\partial E(a_0, a_1, \ldots, a_n)}{\partial a_i} &= 0 \\ \left| \frac{\partial^2 E(a_0, a_1, \ldots, a_n)}{\partial a_i \partial a_j} \right| &> 0 \end{aligned}$$

La primera de las condiciones da

$$\frac{\partial E(a_0, a_1, \ldots, a_n)}{\partial a_i} = -2 < f(x)|\psi_i(x) > +2 \sum_{r=0}^{n} a_r < \psi_r(x)|\psi_i(x) >= 0$$

Esta condición implica el cumplimiento de un sistema de ecuaciones

$$\sum_{r=0}^{n} a_r < \psi_r(x)|\psi_i(x) >=< f(x)|\psi_i(x) > \tag{7.1}$$

que se conocen como ecuaciones normales. Constituyen un sistema lineal para los parámetros

$$Aa = b$$

donde a es el vector de parámetros, b el vector de términos independientes y A la matriz de coeficientes. La segunda condición se cumple siempre, lo que se puede ver explícitamente suponiendo que variamos los parámetros $a_r \to a_r + \delta a_r$ y calculamos la diferencia

$$E(a_0+\delta a_0, a_1+\delta a_1, \ldots, a_n+\delta a_n) - E(a_0, a_1, \ldots, a_n) =$$

$$< f(x) - \sum_{r=0}^{n} (a_r+\delta a_r)\psi_r(x) | f(x) - \sum_{r=0}^{n} (a_r+\delta a_r)\psi_r(x) >$$
$$- < f(x) - \sum_{r=0}^{n} a_r\psi_r(x) | f(x) - \sum_{r=0}^{n} a_r\psi_r(x) > \quad =$$

$$= -2\sum_{r=0}^{n} \delta a_r < \psi_r(x)|f(x) - \sum_{s=0}^{n} a_s\psi_s(x) > + < \sum_{r=0}^{n} \delta a_r\psi_r(x)| \sum_{r=0}^{n} \delta a_r\psi_s(x) >$$

El primer término se anula por el cumplimiento de las ecuaciones normales y el segundo es estrictamente positivo, puesto que es la norma de un vector no nulo.

El caso más simple es cuando tenemos únicamente dos funciones base ψ_0 y ψ_1. Entonces las ecuaciones normales quedan como

$$\begin{aligned} a_0 &< \psi_0|\psi_0 > + a_1 < \psi_0|\psi_1 >=< \psi_0|f > \\ a_0 &< \psi_1|\psi_0 > + a_1 < \psi_1|\psi_1 >=< \psi_1|f > \end{aligned}$$

cuyas soluciones, aplicando la fórmula de Cramer son

$$a_0 = \frac{\begin{vmatrix} < \psi_0|f > & < \psi_1|\psi_0 > \\ < \psi_1|f > & < \psi_1|\psi_1 > \end{vmatrix}}{\begin{vmatrix} < \psi_0|\psi_0 > & < \psi_0|\psi_1 > \\ < \psi_1|\psi_0 > & < \psi_1|\psi_1 > \end{vmatrix}}$$

$$a_1 = \frac{\begin{vmatrix} < \psi_0|\psi_0 > & < \psi_0|f > \\ < \psi_1|\psi_0 > & < \psi_1|f > \end{vmatrix}}{\begin{vmatrix} < \psi_0|\psi_0 > & < \psi_0|\psi_1 > \\ < \psi_1|\psi_0 > & < \psi_1|\psi_1 > \end{vmatrix}}$$

Si consideramos el caso del ajuste lineal, $\psi_0 = 1$ y $\psi_1 = x$, en el caso de un conjunto discreto de puntos tenemos

$$< \psi_0|\psi_0 >= \sum_{i=1}^{N} 1 = N, \quad < \psi_0|\psi_1 >= \sum_{i=1}^{N} x_i, \quad < \psi_1|\psi_1 >= \sum_{i=1}^{N} x_i^2,$$
$$< \psi_0|f >= \sum_{i=1}^{N} f(x_i) \quad < \psi_1|f >= \sum_{i=1}^{N} x_i f(x_i)$$

Poniendo $y_i = f(x_i)$ tenemos las fórmulas usuales del ajuste de un conjunto de puntos por mínimos cuadrados:

$$a_0 = \frac{\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \qquad a_1 = \frac{\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2 - N \sum_{i=1}^{N} x_i y_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

En el caso de aproximaciones polinómicas de orden más elevado (parabólicas, cúbicas, o combinaciones lineales de varias potencias distintas) procederíamos de forma análoga, resolviendo las ecuaciones por uno de los métodos vistos en el capítulo 4, en vez de por la regla de Cramer. Podemos pensar que podemos continuar de esta forma hasta cualquier orden de aproximación aunque este no es el caso. De hecho para más de 10 funciones, las ecuaciones normales están mal condicionadas, y dan resultados imprecisos con doble precisión. Para orden 100, incluso con cuádruple precisión en procesadores de 64 bits se obtienen resultados muy imprecisos. Sin embargo no es raro que sea necesario aproximar una función por varios centenares de funciones base. Esto ocurre por ejemplo cuando se descompone una onda sonora en armónicos o cuando se estudian imágenes. Si obtenemos una solución imprecisa de las ecuaciones normales los agudos de una onda serían incorrectos y la imagen no sería nítida. Por ello hace falta un método eficaz de evitar el mal condicionamiento. Ello se consigue con funciones ortogonales. Decimos que las funciones ψ_r son ortogonales si

$$< \psi_r|\psi_s >= n_r \delta_{rs}$$

donde n_r es la normalización de la función y δ_{ij} es la delta de Kronecker. En este caso las ecuaciones normales se simplifican a

$$a_r < \psi_r|\psi_r >=< \psi_r|f >$$

con la solución

$$a_r = \frac{< \psi_r|f >}{< \psi_r|\psi_r >}$$

La utilización de funciones ortogonales tiene dos ventajas: la primera es que desaparece el mal condicionamiento, y la segunda es que cada coeficiente es independiente de los demás. Por lo tanto, si deseamos extender la aproximación a un orden superior, los coeficientes ya calculados no varían, por lo se dice que tienen la *propiedad de permanencia*. Esta independencia es muy importante en el caso de datos experimentales, puesto que implica que los distitos coeficientes obtenidos ajustando mediante funciones ortogonales no estan correlacionados esdadísticamente.

