Apunte 3:

Ajustes

- 1. Moneda de probabilidades desconocidas.
  - 1.1 Medición todas cara.
  - 1.2 Balance cara-ceca.
  - 1.3 Medición al azar.
- 2 Aproximación Gaussiana.
  - 2.1 Una variable.
  - 2.2 Distribución Normal Multivariada.
    - 2.2.1 Variables independientes.
    - 2.2.2 Variables correlacionadas.
  - 2.3 Desarrollo en el máximo.
- 3. Ajuste de curvas - Cuadrados mínimos.
- 4. Cuidado ! : incluir el error de los datos.
- 5. Recta.

Al proceso de calcular la distribución de los parámetros del modelo, dado el resultado de una medición, lo llamamos *ajuste*.
Para los físicos, un trabajo glorioso termina dibujando una curva hermosamente ajustada por una teoría, pero es bueno saber que la noción de ajuste no se limita a ajustar una curva.

Por eso vamos a empezar con un ejemplo que no sea una curva.

La única matemática que necesitamos es que

$$\boxed{P(\theta|D) \propto P(D|\theta)\, P(\theta)}$$

1. Moneda de probabilidades desconocidas

Tengo una moneda que no sé de dónde salió y es posible que esté cargada, es decir que la probabilidad de que salga cara al tirarla, no es necesariamente a la probabilidad de que salga ceca.

Vamos a modelar el problema con un parámetro θ:

$$\theta \equiv \text{probabilidad de que salga cara}$$

Resulta inmediato calcular la *likelihood*. Si tiro una vez la moneda y sale x:

$$P(x|\theta) = \begin{cases} \theta & x = 1 \quad (cara) \\ 1-\theta & x = 0 \quad (ceca) \end{cases}$$

Sólo queda especificar la *prior*. Voy a proponer una prior uniforme, que es una hipótesis *babieca*, ya que significa que nunca en la vida vi una moneda.

Nuestra experiencia cotidiana con monedas tiene mucha más información que eso, pero resulta más simple comenzar siendo un poco sonzo.

Para calcular algo, necesito resultados. Por eso propongo diferentes resultados de mediciones.

1.1 Medición todas cara

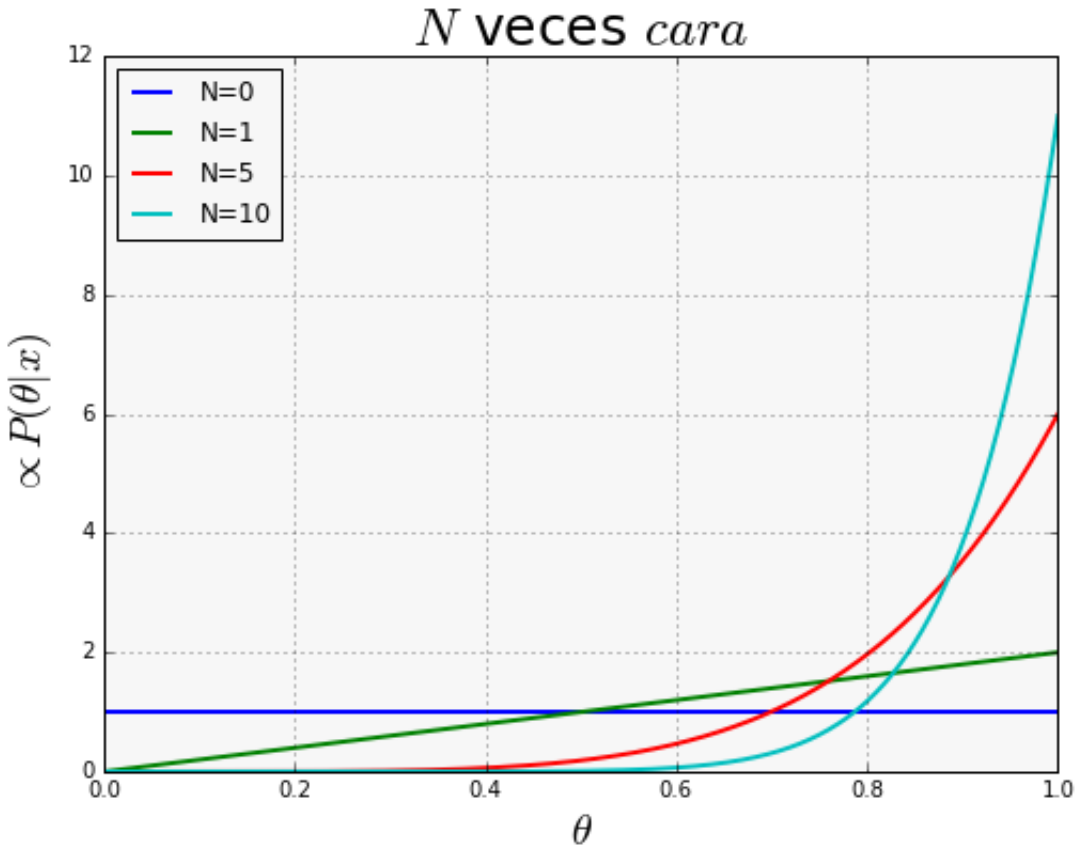
Supongamos ahora que los datos medidos son todas caras. Si mido N tiradas,
resulta que

