Mínimos cuadrados lineales

Versión 22 de octubre de 2020

Este apunte sirve como complemento de la clase teórica sobre mínimos cuadrados y del material que se encuentra en el libro rojo del curso de Geometría y Álgebra lineal 2 de Facultad de Ingeniería, UDELAR[1].

1. El problema de mínimos cuadrados

1.1. ¿Qué es una solución?

Se considera el sistema lineal de ecuaciones $AX = b$ donde $A \in \mathcal{M}_{m\times n}$, $b \in \mathbb{R}^m$. Se denota por $\text{col}(A)$ al espacio de columnas de A, es decir,

$$\text{col}(A) = [A_1, \ldots, A_n],$$

siendo A_i la i-ésima columna de A. De GAL1 se sabe que el sistema lineal tiene solución si y sólo si, $b \in \text{col}(A)$, es decir, si b se escribe como combinación lineal de las columnas de A.

Ahora bien, supongamos que $b \notin \text{col}(A)$, es decir, el sistema $AX = b$ no tiene solución. ¿Es posible resolver algún problema similar, de manera que la solución al nuevo problema sea lo más cercana posible al problema original? Se puede también enfocar de otra manera, ¿existe alguna noción de solución del sistema lineal que me permita resolverlo a pesar que no haya solución algebraica? Ambas preguntas pueden combinarse y tener una respuesta positiva si se precisan algunos conceptos. Primero, qué se entiende por similar. Una forma muy natural de valorar si dos vectores son similares es, por ejemplo, preguntarse si están cerca en el sentido de la distancia euclidiana. Bajo esa noción de similaridad, se puede buscar entonces al vector $\hat{X} \in \mathbb{R}^n$ que está más cerca de resolver el problema, es decir, se quiere hallar $\hat{X} \in \mathbb{R}^n$ tal que minimice la función $r\colon \mathbb{R}^n \to \mathbb{R}$, definida por

$$r(Y) = \|AY - b\|_2. \tag{1}$$

Observar que, el sistema lineal $AX = b$ tiene solución si y sólo si el mínimo de la función r es 0. Sin embargo, cuando el sistema no tiene solución, se puede probar que la función r alcanza siempre un mínimo y este mínimo es único[2]. El lugar donde se alcance el mínimo de r es lo que se considerará como la solución más cercana al sistema lineal $AX = b$, puesto que es el vector de $\mathbb{R}^n$ tal que al multiplicarlo por la matriz A, el resultado es lo más cercano posible a b. Se enfatiza que no es una solución algebraica, no verifica la ecuación original. Es un nuevo concepto de solución, un poco más flexible y general que la solución algebraica, puesto que cada vez que hay solución algebraica el mínimo de r es 0 y se alcanza en ella, pero el mínimo puede no ser 0.

[1] Sugerencias y/o correcciones, escribir a nfrevenza@fing.edu.uy.

[2] Mostrar que la función r siempre tiene un mínimo se puede ver vía proyección ortogonal (lo que se hará en estas notas) o usando con cierto cuidado el teorema de Weierstrass sobre que una función continua en un compacto tiene máximo y mínimo.

Veamos que hallar el mínimo de (1) puede resolverse geométricamente con las herramientas del curso. Notar que minimizar la función (1) es equivalente a minimizar su cuadrado. Ahora bien,

$$\min_{Y\in\mathbb{R}^n} \|AY-b\|_2^2 = \min_{u\in\mathrm{col}(A)} \|u-b\|_2^2 = \mathrm{dist}^2(\mathrm{col}(A),b), \tag{2}$$

puesto que $AY \in \mathrm{col}(A)$ para cualquier elección de $Y \in \mathbb{R}^n$ y todo vector de $\mathrm{col}(A)$ se escribe como combinación lineal de las columnas de A (es decir, de la forma AY para algún Y). El problema de minimizar la distancia entre el vector b y un subespacio ($\mathrm{col}(A)$ en este caso), ya lo tenemos resuelto y su solución está dada por la proyección ortogonal al subespacio. En otras palabras, el mínimo de $\|u-b\|_2^2$ con $u \in \mathrm{col}(A)$ se alcanza cuando $u = P_{\mathrm{col}(A)}(b)$. Esto nos dice quién debe ser $u \in \mathrm{col}(A)$ para minimizar la distancia entre b y $\mathrm{col}(A)$, pero nuestro interés residía en hallar $Y \in \mathbb{R}^n$ tal que $\|AY-b\|_2^2$ sea mínimo.

Notar que el sistema de ecuaciones $AX = P_{\mathrm{col}(A)}(b)$ siempre tiene solución, porque el término independiente pertenece al espacio de columnas de A. Hallar Y de forma que $r^2(Y) = \|AY-b\|_2^2$ sea mínimo, es equivalente a resolver el sistema lineal

$$AX = P_{\mathrm{col}(A)}(b). \tag{3}$$

Por tanto, si llamamos $\hat{X}$ a la solución de (3), $\hat{X}$ es el vector de $\mathbb{R}^n$ buscado. Observar que cuando la matriz A tiene rango n, esto es, el rango coincide con la cantidad de columnas, la solución es única porque las columnas de A forman una base de $\mathrm{col}(A)$.

El vector $\hat{X}$ es lo que se denomina, *la solución por mínimos cuadrados del sistema* $AX = b$, y este nuevo concepto de solución coincide con la solución usual (algebraica) cuando el sistema lineal es compatible, pero que en el caso que el sistema no sea compatible, nos brinda una respuesta alternativa con ciertas propiedades interesantes (una "solución" en otro sentido).

Lo que se ha hecho aquí es generalizar el concepto de qué se entiende por solución. No se habla de una solución en el sentido algebraico de verificar una ecuación, sino en un sentido variacional, la solución es aquella que minimiza cierta función objetivo (la función r en este caso); y cuando existe solución en el sentido usual (algebraico), ésta también es solución en el sentido variacional. Esta idea de modificar (debilitando) qué se entiende por solución es muy habitual en matemática y es lo que permite encontrar soluciones a problemas que *a priori* parecerían no tenerlas[3].

La función $r(Y)$ recibe el nombre de error cuadrático.

1.2. Ecuaciones normales

Comencemos con un sencillo lema.

Lema. Sea $A \in \mathcal{M}_{m\times n}$ y $v \in \mathbb{R}^m$. Se tiene que, $v \in (\mathrm{col}(A))^\perp$ si y sólo si $A^T v = 0_n$, siendo 0_n el vector nulo de $\mathbb{R}^n$.