7.2.3. Series de Fourier

Sin duda alguna, las funciones ortogonales más utilizadas son las funciones trigonométricas $\sin(x)$ y $\cos(x)$. El conjunto de funciones $\{1,\cos(x),\sin(x),\cos(2x),\ldots\}$ son ortogonales en el intervalo $[-\pi,\pi]$ con las relaciones de ortogonalidad

$$\int_{-\pi}^{\pi} dx\cos kx\cos mx = \int_{-\pi}^{\pi} dx\cos kx\sin mx = \int_{-\pi}^{\pi} dx\sin kx\sin mx = 0 \quad m\neq k$$

$$\int_{-\pi}^{\pi} dx\cos kx = \int_{-\pi}^{\pi} dx\sin kx = 0 \quad k>0$$

$$\int_{-\pi}^{\pi} dx(\cos kx)^2 = \int_{-\pi}^{\pi} dx(\sin kx)^2 = \pi \qquad \int_{-\pi}^{\pi} dx = 2\pi$$

El desarrollo de una función como

$$f(x)\sim\frac{a_0}{2}+\sum_{r=1}^{\infty}(a_r\cos rx+b_r\sin rx)$$

se conoce como serie de Fourier. Converge en la norma de mínimos cuadrados siempre que la función sea periódica en $[-\pi,\pi]$ y continua. Cuando la serie se trunca a un número finito de términos, frecuentemente grande, tenemos la aproximación de Fourier. Los coeficientes vienen dados por

$$a_0=\frac{1}{\pi}\int_{-\pi}^{\pi}dxf(x) \qquad a_r=\frac{1}{\pi}\int_{-\pi}^{\pi}dxf(x)\cos rx \qquad b_r=\frac{1}{\pi}\int_{-\pi}^{\pi}dxf(x)\sin rx$$

En casos analíticamente sencillos los coeficientes de Fourier se calculan fácilmente. Consideremos por ejemplo una onda cuadrada, que se utiliza frecuentemente en electrónica.

$$f(x)=\begin{cases} -1 & -\pi\leq x<0 \\ 1 & 0\leq x<\pi \end{cases}$$

Esta función es una función impar. También es discontinua, pero a pesar de esto la serie de Fourier converge. Como $\cos x$ es par, los coeficientes a_r se anulan. Los coeficientes b_r vienen dados por

$$b_r=\frac{1}{\pi}\int_{-\pi}^{\pi}dxf(x)\sin rx=\frac{2}{\pi}\int_0^{\pi}dx\sin rx=\frac{2}{\pi}\cos rx\Big|_0^{\pi}=\begin{cases} 0 & r\ par \\ \dfrac{4}{\pi r} & r\ impar \end{cases}$$

$$f(x)\sim\frac{2}{\pi}\sum_{r=0}^{\infty}\frac{\sin[(2r+1)x]}{2r+1}$$

En el caso de una función periódica de período T, el desarrollo toma la forma

$$f(x)\sim\frac{a_0}{2}+\sum_{r=1}^{\infty}(a_r\cos\frac{2\pi rt}{T}+b_r\sin\frac{2\pi rt}{T})$$

con

$$a_0=\frac{2}{T}\int_{-T/2}^{T/2}dtf(t) \qquad a_r=\frac{2}{T}\int_{-T/2}^{T/2}dtf(t)\cos\frac{2\pi rt}{T} \qquad b_r=\frac{2}{T}\int_{-T/2}^{T/2}dtf(t)\sin\frac{2\pi rt}{T} \quad (7.2)$$

Serie de Fourier discreta

Las funciones trigonométricas también son ortogonales sobre un conjunto finito de puntos. Dada una función $f(t)$ periódica con período T, si tomamos un conjunto de $N+1$ puntos igualmente espaciados entre 0 y T ($t_s = sT/(N+1)$, $s = 0,\ldots,N$) se satisfacen las siguientes relaciones de ortogonalidad

$$\begin{aligned}
< \sin\frac{2\pi kt}{T} | \sin\frac{2\pi mt}{T} >= \sum_{s=0}^{N} \sin\frac{2\pi ks}{N+1}\sin\frac{2\pi ms}{N+1} &= \begin{cases} 0 & k \neq m, k = m = 0, N+1 \\ \frac{N+1}{2} & k = m \neq 0, N+1 \end{cases} \\
< \sin\frac{2\pi kt}{T} | \cos\frac{2\pi mt}{T} >= \sum_{s=0}^{N} \sin\frac{2\pi ks}{N+1}\cos\frac{2\pi ms}{N+1} &= 0 \\
< \cos\frac{2\pi kt}{T} | \cos\frac{2\pi mt}{T} >= \sum_{s=0}^{N} \cos\frac{2\pi ks}{N+1}\cos\frac{2\pi ms}{N+1} &= \begin{cases} 0 & k \neq m \\ \frac{N+1}{2} & k = m \neq 0, N+1 \\ N+1 & k = m = 0, N+1 \end{cases}
\end{aligned}$$

El desarrollo

$$f(t) \sim \frac{a_0}{2} + \sum_{k=1}^{n} \left(a_k \cos\frac{2\pi ks}{N+1} + b_k \sin\frac{2\pi ks}{N+1} \right)$$

converge a $f(t)$ sobre el conjunto de $N+1$ puntos en el sentido de mínimos cuadrados. Cuanto tomamos $N+1$ coeficientes, el desarrollo interpola a la función $f(t)$ en el conjunto de $N+1$ puntos. Si N es par (número de puntos impar), la función interpoladora es

$$F_{N+1}\left(\frac{sT}{N+1}\right) = \frac{a_0}{2} + \sum_{k=1}^{N/2} \left(a_k \cos\frac{2\pi ks}{N+1} + b_k \sin\frac{2\pi ks}{N+1} \right)$$

mientras que si N es impar (número par de puntos)

$$F_{N+1}(\frac{sT}{N+1}) = \frac{a_0}{2} + \sum_{k=1}^{(N-1)/2} \left(a_k \cos\frac{2\pi ks}{N+1} + b_k \sin\frac{2\pi ks}{N+1} \right) + \frac{a_{(N+1)/2}}{2}\cos\pi s$$

Los coeficientes del desarrollo vienen dados por

$$\begin{aligned}
a_k &= \frac{2}{N+1} \sum_{s=0}^{N} f\left(\frac{sT}{N+1}\right) \cos\frac{2\pi ks}{N+1} \\
b_k &= \frac{2}{N+1} \sum_{s=0}^{N} f\left(\frac{sT}{N+1}\right) \sin\frac{2\pi ks}{N+1}
\end{aligned}$$

Es interesante notar que a_k y b_k vienen dados por la evaluación numérica mediante la regla trapezoidal para $N+1$ intervalos ($N+2$ puntos, ampliando con el extremo del $t = T$) de las integrales de las ecuaciones 7.2, notando que $f(0) = f(T)$, $T = (N+1)h$, y que los senos se

anulan en los extremos del intervalo:

$$\begin{aligned} a_k &= \frac{2}{N+1}\left[\frac{f(0)+f(T)}{2}+\sum_{s=1}^{N} f\left(\frac{sT}{N+1}\right)\cos\frac{2\pi ks}{N+1}\right] \\ b_k &= \frac{2}{N+1}\left[\sum_{s=1}^{N} f\left(\frac{sT}{N+1}\right)\sin\frac{2\pi ks}{N+1}\right] \end{aligned}$$