Demostración: La prueba de este lema se reduce a observar que un vector pertenece a $(\mathrm{col}(A))^\perp$ si y sólo si es perpendicular a todas las columnas de A, es decir, a todas las filas de A^T, por tanto, si y sólo si $A^T v = 0_n$. $\square$

Una forma alternativa de hallar la solución por mínimos cuadrados (3), es decir, de hallar el mínimo de la función r definida en parte anterior, es la asociada a las ecuaciones normales que se muestra a continuación.

[3]Solo a modo informativo, es posible resolver ecuaciones diferenciales donde las soluciones no sean funciones derivables, lo que parece un oxímoron, porque para verificar la ecuación habría que derivar la función.

Teorema (Ecuaciones normales). Sea $A \in \mathcal{M}_{m\times n}$ y $b \in \mathbb{R}^m$. Se tiene que el vector $\hat{X} \in \mathbb{R}^n$ minimiza la función $r^2(Y) = \|AY - b\|_2^2$ si y sólo si $A^T(A\hat{X} - b) = 0_n$.

Demostración: ($\Rightarrow$). En la sección anterior se mostró que el mínimo de r^2 es la solución por mínimos cuadrados $\hat{X}$ y que proviene de resolver el sistema $AX = P_{\mathrm{col}(A)}(b)$. Por tanto se verifica que $A\hat{X} = P_{\mathrm{col}(A)}(b)$. Se tiene entonces que:

$$A\hat{X} - b = P_{\mathrm{col}(A)}(b) - b = -P_{\mathrm{col}(A)^\perp}(b),$$

ya que el término independiente (y cualquier vector) se puede escribir como

$$b = P_{\mathrm{col}(A)^\perp}(b) + P_{\mathrm{col}(A)}(b).$$

Finalmente, como $A\hat{X} - b \in \mathrm{col}(A)^\perp$ el Lema previo nos dice que $A^T(A\hat{X} - b) = 0_n$ como se quería probar.

($\Leftarrow$). Sea ahora $\hat{X}$ tal que $A^T(A\hat{X} - b) = 0_n$. Se mostrará que minimiza a la función r^2. En efecto,

$$\begin{aligned} r^2(Y) &= \|AY - b\|_2^2 \\ &= \|A(Y - \hat{X}) + (A\hat{X} - b)\|_2^2 \\ &= \|A(Y - \hat{X})\|_2^2 + \|(A\hat{X} - b)\|_2^2 + 2\langle A(Y - \hat{X}), (A\hat{X} - b)\rangle. \end{aligned}$$

Ahora bien, el producto interno usual se puede escribir de forma matricial como $\langle U, V\rangle = U^T V$, por tanto

$$\begin{aligned} r^2(Y) &= \|A(Y - \hat{X})\|_2^2 + \|(A\hat{X} - b)\|_2^2 + 2\left[A(Y - \hat{X})\right]^T (A\hat{X} - b) \\ &= \|A(Y - \hat{X})\|_2^2 + \|(A\hat{X} - b)\|_2^2 + 2(Y - \hat{X})^T A^T (A\hat{X} - b) \\ &= \|A(Y - \hat{X})\|_2^2 + \|(A\hat{X} - b)\|_2^2 \\ &= \|A(Y - \hat{X})\|_2^2 + r^2(\hat{X}), \end{aligned}$$

donde en la segunda línea se usó que $A^T(A\hat{X} - b) = 0_n$. Se sigue que $r^2(Y) \geq r^2(\hat{X})$ para todo $Y \in \mathbb{R}^n$, y por tanto en $\hat{X}$ se alcanza el mínimo de r^2. $\square$

Ecuaciones normales. Se le llama ecuación normal o ecuaciones normales al sistema lineal $A^TAX = A^Tb$, que es equivalente a la condición $A^T(AX - b) = 0_n$. La solución por mínimos cuadrados es la solución a este sistema de ecuaciones normales.

1.3. La importancia de la norma $\|\cdot\|_2$

El problema de mínimos cuadrados se puede resolver analíticamente de forma sencilla porque podemos pensarlo geométricamente. Para ello es fundamental que la norma que se utiliza provenga de un producto interno. Si esto no es así, no se puede definir una proyección ortogonal que se vincule con la norma y la solución cambia completamente. Para ver ello, analicemos el siguiente ejemplo.

Se considera el subespacio $S = [(1, 1)]$ de $\mathbb{R}^2$. Se busca hallar el punto de S que minimiza la distancia en cierta norma al punto $(2, 1)$, es decir se busca el punto de la función $R\colon S \to \mathbb{R}$ dada por

$$R(u) = \|u - (2, 1)\| \qquad \text{con } u \in S, \tag{4}$$

alcanza su mínimo, pero teniendo en cuenta diferentes normas.

Caso $\|\cdot\|_2$. La solución está dada por la proyección ortogonal de $(2,1)$ a S. Una base ortonormal de S es $\{(1/\sqrt{2}, 1/\sqrt{2})\}$, por lo que el punto buscado es

$$P_S(2,1) = \langle (2,1), (1/\sqrt{2}, 1/\sqrt{2})\rangle (1/\sqrt{2}, 1/\sqrt{2}) = (3/2, 3/2).$$

El mínimo es $\|(2,1) - (3/2, 3/2)\|_2 = 1/\sqrt{2}$ y se alcanza en $(3/2, 3/2)$.

Caso $\|\cdot\|_1$. Cualquier elemento de S se escribe como (a,a) con $a \in \mathbb{R}$, por lo que minimizar la distancia entre $(2,1)$ y un punto de S se reduce a buscar el mínimo (en a) de la siguiente expresión:

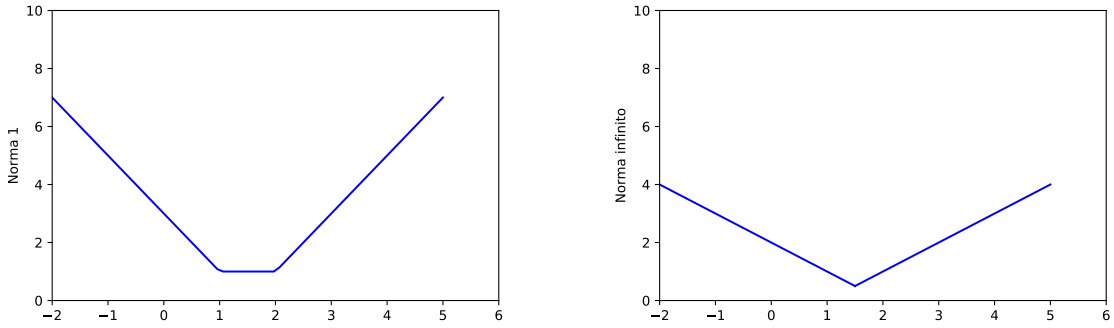
$$\|(2,1) - (a,a)\|_1 = |2-a| + |1-a|.$$

Se puede observar que la función anterior alcanza su mínimo, que vale 1, en infinitos valores (ver las gráficas de la figura 1). Cualquier punto de la forma (a,a) con $a \in [1,2]$ minimiza la distancia a $(2,1)$ a S.

Caso $\|\cdot\|_\infty$. Análogamente, para $\|\cdot\|_\infty$ hay que minimizar la función:

$$\|(2,1) - (a,a)\|_\infty = \text{máx}\{|2-a|, |1-a|\} = \begin{cases} |2-a| & a \leq 3/2 \\ |1-a| & a > 3/2 \end{cases},$$

que alcanza su mínimo en un único valor que es $a = 3/2$, es decir, el punto $(3/2, 3/2) \in S$ es el que minimiza la distancia con la norma ∞ de $(2,1)$ al subespacio S y ese mínimo vale $1/2$.



(a) $\|\cdot\|_1$. (b) $\|\cdot\|_\infty$.

Figura 1: Funciones a optimizar en cada caso.

Recapitulando, con las tres normas existe el mínimo; este es un principio general que tiene que ver con la convexidad de la función R definida en (4). Sin embargo, los puntos donde se alcanza los mínimos pueden ser diferentes, el valor del mínimo es también diferente en general. Más aún, hay ejemplos sencillos donde la solución es única con algunas normas pero no es única si se considera otra norma.

2. La potencia de los mínimos cuadrados

En esta sección se verá como es posible aplicar la formulación abstracta de la parte anterior y la solución por mínimos cuadrados a la resolución de muchísimas aplicaciones. Como sucede en la vida real se deberá hacer algún tipo de elección para tratar el problema.

Comencemos con un planteo general pero lo suficientemente rico como para ilustrar muchos ejemplos. Los pares $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^k$ con $i = 1, \ldots, N$ son el resultado de mediciones (de cierto experimento, de una encuesta, etc). Nos referiremos a ellos como los datos. Notar que

tanto x_i como y_i puede ser multidimensionales, aunque inicialmente podemos pensar que ambas cantidades corresponden a números reales (en ese caso el par (x_i, y_i) lo podemos ubicar en el plano y el análisis es más sencillo).

El primer supuesto es que detrás de estos datos hay cierto modelo que lo explica, es decir, se sabe (o se intuye) que las variables x_i pueden explicar el comportamiento de las y_i. Por ello a las variables que representan las x_i se les llama variables independientes y a las y_i variables dependientes o de respuesta. En definitiva, lo que se buscar es una función $f\colon \mathbb{R}^p \to \mathbb{R}^k$ que explique esta relación, es decir, que $f(x_i) = y_i$.

Ahora bien, este problema así planteado y sin ninguna otra hipótesis adicional, es casi imposible de resolver. ¿Por qué? Porque la función que estoy buscando puede tener cualquier forma *a priori*, podría ser un polinomio de grado 1, 2, 18 o una exponencial, una combinación de senos y cosenos, o directamente una función que no tiene una expresión a partir de funciones elementales. El espacio vectorial de funciones de $\mathbb{R}^p$ a $\mathbb{R}^k$ (aunque sean continuas y diferenciables) es un espacio de dimensión infinita. Entonces, el horizonte de búsqueda debe restringirse de alguna manera razonable, haciendo hipótesis sobre el modelo (modelado) o imponiendo condiciones que se saben *a priori*. Veamos algunos ejemplos.

2.1. Relación lineal entre los datos: la ley de Hooke

Si se asume que hay una relación lineal entre x_i e y_i, entonces tiene sentido buscar solamente entre las funciones de la forma $y = f(x) = \alpha x + \beta$. Los valores de $\alpha, \beta \in \mathbb{R}$ se tienen que determinar para que los datos se ajustar para que el modelo describa lo mejor posible a los datos. Notar que el espacio vectorial de las funciones lineales tiene dimensión 2 y es intuitivamente mucho más sencillo buscar una función allí que en el espacio de funciones reales.

Un ejemplo de modelo lineal (aún más simple) es el de un sistema masa-resorte en equilibrio, donde se estudia el estiramiento del resorte en función de la fuerza que se ejerce sobre el mismo. La ley de Hooke dice que $F = -kL$, donde F es la fuerza ejercida sobre el resorte, L el desplazamiento desde la posición natural y $k > 0$ una constante que depende del resorte. Supongamos que un resorte está amarrado a un techo y se cuelgan de este diferentes masas (ver figura 2) y se mide el desplazamiento desde la posición natural; es decir, para cada masa m_i que se coloca se mide L_i, lo que en los términos de la nomenclatura que mencionamos antes significa que m_i es la variable independiente y L_i es la variable de respuesta. El problema consiste en estimar razonablemente k dado que al medir se cometen errores. Primero, se sabe

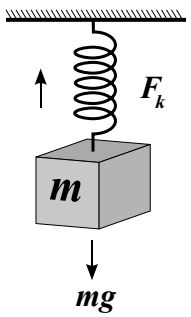


Figura 2: Masa colgando del resorte (tomada de Wikipedia).

que k es siempre la misma constante porque el resorte no cambia (y además, no se lo deforma de manera que cambie su longitud natural). Físicamente se sabe que $L_i = g/km_i$, por lo que se buscará, entre las funciones lineales con forma $f(m) = \alpha m$, a la que mejor aproxime los

datos (m_i, L_i) con $i = 1 \ldots, N$. Como se tiene una forma genérica para f se puede plantear un sistema lineal imponiendo la condición $f(m_i) = L_i$, cuyas incógnitas son los parámetros, en este caso únicamente α. Queda entonces:

$$\left\{ \begin{array}{rcl} \alpha m_1 & = & L_1 \\ \alpha m_2 & = & L_2 \\ \vdots & = & \vdots \\ \alpha m_N & = & L_N \end{array} \right. . \tag{5}$$

El problema inmediato es que el sistema no debería tener solución puesto que las mediciones no son exactas, tienen un margen de error; por lo que el valor de α que se obtiene usando la primera ecuación no será el mismo que el se obtiene usando otra de las ecuaciones. Se está en una situación similar a la de la sección anterior: un sistema lineal de ecuaciones que no tiene solución algebraica. De cualquier forma podemos buscar la solución por mínimos cuadrados (que siempre existe!), es decir, se busca $\alpha \in \mathbb{R}$ de forma que el cuadrado del error cuadrático

$$r^2(\alpha) = \sum_{i=1}^{N} (\alpha m_i - L_i)^2$$

sea mínimo. Conviene remarcar que nuestro interés es en α, que es nuestra incógnita para definir la ecuación que gobierna el fenómeno, el resto son datos dados o que se midieron.