7.3. Polinomios ortogonales

El conjunto más sencillo de funciones ortogonales son los polinomios. Se pueden definir sobre un conjunto discreto de puntos o sobre un intervalo continuo. Vamos a definirlos por ahora con coeficiente de la potencia más elevada igual a la unidad. De esta forma siempre existe una relación de recurrencia del tipo $(p_{k+1}(x) - xp_k(x))$

$$p_{k+1}(x) = xp_k(x) + \sum_{s=0}^{k} c_s^{k+1} p_s(x) \tag{7.3}$$

ya que $(p_{k+1}(x) - xp_k(x))$ es un polinomio de grado k, y por lo tanto siempre se puede expresar como combinación lineal de $p_0(x), \ldots, p_k(x)$. Vamos a suponer únicamente la existencia de un producto escalar sobre un intervalo $[a,b]$ o sobre un conjunto discreto de $N+1$ puntos. Dicho producto escalar los supondremos de la forma más general con una función peso $w(x)$ en el caso continuo y un conjunto de pesos w_s en el caso discreto

$$< p_k(x)|p_j(x) >= \left\{ \begin{array}{l} \int_a^b dx w(x) p_k(x) p_j(x) \\ \sum_{s=1}^{N} w_s p_k(x_s) p_j(x_s) \end{array} \right\}$$

Tenemos que determinar los coeficientes c_s^{k+1}. Para ello multiplicamos escalarmente la ec. 7.3 por un polinomio dado $p_r(x)$, $r \leq k$,

$$< p_r|p_{k+1} >= 0 =< p_r|xp_k > + \sum_{s=0}^{k} c_s^{k+1} < p_r|p_s >=< p_r|xp_k > + c_r^{k+1} < p_r|p_r >$$

de donde

$$c_r^{k+1} = -\frac{< p_r|xp_k >}{< p_r|p_r >}$$

Como $< p_r|xp_k >=< p_r x|p_k >$ y $xp_r(x)$ es un polinomio de grado $r+1$, que se puede expresar como una combinación lineal de $p_0, \ldots, p_{r+1}$, $< p_r|xp_k >= 0$ para $r = 0, 1, \ldots, k-2$. Por lo tanto sólo c_{k-1}^{k+1} y c_k^{k+1} pueden ser distintos de 0. Vienen dados por

$$c_{k-1}^{k+1} = -\frac{< p_{k-1}|xp_k >}{< p_{k-1}|p_{k-1} >}$$

y

$$c_k^{k+1} = -\frac{< p_k|xp_k >}{< p_k|p_k >}$$

Los polinomios ortogonales satisfacen por lo tanto la relación de recurrencia

$$p_{k+1}(x) = (x + c_k^{k+1})p_k(x) + c_{k-1}^{k+1}p_{k-1}(x)$$

Para que esta relación se cumpla también para $p_1(x)$ se define $p_{-1}(x) = 0$. Para obtener el ajuste por mínimos cuadrados de una función dada $f(x)$, sólo tenemos que calcular los coeficientes c_k^{k+1} y c_{k-1}^{k+1} mediante las ecuaciones anteriores para obtener los polinomios necesarios mediante la relación de recurrencia. El ajuste de mínimos cuadrados de orden n viene dado por

$$\sum_{r=0}^{n} a_r p_r(x)$$

donde a_r se obtiene de

$$a_r = \frac{< f|p_r >}{< p_r|p_r >}$$

El incremento del orden de aproximación en una unidad implica, por lo tanto, el cálculo de un nuevo polinomio y un coeficiente, lo que equivale a realizar 6 productos escalares, que se reducen a 4 dado las constantes de normalización de los polinomios $< p_r|p_r >$ se han calculado durante la obtención del coeficiente previo. Esta es la forma más eficiente de ajustar datos mediante polinomios de orden elevado, tanto para datos discretos como continuos, pues se evitan errores debidos al mal condicionamiento de las ecuaciones normales, y por otro lado el esfuerzo numérico es menor, y se puede elevar el orden aprovechando los cálculos realizados para un orden inferior. En el caso de datos discretos, el único inconveniente es la dependencia de los polinomios del conjunto de puntos, lo cual no es importante, pues la suma de polinomios ortogonales se puede expresar de forma inmediata como un polinomio ordinario.

Para datos definidos en intervalos continuos hay polinomios ortogonales bien conocidos para diversos pesos e intervalos, algunos de los cuales se dan en la tabla 7.1

Tabla 7.1: Principales polinomios ortogonales

Nombre	Peso	Intervalo	Símbolo
Legendre	1	$[-1,1]$	$P_n(x)$
Hermite	$\exp(-x)$	$[-\infty,\infty]$	$H_n(x)$
Laguerre	$\exp(-x^2)$	$[0,\infty]$	$L_n(x)$
Chebychev	$1/\sqrt{1-x^2}$	$[-1,1]$	$T_n(x)$
Chebychev 2ª especie	$\sqrt{1-x^2}$	$[-1,1]$	$U_n(x)$

Si la función f se conoce analíticamente o se puede calcular con facilidad en cualquier punto que se desee, los coeficientes del desarrollo de la función en serie de polinomios ortogonales se pueden calcular por cualquiera de los métodos de integración vistos en el capítulo anterior.

7.3.1. Serie de Chebychev discreta

Otro conjunto de funciones que satisfacen relaciones de ortogonalidad sobre un conjunto discreto de puntos son los polinomios de Chebychev.