El sistema lineal (5) se escribe matricialmente como $AX = b$ siendo:

$$A = \begin{pmatrix} m_1 \\ \vdots \\ m_N \end{pmatrix}, \qquad b = \begin{pmatrix} L_1 \\ \vdots \\ L_N \end{pmatrix}, \qquad X = (\alpha).$$

Conviene observar que la cantidad de filas de la matriz A y el tamaño del vector b, están dados por la cantidad de datos que se disponen (N en total) y que la cantidad de columnas de la matriz A es siempre igual a la cantidad de parámetros con los que se quiere modelar (uno en este caso). Las ecuaciones normales $A^T AX = A^T b$ en este caso nos quedan:

$$(m_1, \ldots, m_N) \begin{pmatrix} m_1 \\ \vdots \\ m_N \end{pmatrix} (\alpha) = (m_1, \ldots, m_N) \begin{pmatrix} L_1 \\ \vdots \\ L_N \end{pmatrix},$$

es decir, $\alpha \sum_{i=1}^{N} m_i^2 = \sum_{i=1}^{N} m_i L_i$. Entonces el valor de α que resuelve las ecuaciones normales, y en consecuencia, la solución por mínimos cuadrados del sistema (5), es:

$$\alpha = \frac{\sum_{i=1}^{N} m_i L_i}{\sum_{i=1}^{N} m_i^2}.$$

Utilizando que α debe ser igual a g/k, se obtiene que la constante del resorte estimada es:

$$k = \frac{g \sum_{i=1}^{N} m_i^2}{\sum_{i=1}^{N} m_i L_i},$$

siendo g la gravedad.

2.2. Ajuste polinomial: el cómputo de un determinante

Nuevamente se tienen datos (x_i, y_i) con $i = 1, \ldots, N$, pero ahora se asume o se sabe por alguna propiedad física como en el ejemplo anterior, que el comportamiento de los datos sigue alguna ley de forma polinómica. Esto es, se busca un polinomio f de cierto grado que ajuste a los datos. Surgen dos preguntas básicas: ¿cuál debería ser el grado del polinomio?, y una vez fijado el grado, ¿cómo elegimos el mejor polinomio que represente la relación entre los valores de x e y?

Para elegir el grado del polinomio lo mejor es recurrir a alguna heurística del problema o conocimiento que se tenga sobre el fenómeno del que provienen los datos. Esto es fundamental. Si los datos provienen de un experimento donde se mide la distancia recorrida de una partícula en función del tiempo, donde la partícula siempre parte del mismo lugar y con la misma velocidad y estaba sometida a una aceleración constante (por ejemplo una partícula en caída sin rozamiento), entonces es razonable buscar entre las funciones de la forma $f(t) = \alpha t^2 + \beta t + \gamma$ para describir el fenómeno[4]. Por más que se tengan millares de datos, el problema es razonable describirlo con 3 parámetros (α, β y γ) y no buscar un polinomio de grado 500, porque allí tendría demasiados parámetros para determinar (exactamente 501 parámetros).

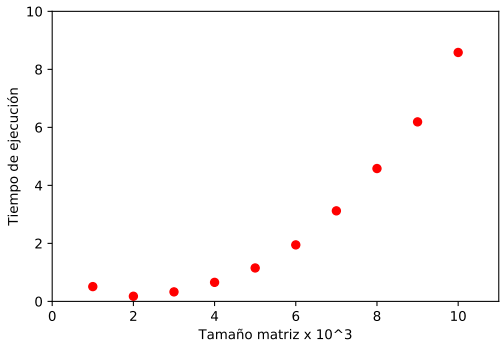
Otro ejemplo natural es el tiempo de ejecución del cálculo del determinante de una matriz $n \times n$. Se sabe que (no es difícil hacer las cuentas) el cálculo del determinante requiere del orden de n^3 multiplicaciones y que el tiempo de ejecución de una computadora es proporcional a la cantidad de multiplicaciones que debe hacer. En este caso, si x_i es el tamaño de la matriz, la variable de respuesta y_i es tiempo de ejecución del programa. Aquí entonces lo razonable parece ser tomar un polinomio de grado 3, $f_4(x) = \alpha x^3 + \beta x^2 + \gamma x + \delta$ (el subíndice indica la cantidad de parámetros porque más adelante volveremos sobre este ejemplo).

La siguiente tabla de valores fue obtenida corriendo un código elemental de *Python* donde se pide que se calcule el determinante de una matriz aleatoria (sin valores nulos, para evitar que sea sencillo) y se calcula el tiempo de ejecución en cada caso. Posteriormente los puntos (x_i, y_i) se grafican en el plano (ver figura 3). En la tabla el tamaño de la matriz será el tamaño real, pero para las cuentas utilizaremos como unidad de medida el tamaño por mil (no cambia absolutamente nada, pero nos permite escribir las matrices y mejorar la visualización de los datos).

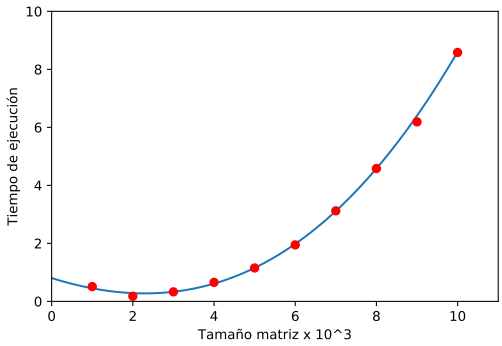
Tamaño matriz (x_i)	Tiempo (seg) (y_i)
1000	0.51
2000	0.177
3000	0.328
4000	0.656
5000	1.152
6000	1.948
7000	3.122
8000	4.581
9000	6.189
10000	8.582

Ahora se plantea el sistema lineal $AX = b$ usando que $y_i = f_4(x_i) = \alpha x_i^3 + \beta x_i^2 + \gamma x_i + \delta$.

[4]Se está pensando que el tiempo t es la variable independiente, es decir, las x_i en la nomenclatura anterior serán medidas de tiempo, y la variable de respuesta es la distancia recorrida y_i.



(a) Datos.



(b) Datos y ajuste con f_4.

Figura 3: A la izquierda se encuentra los datos y a la derecha se grafican en conjunto con el polinomio hallado.