7.4. Aproximación minimax

7.5. Aproximación por funciones racionales

7.6. Modelado de datos experimentales

7.6.1. Variables aleatorias, valores esperados y varianzas

Una variable aleatoria es una variable que puede tomar un conjunto de valores (continuo o discreto) y que cada valor aparece con una probabilidad determinada. Por ejemplo el valor de la cara de un dado puede tomar 6 valores con probabilidad $1/6$. El número de desintegraciones de una muestra radioactiva en la unidad de tiempo toma valores enteros. La variable puede tomar valores continuos, en cuyo caso existe una distribución de probabilidad o densidad de probabilidad $p(x)$, definida en $[-\infty,+\infty]$. La probabilidad de que x tome un valor comprendido entre dos valores a y b viene dada por

$$P(a<x<b)=\int_a^b p(x)dx$$

Se define el valor esperado de x, $E[x]$, también denominado valor medio, como

$$E[x]=\bar{x}=\int_{-\infty}^{\infty} xp(x)dx$$

y la varianza $\sigma^2(x)$ como

$$\sigma^2(x)=E[(x-\bar{x})^2]=\int_{-\infty}^{\infty}(x-\bar{x})^2p(x)dx$$

Frecuentemente tenemos varias variables aleatorias que pueden aparecer simultáneamente. En este caso tenemos una distribución de probabilidad conjunta $p(x_1,x_2,...,x_n)$. Si tenemos dos variables aleatorias x_1 y x_2, se define la covarianza $\sigma(x_1,x_2)$ como

$$\sigma(x_1,x_2)=E[(x_1-\bar{x}_1)(x_2-\bar{x}_2]=\int_{-\infty}^{\infty}(x_1-\bar{x}_1)(x_2-\bar{x_2})p(x_1,x_2)dx_1dx_2$$

Si dos variables son independientes, su covarianza se anula, ya que en este caso $p(x_1,x_2)=p(x_1)p(x_2)$ y la integral anterior se descompone en el producto de dos integrales que se anulan, lo cual se demuestra fácilmente teniendo en cuenta la definición del valor medio.

Los datos experimentales se comportan como variables aleatorias. Cada vez que medimos una magnitud física con suficiente precisión obtenemos un valor distinto. El conjunto de valores de una serie de medidas se distribuye con una función de distribución de probabilidad. Una serie de medidas x_i se caracteriza por su valor medio $\bar{x}$ y su desviación típica σ_x.

7.6.2. Comportamiento estadístico de los datos experimentales

Un caso particularmente importante es cuando deseamos ajustar datos experimentales mediante una función dependiente de parámetros ajustables. Esta función puede estar inspirada en un modelo teórico, o bien puede ser de carácter empírico, motivada únicamente por el comportamiento de los datos.

Los datos experimentales vienen siempre afectados de errores de medida. Estos errores pueden ser sistemáticos o aleatorios. Los errores sistemáticos son debidos al sistema o aparato de medida y generalmente sólo actúan en una dirección. Tienen un número reducido de causas y se pueden determinar frecuentemente a partir del análisis del método de medida, comparando con otras medidas conocidas, o mediante un procedimiento de calibrado. Un ejemplo de error sistemático es el error de la medida de una longitud con una regla debido a la variación de la longitud de la regla con la temperatura. La corrección de este error se consigue conociendo el coeficiente de dilatación térmica de la regla con la temperatura (análisis del método de medida) o comparando la longitud medida con una longitud conocida. Los errores aleatorios por otro lado tienen un número muy elevado de causas, difíciles de identificar por separado, y que producen una contribución aleatoria en cada medida independiente. Cada una de las causas produce una pequeña contribución y el error aleatorio total es la suma de todas las causas por separado. El error aleatorio se puede representar matemáticamente por una suma de variables aleatorias.

El teorema del límite central establece que una suma de variables aleatorias independientes con distribuciones arbitrarias tiende a la distribución normal. En términos matemáticos:

Si $x_1, x_2, x_3, \ldots$ es una sucesión de variables aleatorias independientes, con distribuciones de probabilidad arbitrarias con medias μ_i y desviaciones típicas σ_i, y formamos la nueva sucesión de variables aleatorias y_k definidas por

$$y_k = \frac{\sum_{i=1}^{k}(x_i - \mu_i)}{\left(\sum_{i=1}^{k}\sigma_i^2\right)^{1/2}}$$

la función de distribución de y_k tiende a una distribución normal con media 0 y desviación típica 1 cuando k tiende a ∞.

La distribución de probabilidad de una distribución normal de media μ y desviación típica σ viene dada por

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp -\frac{(x-\mu)^2}{2\sigma^2}$$

7.6.3. Principio de máxima verosimilitud

El principio de máxima verosimilitud establece que si obtenemos los valores $x_1, x_2, \ldots, x_n$ en N medidas de una variable aleatoria x, ese conjunto de valores tenía una probabilidad máxima de ocurrir. Vamos a ver como podemos utilizar este principio para obtener parámetros de distribuciones. La probabilidad de obtener el anterior conjunto de medidas la podemos escribir, en el caso de que la variable x satisface la distribución normal, como

$$P(x_1, x_2, \ldots, x_N) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp -\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2\sigma^2}$$

Si esta probabilidad es máxima, los parámetros μ y σ deben ser tales que se satisfagan las ecuaciones

$$\begin{aligned}\frac{\partial P(x_1,x_2,\ldots,x_N)}{\partial \mu} &= 0\\ \frac{\partial P(x_1,x_2,\ldots,x_N)}{\partial \sigma} &= 0\end{aligned}$$

La primera de las ecuaciones da

$$-2\sum_{i=1}^{N}(x_i-\mu) = 0$$

y por lo tanto la solución es

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

La segunda ecuación queda como

$$-N\sigma^{-(N+1)}\exp-\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}+\sigma^{-N}\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{\sigma^3}\exp-\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}=0$$

$$-\frac{N}{\sigma}+\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{\sigma^3}=0$$

dando como solución

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

que son los estimadores usuales de la media y desviación típica.

Vamos a aplicar ahora el principio de máxima verosimilitud a un conjunto de datos experimentales que satisfacen una ley que depende de *n* parámetros:

$$y = f(x,a_0,\ldots,a_n)$$

Si medimos *N* puntos (x_i,y_i) entonces el principio de máxima verosimilitud establece que

$$P(x_1,x_2,\ldots,x_N) = \frac{1}{(2\pi)^{N/2}\sigma_1\sigma_2\cdots\sigma_N}\exp-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i-f(x_i,a_0,\ldots,a_n)}{\sigma_i}\right)^2$$

es máximo lo que implica que el término

$$\chi^2(a_0,a_1,\ldots a_n) = \sum_{i=1}^{N}\left(\frac{y_i-f(x_i,a_0,\ldots,a_n)}{\sigma_i}\right)^2$$

es mínimo. En el caso de una función lineal de los parámetros

$$f(x_i,a_0,\ldots,a_n) = \sum_{r=0}^{n} a_r\psi_r(x_i)$$

obtenemos, derivando con respecto de los parámetros de forma análoga al caso de un conjunto de puntos, las ecuaciones normales

$$\sum_{r=0}^{n} a_r \sum_{i=1}^{N} \frac{\psi_r(x_i)\psi_s(x_i)}{\sigma_i^2} = \sum_{i=1}^{N} \frac{y_i \psi_s(x_i)}{\sigma_i^2}$$

que se pueden poner en la forma 7.1 con la definición de producto escalar

$$< \psi_r | \psi_s >= \sum_{i=1}^{N} \frac{\psi_r(x_i)\psi_s(x_i)}{\sigma_i^2}$$