Entonces se tiene que,

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 4 & 2 & 1 \\ 27 & 9 & 3 & 1 \\ 64 & 16 & 4 & 1 \\ 125 & 25 & 5 & 1 \\ 216 & 36 & 6 & 1 \\ 343 & 49 & 7 & 1 \\ 512 & 64 & 8 & 1 \\ 729 & 81 & 9 & 1 \\ 1000 & 100 & 10 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix}, \quad b = \begin{pmatrix} 0{,}51 \\ 0{,}177 \\ 0{,}328 \\ 0{,}656 \\ 1{,}152 \\ 1{,}948 \\ 3{,}122 \\ 4{,}581 \\ 6{,}189 \\ 8{,}582 \end{pmatrix}.$$

Las ecuaciones normales quedan

$$A^T A = \begin{pmatrix} 1978405 & 220825 & 25333 & 3025 \\ 220825 & 25333 & 3025 & 385 \\ 25333 & 3025 & 385 & 55 \\ 3025 & 385 & 55 & 10 \end{pmatrix}, \qquad A^T b = \begin{pmatrix} 17127{,}633 \\ 1919{,}265 \\ 221{,}943 \\ 27{,}245 \end{pmatrix}.$$

De aquí se obtiene que la solución es $\alpha = 0{,}004$, $\beta = 0{,}082$, $\gamma = -0{,}44$ y $\delta = 0{,}805$. Por lo tanto, el polinomio buscado es $f_4(x) = 0{,}004x^3 + 0{,}082x^2 - 0{,}44x + 0{,}805$. El error cuadrático, que recordemos que se define como

$$\sum_{i=1}^{N} \big(y_i - f_4(x_i)\big)^2,$$

en este caso es $r = 0{,}254$.

A veces, en vez de trabajar con el error cuadrático global, se trabaja con el error cuadrático medio, que no es otra cosa que dividir al error cuadrático r entre la cantidad de datos disponibles, es decir, r/N. En este caso el error cuadrático medio es 0,025. La razón para ello es que el error cuadrático aumenta con la cantidad de datos porque se agregan sumandos por lo que muchas veces conviene normalizarlo (para tener idea de si es grande o no, en relación al problema que se quiere modelar).

2.3. Ajuste lineal múltiple: el índice de Gini

En esta subsección vamos a considerar un ejemplo donde la variable independiente x_i es multidimensional, es decir, es un vector de $\mathbb{R}^p$ y la variable de respuesta y_i es real. En la mayoría de las ocasiones se dice que el modelo tiene p variables independientes, dadas por las coordenadas del vector $x_i = (x_{i,1}, \ldots, x_{i,N})$, es decir, cada $x_{i,j}$ es una variable independiente real. Mostremos que a pesar de la multidimensionalidad, funciona de forma similar a lo que se ha visto.

Se cuenta con una serie de datos obtenidos de World Development Indicators del Banco Mundial (databank.worldbank.org) de un conjunto de 78 países para el año 2017. La idea es tratar de analizar un indicador de concentración del ingreso (o de forma dual, de la desigualdad del ingreso) muy conocido en economía que se denomina índice de Gini. En particular nos interesa analizar los determinantes de la concentración del ingreso en función de una serie de variables (las variables independientes) que se consideran potencialmente explicativas del índice de Gini.

Las variables disponibles para cada uno de los 78 países son:

- $x_{i,1}$: es el porcentaje de inscriptos en educación secundaria con respecto al total de jóvenes en edad de estar escolarizados en ese nivel[5].
- $x_{i,2}$: es el porcentaje del ingreso derivado de la agricultura en relación al PIB.
- $x_{i,3}$: es el porcentaje de la población que vive en zonas urbanas sobre la población total del país.
- $x_{i,4}$: es el crecimiento de la población; es una tasa (por tanto puede ser negativa).
- $x_{i,5}$: es el crecimiento del PIB; es también una tasa.
- $x_{i,6}$: es el producto Interno Bruto per cápita.
- y_i : es el coeficiente de Gini, que está representado por un valor de 0 a 100). El índice de Gini es nuestra variable de respuesta (la que se busca modelar a través de las variables independientes anteriores).

Antes de plantear el modelo vamos a mostrar algunas de estas variables para algunos de los países que están en la base de datos, solamente para tener una idea de cómo son nuestros datos.

País	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	$x_{i,5}$	$x_{i,6}$	y_i
Argentina	108.0153	5.4784	91.75	1.0371	2.6686	20310.0	40.6
Chile	99.6577	3.8742	87.49	1.4252	1.2792	23260.0	46.6
Estados Unidos	98.7699	0.9165	82.06	0.6405	2.217	61120.0	41.5
Estonia	115.0141	2.3246	68.78	0.1211	4.8567	32790.0	32.7
Finlandia	152.1684	2.3753	85.33	0.2347	2.6525	46480.0	27.1
Honduras	54.1612	12.8449	56.46	1.6922	4.788	4570.0	50.5
Portugal	117.5025	1.9771	64.65	-0.2439	2.795	31840.0	35.5
St. Lucia	88.1776	1.9854	18.61	0.5158	3.6719	12760.0	51.2
Togo	61.6865	23.607	41.16	2.4793	4.4494	1680.0	43.1
Uruguay	115.2387	5.1088	95.24	0.3648	2.5913	21380.0	39.5

[5]Ello puede llevar a que algunos porcentajes sean mayores a 100 % por la población extra edad.

Se plantea buscar entre las funciones afines (lineales + una constante) de 6 variables a la función que explique el índice de Gini como consecuencia de las variables independientes antes definidas. Esto es, se busca $f\colon \mathbb{R}^6 \to \mathbb{R}$ dada por

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6, \tag{6}$$

donde $\beta_0, \ldots, \beta_6 \in \mathbb{R}$ son los parámetros del modelo, que precisamos determinar a partir de nuestros datos. Imponiendo la condición $y_i = f(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5}, x_{i,6})$ para cada uno de los 78 países se obtiene un sistema lineal de ecuaciones donde las incógnitas son los 7 parámetros. Si A es la matriz asociada a este sistema lineal de ecuaciones sus dimensión es 78×7 y el término independiente b, dado por los índices de Gini, es de tamaño 78×1 (notar que la primer columna de A tendrá todas las entradas iguales a 1). Las ecuaciones normales sí las vamos a escribir puesto que nos queda un sistema lineal 7×7, con

$$A^T A = \begin{pmatrix} 78 & 7484 & 621{,}5 & 4987 & 73{,}09 & 300{,}6 & 1900000 \\ 7484 & 767900 & 46550 & 504500 & 6036 & 27060 & 206500000 \\ 621{,}5 & 46550 & 10330 & 30770 & 932{,}2 & 2989 & 6728000 \\ 4987 & 504500 & 30770 & 351500 & 4340 & 17750 & 139400000 \\ 73{,}09 & 6036 & 932{,}2 & 4340 & 141{,}1 & 326{,}7 & 1345000 \\ 300{,}6 & 27060 & 2989 & 17750 & 326{,}7 & 1455 & 6284000 \\ 1900000 & 206500000 & 6728000 & 139400000 & 1345000 & 6284000 & 72870000000 \end{pmatrix},$$

y

$$A^T b = \begin{pmatrix} 2852{,}6 \\ 268359{,}148 \\ 23993{,}253 \\ 180996{,}251 \\ 2874{,}689 \\ 11081{,}972 \\ 65023682 \end{pmatrix}.$$

donde la solución es

$$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) = (47{,}944, -0{,}091, -0{,}309, 0{,}062, 2{,}362, -0{,}382, 0).$$

En este caso el error cuadrático es 48,937 y el error cuadrátrico medio es 0,627.