Vemos que la condición de que χ^2 sea mínimo implica unas ecuaciones normales con un producto escalar cuyos pesos son los inversos de las varianzas de los errores de los puntos. El producto escalar con pesos se puede escribir como

$$< \psi_r | \psi_s >= \begin{pmatrix} \psi_r(x_1) & \psi_r(x_2) & \cdots & \psi_r(x_N) \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1{}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2{}^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_N{}^2} \end{pmatrix} \begin{pmatrix} \psi_r(x_1) \\ \psi_r(x_2) \\ \vdots \\ \psi_r(x_N) \end{pmatrix}$$

con lo que las ecuaciones normales se pueden escribir como

$$\begin{pmatrix} \psi_0(x_1) & \psi_0(x_2) & \cdots & \psi_0(x_N) \\ \psi_1(x_1) & \psi_1(x_2) & \cdots & \psi_1(x_N) \\ \vdots & \vdots & \vdots & \vdots \\ \psi_n(x_1) & \psi_n(x_2) & \cdots & \psi_n(x_N) \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1{}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2{}^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_N{}^2} \end{pmatrix} \begin{pmatrix} \psi_0(x_1) & \psi_1(x_1) & \cdots & \psi_n(x_1) \\ \psi_0(x_2) & \psi_1(x_2) & \cdots & \psi_n(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \psi_0(x_N) & \psi_1(x_N) & \cdots & \psi_n(x_N) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}$$

$$= \begin{pmatrix} \psi_0(x_1) & \psi_0(x_2) & \cdots & \psi_0(x_N) \\ \psi_1(x_1) & \psi_1(x_2) & \cdots & \psi_1(x_N) \\ \vdots & \vdots & \vdots & \vdots \\ \psi_n(x_1) & \psi_n(x_2) & \cdots & \psi_n(x_N) \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1{}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2{}^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_N{}^2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Definiendo

$$W = \begin{pmatrix} \frac{1}{\sigma_1{}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2{}^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_N{}^2} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad \Psi = \begin{pmatrix} \psi_0(x_1) & \psi_1(x_1) & \cdots & \psi_n(x_1) \\ \psi_0(x_2) & \psi_1(x_2) & \cdots & \psi_n(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \psi_0(x_N) & \psi_1(x_N) & \cdots & \psi_n(x_N) \end{pmatrix} \quad a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}$$

la matriz de coeficientes de las ecuaciones normales queda como

$$A = \Psi^T W \Psi$$

y podemos escribir las ecuaciones normales en forma compacta como

$$\Psi^T W \Psi a = \Psi^T W y$$

con lo que

$$a = (\Psi^T W \Psi)^{-1} \Psi^T W y = \Psi^{-1} W^{-1} (\Psi^T)^{-1} \Psi W y = \Psi^{-1} y$$

Obtenenos una ley lineal para la dependencia de a co y. Llamando

$$S = \Psi^{-1}$$

podemos escribir la ley lineal como

$$a = Sy$$

7.6.4. Errores de los parámetros

Si tenemos una ley lineal

$$a_i = \sum_{j=1}^{N} S_{ij} y_j$$

los valores medios de los parámetros viene dado por

$$\bar{a_i} = \sum_{j=1}^{N} S_{ij} \bar{y_j}$$

Las medidas y_i son independientes, y por lo tanto, sus covarianzas

$$\sigma^2(y_l, y_m) = E[(y_l - \bar{y_l})(y_m - \bar{y_m}) = \delta_{lm} \sigma_l^2$$

son nulas. Las varianzas y covarianzas de los parámetros vienen dadas por

$$\sigma^2(a_i) = E[(a_i - \bar{a_i})^2] = \sum_{lm} S_{il} S_{im} E[(y_l - \bar{y_l})(y_m - \bar{y_m})] = \sum_{lm} S_{il} S_{im} \delta_{lm} \sigma^2(y_m) = \sum_{m} S_{mi}^T S_{im} \sigma_m^2 = [SW^{-1}S^T]_{ii}$$

$$\begin{aligned} \sigma(a_i, a_j) &= E[(a_i - \bar{a_i})(a_j - \bar{a_j})] = S_{il} S_{jm} E[(y_l - \bar{y_l})(y_m - \bar{y_m})] = S_{il} S_{jm} \delta_{lm} \sigma^2(y_m) = S_{mi}^T S_{jm} \sigma_m^2 = \\ &[SW^{-1}S^T]_{ij} \end{aligned}$$

La matriz $SW^{-1}S^T$ cumple

$$SW^{-1}S^T = \Psi^{-1} W^{-1} (\Psi^{-1})^T = (\Psi^T W \Psi)^{-1} = A^{-1}$$

por lo que que la matriz de covarianzas es la inversa de la matriz de coeficientes. Podemos expresar los parámetros con su error como

$$a_i \pm \sqrt{[A^{-1}]_{ii}}$$

Si el término (i,j) de A^{-1} es elevado, entonces los parámetros a_i y a_i están muy correlacionados, y la supresión de uno de ellos debe ser considerada. Notemos que, en el caso de ajuste por funciones ortogonales, la matriz de coeficientes es diagonal y por lo tanto también su inversa, la matriz de covarianzas. Por lo tanto, los coeficientes de los ajustes por funciones ortogonales no están correlacionados, lo cual es una ventaja adicional obtenida en el empleo de este tipo de funciones. Los errores de los parámetros vienen dados, en el caso de ajustes mediante funciones ortogonales, por

$$a_i \pm \sqrt{< p_i | p_i >^{-1}}$$

Ajuste de puntos experimentales mediante una línea recta

En el caso del ajuste lineal tenemos el sistema de ecuaciones para el vector de parámetros a

$$Aa = b$$

donde

$$A = \begin{bmatrix} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} & \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} & \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \end{bmatrix} = \begin{bmatrix} S & S_x \\ S_x & S_{xx} \end{bmatrix}$$

y

$$b = \begin{bmatrix} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \\ \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \end{bmatrix} = \begin{bmatrix} S_y \\ S_{xy} \end{bmatrix}$$

Las soluciones de los parámetros son

$$a_0 = \frac{S_y S_{xx} - S_x S_{xy}}{S S_{xx} - S_x^2} \qquad a_1 = \frac{S S_{xy} - S_x S_y}{S S_{xx} - S_x^2}$$

y la matriz de covarianzas es

$$A^{-1} = \frac{1}{\Delta} \begin{bmatrix} S_{xx} & -S_x \\ -S_x & S \end{bmatrix}$$

donde el determinante de la matriz de coeficientes $\Delta = S S_{xx} - S_x^2$. Si el ajuste es $y = a_0 + a_1 x$ los errores de a_0 y $a_1$valen $\sigma(a_0) = \sqrt{\frac{S_{xx}}{\Delta}}$ y $\sigma(a_1) = \sqrt{\frac{S}{\Delta}}$ mientras que la covarianza de a_0 y $a_1$viene dada por

$$\sigma^2(a_0, a_1) = \frac{-S_x}{\Delta}$$

Se define el coeficiente de correlación de los parámetros $r(a_0, a_1)$ como la covarianza dividida por el producto de desviaciones típicas

$$r(a_0, a_1) = \frac{\sigma^2(a_0, a_1)}{\sigma(a_0)\sigma(a_1)} = \frac{-S_x}{\sqrt{S S_{xx}}}$$

y está comprendido entre -1 y 1. Si es positivo los errores de a_0y $a_1$tienen el mismo signo y si es negativo, signo contrario.

7.6.5. La distribución χ^2

La variable aleatoria

$$\chi^2(a_0, a_1, \ldots a_n) = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i, a_0, \ldots, a_n)}{\sigma_i} \right)^2$$

se distribuye mediante un distribución de probabilidad bien conocida en Estadística, conocida como distribución χ^2(de ahí nuestra notación). Su valor nos indica la bondad del ajuste.

En general, si tenemos k variables aleatorias y_i distribuidas normalmente con media μ_i y desviación típica σ_i, la variable

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{y_i - \mu_i}{\sigma_i} \right)^2$$

se distribuye según la distribución χ^2 con k grados de libertad. Esta distribución depende de dos parámetros, la variable χ^2 y el número de grados de libertad ν, que en nuestro caso es el número de puntos menos el número de parámetros, $\nu = N - n - 1$. La distribución χ^2 está definida como

$$f(\chi^2, \nu) = \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} \left(\chi^2\right)^{\nu/2-1} e^{-\frac{\chi^2}{2}} \qquad \chi^2 > 0$$

donde $\Gamma(x) = \int_0^\infty due^{-u}u^{x-1}$. Esta función de distribución tiene media $\mu = \nu$ y varianza $\sigma^2 = 2\nu$, con un máximo en $\nu - 2$. En la figura se muestra la distribución χ^2 con 6 grados de libertad.

La distribución χ^2 se aproxima de la distribución normal para grandes valores de ν. En la práctica, para $\nu > 30$ es aproximadamente normal. La probabilidad de que $\chi^2 < \chi_0^2$ es

$$F(\chi_0^2, \nu) = \int_0^{\chi_0^2} f(\chi^2, \nu) d\chi^2$$

y la probabilidad de que $\chi^2 > \chi_0^2$ es

$$P(\chi^2 > \chi_0^2) = 1 - F(\chi_0^2, \nu)$$

Si $P(\chi^2 > \chi_0^2) < 0{,}01$ tenemos menos un 1 % de probabilidad de encontrar este valor de χ_0^2 y podemos rechazar el ajuste con un nivel de confianza de un 1 %. En general si $\chi^2/\nu > 2$ podemos pensar que el ajuste no es aceptable para $\nu > 30$. Generalmente esto significa que nuestro modelo no describe adecuadamente los datos, sea porque el número de funciones base empleadas es insuficiente o porque las funciones base empleadas son inadecuadas. El valor medio de χ^2/ν es 1 y su desviación típica es $\sqrt{2/\nu}$. Si obtenemos $\chi^2 \ll 1$ entonces lo más probable es que estemos sobreestimando los errores experimentales.

7.7. Tests estadísticos basados en la distribución χ^2

La distribución χ^2 es una herramienta poderosa para decidir si una ley determinada describe adecuadamente unos datos experimentales. Si tenemos N datos con $n+1$ parámetros, el valor de

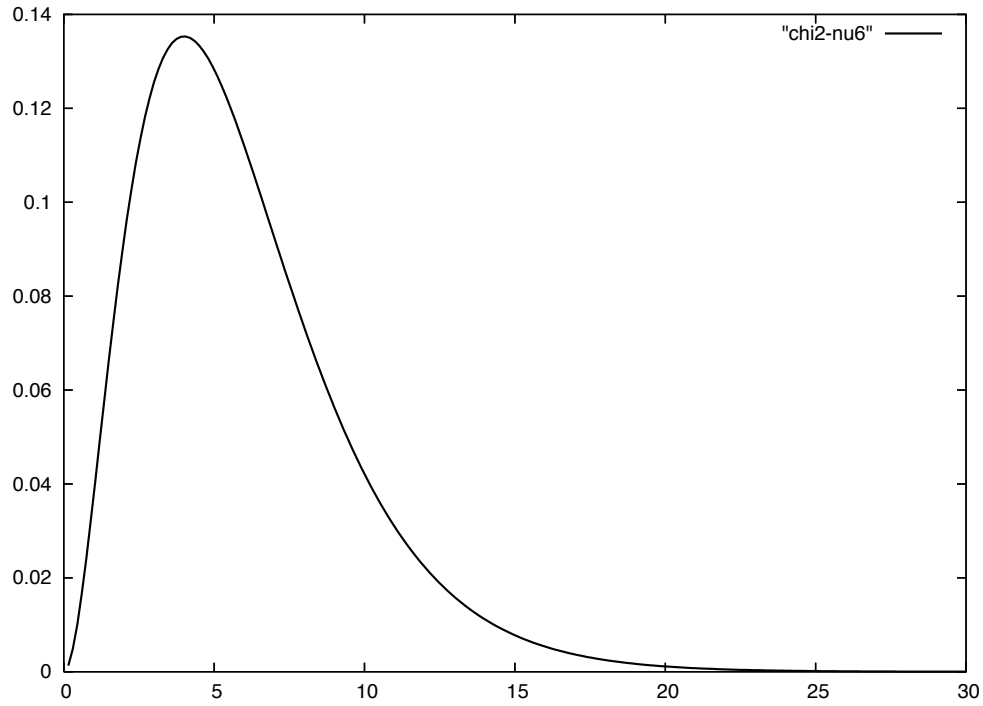

Figura 7.1: Distribución χ^2 con 6 grados de libertad

χ^2 debe de satisfacer la distribución χ^2 con $\nu = N - n - 1$ grados de libertad. Esto quiere decir que si obtenemos un valor de χ^2 muy pequeño o muy grande, este valor es muy poco probable y la ley no es satisfactoria. Cuando se obtienen valores muy pequeños, lo que sucede en general es que los errores están sobreestimados. Por lo tanto se presta atención en general a los valores muy grandes de χ^2. Si por ejemplo, obtenemos un valor χ^2_0 tal que