Una pregunta razonable que nos podemos hacer es, qué tan bueno es el modelo hallado vía mínimos cuadrados. Obviamente podemos considerar el error cuadrátrico medio (o global) como una medida de ello, pero también se pueden hacer otras cosas. Una idea que se utiliza a veces para testear esto es la siguiente: vamos a hacer de nuevo el método de mínimos cuadrados y para calcular $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ pero omitiendo algunos países en los datos. Supongamos por ejemplo que omitimos las ecuaciones de los datos correspondientes a Estonia y Uruguay, es decir, ahora tenemos un modelo donde la matriz A tiene tamaño 75×7. Resolviendo vía ecuaciones normales se obtiene una función f como en (6) con coeficientes

$$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) = (48{,}23, -0{,}092, -0{,}315, 0{,}058, 2{,}392, -0{,}382, 0).$$

Ahora calculemos los valores de f(Estonia) y f(Uruguay), donde obviamente se evalúa f en las variables independientes correspondientes a Estonia y lo mismo para Uruguay. El resultado se resume en el siguiente cuadro.

País	Gini real	Gini predecido
Estonia	32.7	32.84
Uruguay	39.5	37.16

Para usar esta idea de forma correcta, lo conveniente es quitar los datos de varios países, y usar el resto para estimar (en el lenguaje de la ciencia de datos, se dice que los países que no eliminamos de nuestros datos son los que usamos para entrenar el modelo y los que guardamos, Estonia y Uruguay en este caso, son para testear).

Este modelo se conoce también como regresión lineal múltiple, y el ejemplo es apenas un ejercicio muy básico de una herramienta muy utilizada en la economía y estadística, que se combina con temas que verán más adelante en sus carreras. Los coeficientes β_i se conocen como regresores.

2.4. Linealización y problemas no lineales

En todos los ejemplos anteriores, la relación entre los parámetros era lineal, lo que dados los datos (x_i, y_i) con $i = 1, \ldots, N$ nos permitía plantear un sistema lineal de ecuaciones. Ello no significa que la relación entre x e y sea lineal; en el caso polinómico no lo era, pero sí era la relación entre los parámetros. ¿Qué sucede cuándo los parámetros no se relacionan linealmente? Como hemos visto en el último tiempo, muchos modelos de sobre la difusión de una epidemia poseen un comportamiento exponencial (por ejemplo, si miramos el número de contagiados por Covid-19 en algunos países en función de los días transcurridos de la epidemia). También muchos modelos sobre el crecimiento de colonias de bacterias o células, o la velocidad de las reacciones químicas, asumen un comportamiento exponencial.

2.4.1. Modelos exponenciales

En un modelo exponencial la relación entre las variables x e y viene dada por $y = f(x) = \alpha e^{\beta x}$, donde claramente la relación entre los parámetros α y β no es lineal. ¿Cómo trabajar con este modelo?

Notar que se puede linealizar manipulando algebraicamente las expresiones. En efecto, la expresión $y_i = \alpha e^{\beta x_i}$ se puede transformar en $\ln(y_i) = \ln(\alpha) + \beta x_i$, donde en esta segunda expresión la relación entre los parámetros es lineal (podemos renombrar $\ln(\alpha)$ como α'). Lo que se hizo fue transformar los datos: en vez de trabajar con el par (x_i, y_i) se trabaja con el par $(x_i, \ln(y_i))$ y se busca ajustar a una recta[6]. La pendiente de esa recta será el coeficiente β de la exponencial y el término independiente α' de la recta se relaciona con el factor que multiplica a la exponencial.

2.4.2. Leyes de potencia

Las leyes de potencia son usuales en modelos sobre redes y también en la descripción de fenómenos críticos. Algunos ejemplos son la ley de Stefan-Boltzmann, que relaciona la potencia emisiva de la radiación térmica de un objeto con la temperatura del mismo, o los fenómenos de criticalidad auto-organizada conocidos también como fenómenos de avalancha. También aparecen leyes de potencia en sistemas que tienen transición de fase, cuando se encuentran cerca de su punto crítico (por ejemplo, en el caso de materiales ferromagnéticos a muy baja temperatura).

En estos modelos, donde la relación esperada es de la forma $y_i = \alpha x_i^{\beta}$ se puede utilizar la misma idea que en los modelos exponenciales para llevar el problema de ajustar a una función potencial, donde la relación de los parámetros no es lineal entre sí, a un problema donde sí lo sea. En efecto, luego de aplicar el logaritmo a la expresión inicial, nos queda: $\ln(y_i) = \beta \ln(x_i) + \ln(\alpha)$

[6]Se asume implícitamente que los y_i tiene todos el mismo signo y este es positivo; si fueran todos negativos alcanza con trabajar con $(x_i, \ln(-y_i))$ y se llega a una expresión similar. Si los y_i no fuesen todos del mismo signo, entonces el ajuste a un modelo exponencial no tiene sentido porque la exponencial no cambia de signo.

y ahora la relación entre los datos en escala logarítmica es lineal, es decir, se busca una recta que relacione a los datos $(\ln(x_i), \ln(y_i))$ con $i = 1, \ldots, N$.

2.5. Problemas no lineales

En algunos casos el tipo de modelo al que se quiere ajustar no tiene una relación lineal entre los parámetros y no es posible linealizarlo con manipulaciones algebraicas exactas. Por ejemplo, es el caso de una superposición de exponenciales, donde la clase de funciones son de la forma $f(x) = e^{\alpha x} + e^{\beta x}$ con α y β desconocidos y a determinar. Otros ejemplos pueden ser el caso de un oscilador forzado o el fenómeno de caída libre pero considerando ahora el rozamiento dado por el aire.