$$P(\chi^2 > \chi^2_0) = 0{,}05$$

este valor sólo tiene un 5 % de probabilidad de ocurrir. En este caso, podemos rechazar la ley (el conjunto de parámetros) con un nivel de significación del 5 %. Decimos que χ^2 está fuera del intervalo de confianza de 95 %. En el ajuste de datos experimentales se suele prestar atención sólo a valores de χ^2 grandes, por lo que decimos que hacemos un test de una cola. Sin embargo, si no hay evidencias de sobreestimación de los errores, se debe hacer un test de dos colas. Elegimos un nivel de significación α (generalmente de 0.05 o 0.01) y determinamos (mediante tablas estadísticas o un programa) los valores de $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ tales que

$$P(\chi^2 < \chi^2_{\alpha/2}) = P(\chi^2 > \chi^2_{1-\alpha/2}) = \alpha/2$$

Al intervalo $[\chi^2_{\alpha/2}, \chi^2_{1-\alpha/2}]$ le denominamos intervalo de confianza de nivel $1-\alpha$. Si el valor de χ^2 obtenido cae dentro de este intervalo, aceptamos la ley con un nivel de confianza $1-\alpha$ mientras que si cae fuera la rechazamos con un nivel de significación de α(normalmente se expresa en %). Por ejemplo, si tenemos 13 puntos ajustados por una parábola, tenemos $\nu = 10$. Si queremos hacer un test con un intervalo de confianza del 5 %, encontramos en las tablas que para $\nu = 10$ $\chi^2_{0{,}025} = 3{,}25$ y $\chi^2_{0{,}975} = 20{,}5$. Por lo tanto, si obtenemos valores de χ^2 menores que 3.25 o mayores que 20.5, rechazamos los parámetros con un nivel de significación del 5 %, mientras que si χ^2 cae en este intervalo, aceptamos los parámetros con un nivel de confianza del 95 %. Valores de χ^2 muy pequeños pueden ser indicativos de datos fraudulentos (“amañados”).

7.8. Ajuste de funciones que dependen en forma no lineal de los parámetros

En diversas ciencias aparecen frecuentemente leyes con una dependencia no lineal de los parámetros. En este caso, no existe un sistema de ecuaciones cuya solución de el valor óptimo de los parámetros. En el caso no lineal, la solución a menudo no es única, sino que existen varios mínimos relativos.

7.8.1. Reducción a la forma lineal mediante cambio de variables

En lagunas ocasiones una ley no lineal se puede reducir a otra lineal mediante un cambio de variables. Esto sucede por ejemplo en el caso de leyes exponenciales

$$y = ce^{ax}$$

En este caso el cambio de variables de y a $y' = \ln y$ reduce el problema a la ley lineal

$$y' = c' + ax$$

con $c' = \ln c$. Este cambio de variables tiene la ventaja adicional de resalta los detalles de la ley para valores pequeños de y (si y varía entre 1 y 10^6, y' varía entre 0 y 6). En el caso de datos experimentales afectados de errores, también hay que transformar los errores. En el caso de la ley exponencial

$$\varepsilon y' = \frac{dy'}{dy}\varepsilon y = \frac{\varepsilon y}{y}$$

7.8.2. Método de la máxima pendiente

Frecuentemente tenemos una ley no lineal

$$y = f(\mathbf{x}, \mathbf{a})$$

donde $\mathbf{a} = (a_0, a_1, ..., a_n)$ es el vector de parámetros y $\mathbf{x} = (x_1, x_2, ..., x_m)$ es un vector de coordenadas que toma valores en un espacio de m dimensiones (no necesariamente coordenadas físicas). La función f es una función no lineal de los parámetros a_i. Si realizamos una serie de N medidas y_i con errores experimentales σ_i en N puntos $\mathbf{x_i}$, la función χ^2 es también no lineal

$$\chi^2(a_0, a_1, ..., a_n) = \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x_i}, \mathbf{a})^2}{\sigma_i^2}$$

El conjunto óptimo de parámetros $\mathbf{a}$ es aquel que minimiza χ^2. Sin embargo no tenemos un sistema de ecuaciones para calcularlo. La forma de encontrar el mínimo es avanzar en la dirección del espacio de los parámetros en la dirección en la que χ^2 disminuye, considerando χ^2 como una superficie en un espacio de $n+1$ dimensiones. Como la dirección de máximo aumento viene dada por el gradiente, la dirección de máxima disminución $\mathbf{u}$ es la dirección opuesta al gradiente:

$$\mathbf{u} = -\nabla_{\mathbf{a}} \chi^2(\mathbf{x}, \mathbf{a}) = \left(-\frac{\partial \chi^2}{\partial a_0}, -\frac{\partial \chi^2}{\partial a_1}, -\frac{\partial \chi^2}{\partial a_2}, ..., -\frac{\partial \chi^2}{\partial a_n} \right)$$

Partimos de un punto inicial $\mathbf{a_0}$ dado por razonamientos fenomenológicos o teóricos o incluso arbitrario. Las derivadas se pueden calcular numéricamente si no conocemos la forma analítica de f. Si estamos lejos del mínimo, avanzamos una distancia h en el espacio de los parámetros

$$\mathbf{a_1} = \mathbf{a_0} + h\mathbf{u}$$

y recalculamos el valor de χ^2. Si χ^2 disminuye, aumentamos h por un factor F de éxito (10 es una opción frecuente, pero también se puede elegir un valor menor como por ejemplo 2) mientras que si χ^2 aumenta dividimos h por un factor de fracaso (2 es un valor común). De esta manera nos vamos aproximando al mínimo. Tendremos en nuestro programa una actualización de los valores de h y $\mathbf{a}$ dadas por

$$h = h_{old} F$$

$$\mathbf{a_{nuevo}} = \mathbf{a_{actual}} + h\mathbf{u}$$

La función χ^2 es aproximadamente parabólica cerca del mínimo. Si el mínimo es $\mathbf{a_0}$, podemos desarrollar $\chi^2(\mathbf{a_0})$ en serie de potencias alrededor de $\mathbf{a}$

$$\chi^2(\mathbf{a}_0) = \chi^2(\mathbf{a}) + \nabla_{\mathbf{a}}\chi^2(\mathbf{a}) \cdot (\mathbf{a_0} - \mathbf{a}) + \frac{1}{2}\frac{\partial^2\chi^2(\mathbf{a})}{\partial a_i \partial a_j}(a_{i0} - a_i)(a_{j0} - a_j) + \cdots$$

y retener los tres términos escritos explícitamente. El tercer término del segundo miembro es una forma cuadrática construida con el Hessiano de χ^2. Podemos escribir esta ecuación como una función de la diferencia $\mathbf{d} = \mathbf{a_0} - \mathbf{a}$**:**

$$\chi^2(\mathbf{a}_0) = \chi^2(\mathbf{a}) + \nabla_{\mathbf{a}}\chi^2(\mathbf{a}) \cdot \mathbf{d} + \frac{1}{2}\mathbf{d^T} H \mathbf{d}$$

donde la matriz H viene dada por

$$H_{ij} = \frac{1}{2}\frac{\partial^2\chi^2(\mathbf{a})}{\partial a_i \partial a_j}$$

Reteniendo estos dos términos y calculando el gradiente a ambos lados, imponiendo la condición

$$\nabla_{\mathbf{d}}\chi^2(\mathbf{a}_0) = 0$$

obtenemos una estimación de $\mathbf{d} = \mathbf{a_0} - \mathbf{a}$:

$$\nabla_{\mathbf{a}}\chi^2(\mathbf{a}) + H\mathbf{d} = 0$$

de donde obtenemos

$$\mathbf{d} = -H^{-1}\nabla_{\mathbf{a}}\chi^2(\mathbf{a})$$

de donde obtenemos una estimación de $\mathbf{a_0}$:

$$\mathbf{a_0} = \mathbf{a} - H^{-1}\nabla_{\mathbf{a}}\chi^2(\mathbf{a})$$

Si estamos cerca del mínimo podemos intentar obtener el mínimo mediante el siguiente esquema iterativo inspirado en la anterior ecuación:

$$\mathbf{a_{new}} = \mathbf{a_{old}} - H^{-1}\nabla_{\mathbf{a}}\chi^2(\mathbf{a_{old}})$$

que suele converger si estamos suficientemente próximos del mínimo. Cada iteración implica el cálculo del gradiente y del Hessiano de χ^2. Vamos ahora a obtener las expresiones explícitas del gradiente y Hessiano de χ^2. Tenemos

$$\chi^2(a_0, a_1, ..., a_n) = \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x_i}, \mathbf{a})^2}{\sigma_i^2}$$

con lo que tenemos para las componentes del gradiente

$$\frac{\partial\chi^2(\mathbf{a})}{\partial a_k} = -2\sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x_i}, \mathbf{a})}{\sigma_i^2}\frac{\partial f(\mathbf{x_i}, \mathbf{a})}{\partial a_k}$$

Volviendo a derivar, tenemos para las componentes del Hessiano

$$\frac{\partial^2\chi^2(\mathbf{a})}{\partial a_k \partial a_l} = 2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\frac{\partial f(\mathbf{x_i},\mathbf{a})}{\partial a_k}\cdot\frac{\partial f(\mathbf{x_i},\mathbf{a})}{\partial a_l} - (y_i - f(\mathbf{x_i},\mathbf{a})\cdot\frac{\partial^2 f(\mathbf{x_i},\mathbf{a})}{\partial a_k \partial a_l}\right]$$

El término

$$-2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(y_i - f(\mathbf{x_i},\mathbf{a})\cdot\frac{\partial^2 f(\mathbf{x_i},\mathbf{a})}{\partial a_k \partial a_l}$$

es una suma de valores aleatorios, ya que $y_i - f(\mathbf{x_i},\mathbf{a})$ se distribuye normalmente,por lo que en general es despreciable. De hecho se encuentra que frecuentemente las iteraciones convergen mejor si se elimina este término, por lo que se toma

$$\frac{\partial^2\chi^2(\mathbf{a})}{\partial a_k \partial a_l} = 2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\frac{\partial f(\mathbf{x_i},\mathbf{a})}{\partial a_k}\cdot\frac{\partial f(\mathbf{x_i},\mathbf{a})}{\partial a_l}\right]$$

El método iterativo queda por lo tanto como

$$a_{l,new} = a_l + 2H_{lk}^{-1}\left[\sum_{i=1}^{N}\frac{(y_i - f(\mathbf{x_i},\mathbf{a})}{\sigma_i^2}\frac{\partial f(\mathbf{x_i},\mathbf{a})}{\partial a_k}\right]$$

con

$$H_{lk} = 2\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[\frac{\partial f(\mathbf{x_i},\mathbf{a})}{\partial a_k}\cdot\frac{\partial f(\mathbf{x_i},\mathbf{a})}{\partial a_l}\right]$$

con lo que cada paso implica sólo el cálculo de $f(\mathbf{x_i},\mathbf{a})$, y su gradiente, o sea la evaluación de $n+2$ funciones para cada punto y_i.

7.9. Ejercicios

1. Determínese los parámetros a y b que ajustan la curva $y = a + b\sin(x)$ a la tabla de valores adjunta. Obtener el valor de χ^2 y los errores con los que se determinan los parámetros. Hágase una representación gráfica de los valores ajustados y empíricos.

x	0,0	0,3	0,5	0,7	0,9	1,0
y	1,80	1,71	1,50	1,45	1,17	1,17
σ	0,2	0,2	0,2	0,2	0,2	0,2

2. Se desea ajustar la función modelo $y = \sqrt{ax^2 + b}$ al conjunto de datos especificado en la tabla adjunta. Hacer las transformaciones de variables adecuadas para que el ajuste sea lineal, realizando las transformaciones correspondientes para los errores.

x	0.1	0.6	1.0	1.5	2.0	2.5
y	1.0	1.4	2.1	2.8	3.6	4.4
σ	0.05	0.2	0.05	0.1	0.2	0.1

3. Ajustar minimizando χ^2 la curva $y = Ae^{-x} + Be^{x}$ a la siguiente tabla de valores. Presentar los valores de los parámetros ajustados, sus errores y el valor de χ^2¿Se trata de un buen ajuste?

x	-4	-3	-1	0	1	2	4
y	163	61	11	8	15	37	270
σ	3	1	1	0.2	1	2	4

4. Ajustar minimizando χ^2 la curva $y = A + Be^{x}$ a la siguiente tabla de valores. Presentar los valores de los parámetros ajustados, sus errores y el valor de $\chi^{2.}$¿Se trata de un buen ajuste?

x	0.0	1.0	1.5	2.0	2.5
y	5.1	8.2	22.0	112.1	1039.0
σ	0.2	0.1	0.2	0.1	0.3

5. Ajustar minimizando χ^2 la curva $y = a\sqrt{x} + b\ln(1+x)$ a la siguiente tabla de valores. Presentar los valores de los parámetros ajustados, sus errores y el valor de χ^2¿Se trata de un buen ajuste?

x	1	2	4	8	10	12
y	4	6	9	12	14	15
σ	0.3	0.3	0.3	0.3	0.3	0.3

6. Ajustar minimizando χ^2 la curva $y = ax + be^{-x^2/2}$ a la siguiente tabla de valores. Presentar los valores de los parámetros ajustados, sus errores y el valor de χ^2¿Se trata de un buen ajuste?

x	-2	-1	0	1	2
y	-1	5	10	7	4
σ	0.4	0.2	0.2	0.2	0.4