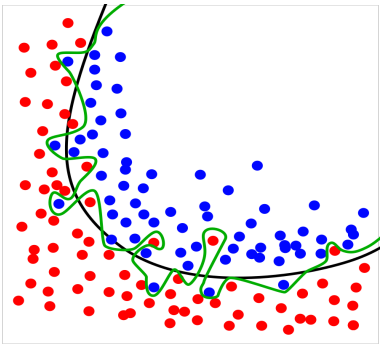
Para resolver un problema de mínimos cuadrados no lineales, es decir, donde los parámetros no se relacionan de forma lineal, es necesario aplicar algunas técnicas de análisis numérico. En particular, el algoritmo de Gauss-Newton se utiliza para esos fines. A *grosso modo*, consiste en dejar de pensar el problema de forma geométrica y atacarlo solamente en términos de optimización (de minimizar el error cuadrátrico). Para ello, se linealiza el error cuadrátrico usando la aproximación del desarrollo de Taylor de orden 1, y luego se aplica el método de mínimos cuadrados lineales. Esta estrategia de buscar linealizar el problema es también muy general en matemática, casi cualquier problema no lineal, se intenta llevar a un problema lineal, donde siempre hay más herramientas para su solución. En particular el tema de mínimos cuadrados no lineales es un tema más avanzado que se estudia en la asignatura Métodos Numéricos.

3. No todo es tan sencillo

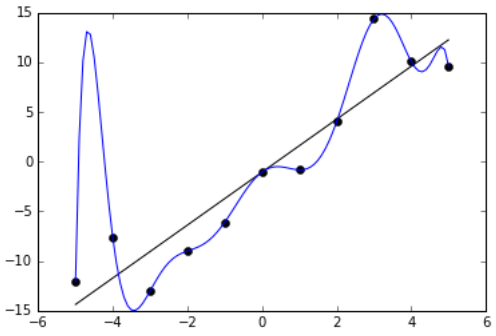
Como ya se mencionó con anterioridad, la cantidad de parámetros necesarios para describir la clase de funciones a buscar para vincular x con y depende fuertemente del conocimiento que se tenga *a priori* del problema o fenómeno. Es quizás, una de las partes más importantes al momento de aplicar el método de mínimos cuadrados a casos reales, puesto que se trata de elegir el modelo que está por detrás de nuestros datos (y no siempre es posible hacerlo con conocimiento, para lo que muchas veces se hace una exploración del comportamiento de los datos corriendo diferentes modelos).

3.1. Overfitting

Una tentación frecuente es aumentar la cantidad de parámetros para reducir el error cuadrátrico. Si bien tiene sentido buscar reducir el error, hay que tener cuidado de no sobre-ajustar nuestros datos (el fenómeno se conoce como *overfitting*). El caso de *overfitting* más sencillo de entender es cuando se sabe que la relación entre las variables es polinómica. Si se tienen N datos de la forma (x_i, y_i) sería posible buscar un polinomio de hasta grado $N - 1$ (que se representa con N parámetros) que pase por esos puntos. Un problema que surge cuando elegimos el grado de polinomio muy cercano a N es que cuando se agrega una observación más, es decir, un nuevo dato (x_{N+1}, y_{N+1}) es muy probable que el polinomio que hallamos no pase cerca de ese punto, y por tanto la supuesta bondad de nuestra estimación (con poco error o error casi cero) se rompe. Por ejemplo en el caso de la segunda imagen de la figura 4 se tiene un polinomio de grado 10, que pasa exactamente por todos los datos, por lo que el error cuadrático es 0, pero sin embargo está lejos de capturar la estructura de los datos que era esencialmente lineal.



(a) La curva verde sobre-ajusta, mientras que la negra no.



(b) Muchos parámetros no dicen más.

Figura 4: Dos ejemplos de *overfitting* de datos (imágenes tomadas de Wikipedia).

Un principio general que aplica es que, las buenas descripciones con poca cantidad de parámetros son mejores que aquellas que tienen muchos parámetros. Veamos cómo funciona este principio en nuestro ejemplo sobre el tiempo de ejecución del cálculo del determinante.

3.2. ¿Más parámetros es mejor?

En la sección anterior se halló el mejor polinomio de grado 3 que explica los datos de la tabla sobre los tiempos de ejecución del determinante. Ahora bien, el modelo de un polinomio completo de grado 3 tiene cuatro parámetros, ¿qué sucede si seguimos buscando polinomios de grado 3 pero disminuimos la cantidad de parámetros?

Comencemos con la opción más sencilla, buscando una función de la forma $f_1(x) = \alpha x^3$ y luego, se buscará un polinomio de la forma $f_2(x) = \alpha x^3 + \beta x^2$ (nuevamente los subíndices indican la cantidad de parámetros) .

Caso f_1 (un parámetro).

Si se imponen las condiciones $y_i = \alpha x_i^3$ para $i = 1, \ldots, 10$, y se plantea el sistema lineal de ecuaciones $AX = b$, este nos queda

$$A = \begin{pmatrix} 1 \\ 8 \\ 27 \\ 64 \\ 125 \\ 216 \\ 343 \\ 512 \\ 729 \\ 1000 \end{pmatrix}, \quad X = (\alpha), \quad b = \begin{pmatrix} 0{,}51 \\ 0{,}177 \\ 0{,}328 \\ 0{,}656 \\ 1{,}152 \\ 1{,}948 \\ 3{,}122 \\ 4{,}581 \\ 6{,}189 \\ 8{,}582 \end{pmatrix}.$$

Ahora se escriben las ecuaciones normales $A^TAX = A^Tb$ y se resuelve el sistema de ecuaciones (como tenemos un parámetro todo nos da de dimensión 1),

$$A^TA = (1978405), \qquad A^Tb = (17127{,}633),$$

obteniendo $\alpha = 0{,}009$. Por tanto, la función que busco es $f_1(x) = 0{,}009x^3$ y el error cuadrátrico que se comete es $r = 0{,}769$ y el error cuadrátrico medio nos da $0{,}077$.

Caso f_2 (dos parámetros).

Ahora se imponen las condiciones $y_i = \alpha x_i^3 + \beta x_i^2$ para $i = 1, \ldots, 10$, y se plantea el sistema lineal de ecuaciones $AX = b$, donde

$$A = \begin{pmatrix} 1 & 1 \\ 8 & 4 \\ 27 & 9 \\ 64 & 16 \\ 125 & 25 \\ 216 & 36 \\ 343 & 49 \\ 512 & 64 \\ 729 & 81 \\ 1000 & 100 \end{pmatrix}, \quad X = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad b = \begin{pmatrix} 0{,}51 \\ 0{,}177 \\ 0{,}328 \\ 0{,}656 \\ 1{,}152 \\ 1{,}948 \\ 3{,}122 \\ 4{,}581 \\ 6{,}189 \\ 8{,}582 \end{pmatrix}.$$

Notar que en la primer columna de A van los términos x_i^3 y en la segunda columnas los correspondientes a x_i^2. El término independiente b son los valores de y_i y es siempre el mismo (no cambia a pesar de que cambiemos la cantidad de parámetros). Ahora se escriben las ecuaciones normales $A^T AX = A^T b$ y se resuelve el sistema de ecuaciones,

$$A^T A = \begin{pmatrix} 1978405 & 220825 \\ 220825 & 25333 \end{pmatrix}, \qquad A^T b = \begin{pmatrix} 17127{,}633 \\ 1919{,}265 \end{pmatrix},$$

obteniendo $\alpha = 0{,}007$, $\beta = 0{,}011$. Por lo tanto la función que se obtiene es $f_2(x) = 0{,}007x^3 + 0{,}011x^2$. El error cuadrátrico es $r = 0{,}801$ y el error cuadrático medio es 0,080.

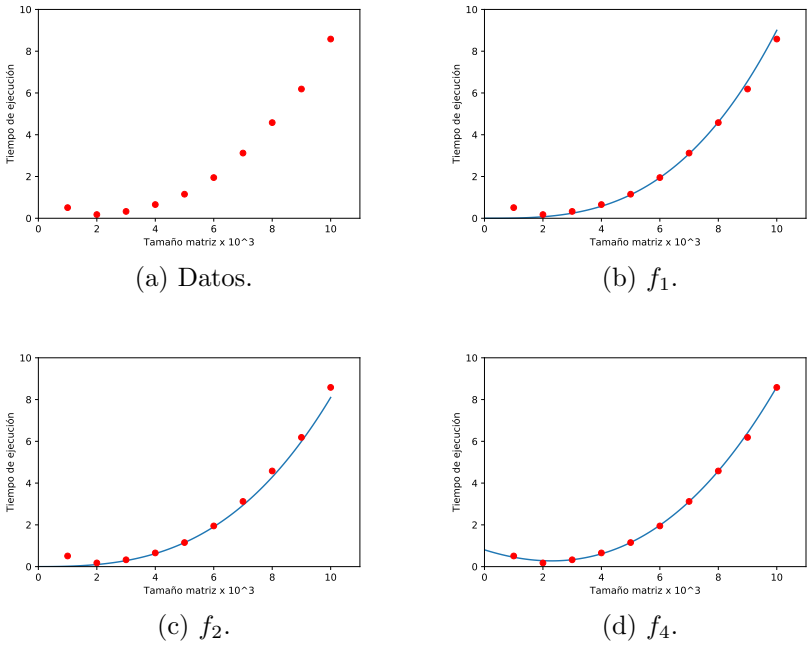
Resumen.

De las tres aproximaciones a los datos que se hicieron con 1, 2 y 4 parámetros, ¿cuál es mejor? Si nos guiamos solamente por el error cuadrátrico, se puede concluir que aproximar con 4 parámetros es mejor que con 1 y 2, y en este caso a su vez, aproximar con 1 parámetro es mejor que hacerlo con 2 (ver cuadro). Fue necesario agregar 3 parámetros para reducir el error a una tercera parte (lo que es una mejora), pero sin embargo, el error sigue siendo del mismo orden de magnitud que con un solo parámetro.

Función	Error cuadrático
f_1	0.769
f_2	0.801
f_4	0.254

Observando las gráficas de la figura 5 se observa que todas son similares, si se quiere la principal diferencia entre f_1 y f_4 se encuentra en los primeros datos, para tamaños pequeños de la matrices, pero no en el comportamiento para valores grandes. Esto se refleja aún más cuando se realiza el cálculo del determinante para algún valor que no está en la escala, por ejemplo, para una matriz de tamaño 20000 (a los efectos de nuestras cuentas sería $x = 20$). Allí f_1 estima el tiempo de ejecución en 72 seg, f_2 lo hace en 56,805 seg y f_4 en 60,4 seg, mientras que el tiempo de ejecución real fue de 81,814 seg. La estimación con un parámetro es la que comete menor error al aproximar un nuevo valor que no estaba en nuestra escala. ¿Por qué? Porque no sobre ajustó, sino que capturó la estructura de los datos. Además es razonable que cuando el tamaño de la matriz sea grande, el término que dominará es el de grado 3 y nuestra estimación con un parámetro es la que tiene coeficiente de grado 3 con mayor valor absoluto.

La moraleja es que la performance de un método de ajuste no debe basarse únicamente en el error cometido, sino que, por vago que parezca, tener en cuenta si captura la esencia de los



(a) Datos. (b) f_1.

(c) f_2. (d) f_4.

Figura 5: En cada imagen se tienen los ajustes.

datos de acuerdo a lo que conozco del experimiento o fenómeno, más que un ajuste perfecto en términos del error. En muchas ocasiones también es necesario contemplar la complejidad y los tiempos de resolución. Resolver las ecuaciones normales para k parámetros requiere de al menos k^2 operaciones, por lo que si k es grande esta cuestión se transforma también en una limitante.

3.3. Penalización a la cantidad de parámetros

El método de mínimos cuadrados tiene también la contra-indicación que en general nos da valores no nulos para todos los parámetros. Si para buscar la función que mejor ajusta a datos obtenidos de un experimento de caída libre le impongo 20 parámetros (por ejemplo en un polinomio de grado 19), mínimos cuadrados me dará en general, 20 parámetros no nulos que describen mi fenómeno. A veces, usando el conocimiento previo del fenómeno o alguna heurística, se puede reducir la cantidad de parámetros, pero cuando no podemos hacerlo, ¿cómo saber cuáles son significativos y cuáles deberían ser nulos (a pesar de que mínimos cuadrados no los halle como nulos)?

Una forma de corregir el exceso de parámetros es utilizar un algoritmo que se llama Lasso. Este algoritmo no busca minimizar solamente el error cuadrátrico (como hace mínimos cuadrados) sino que lo combina con otro error que penaliza tener muchos coeficientes no nulos. El algoritmo Lasso busca el valor de Y donde se alcanza el siguiente mínimo:

$$\min_{Y \in \mathbb{R}^n} \{\|AY - b\|_2^2 + \lambda \|Y\|_1\}, \tag{7}$$

siendo $\lambda \geq 0$ un factor de penalización que debe ser definido por el usuario y $\|Y\|_1 = |Y_1| + \cdots + |Y_n|$ donde $Y = (Y_1, \ldots, Y_n)$. Notar que el primer término es exactamente el de mínimos cuadrados y el otro sumando que se agrega crece cuando hay muchos coeficientes no nulos. El valor de λ modula la penalización: un λ grande penaliza mucho tener coeficientes no nulos y se buscará una solución que tenga pocos parámetros no nulos; en el otro extremo un λ pequeño no impondrá demasiado esa condición. Para $\lambda = 0$ se recupera mínimos cuadrados.

El problema es que ya no se puede resolver el mínimo de Lasso usando la proyección ortogonal y por tanto su solución es bastante más complicada. Notar también que la función que se busca minimizar deja de ser diferenciable por el efecto del factor $\|Y\|_1$. Para hallar el mínimo se deben combinar varios métodos numéricos o buscar una regularización de la función.