$$P\left(\theta \mid \{x\}_N\right) = A\,\theta^N\,P(\theta)$$

y no es difícil obtener el factor de proporcionalidad que sale por normalización:

$$A = N + 1$$

Graficamos.



Vemos que

- antes de medir (curva azul) me da lo mismo cualquier θ
- Si mido una vez cara, es más probable que la próxima vuelva a ser cara.
  Además la probabilidad de que siempre salga ceca, se ha vuelto cero.
- A medida de que salga muchas veces cara, la probabilidad máxima es que vuelva a salir cara, y cada vez se concentra más la probabilidad cerca de $\theta = 1$.

1.2 Balance cara-ceca

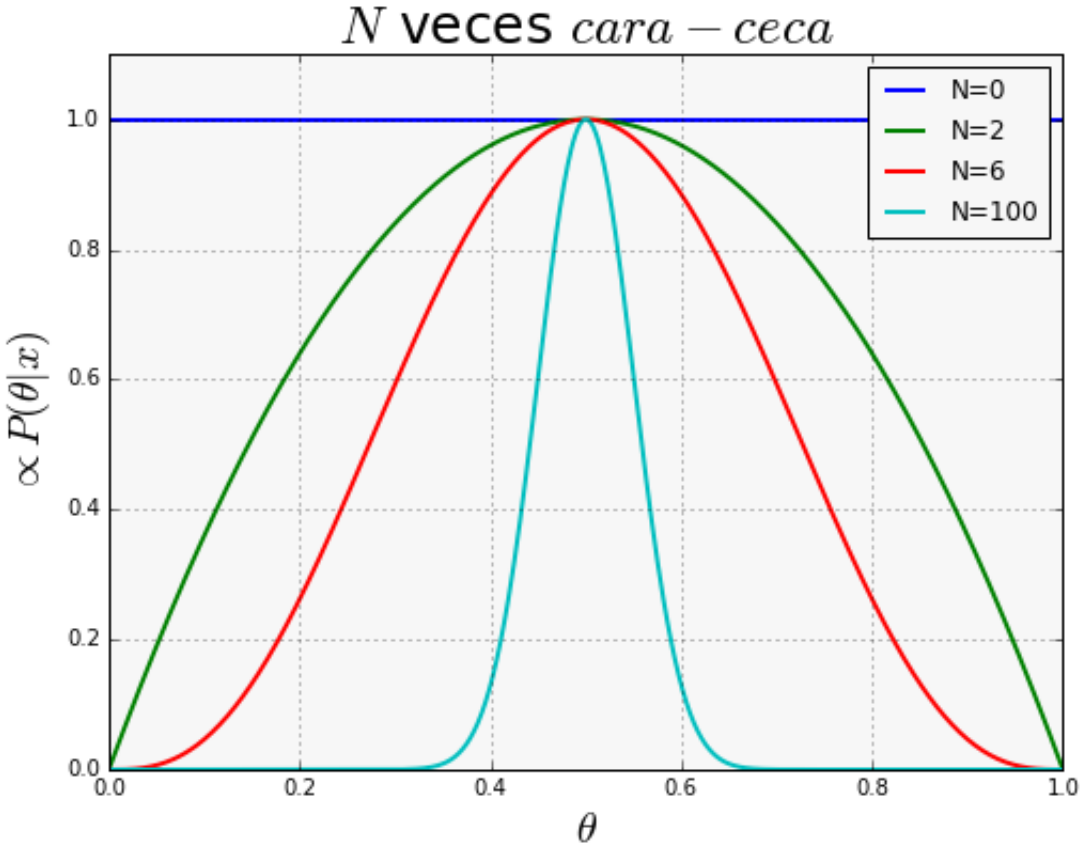
Si sale una vez cara y una vez ceca alternadamente, cada dos veces aparece el factor $\theta(1-\theta)$ y resulta (N par):

$$P\left(\theta|\{x\}\right) = A\,\theta^{N/2}(1-\theta)^{N/2}\,P(\theta)$$

El factor de normalización es muy aburrido de calcular, pero para graficar la forma, podemos hacer que el máximo de la función tenga siempre el mismo valor. Para eso basta hacer

$$A = 2^N$$

Veamos lo que da:

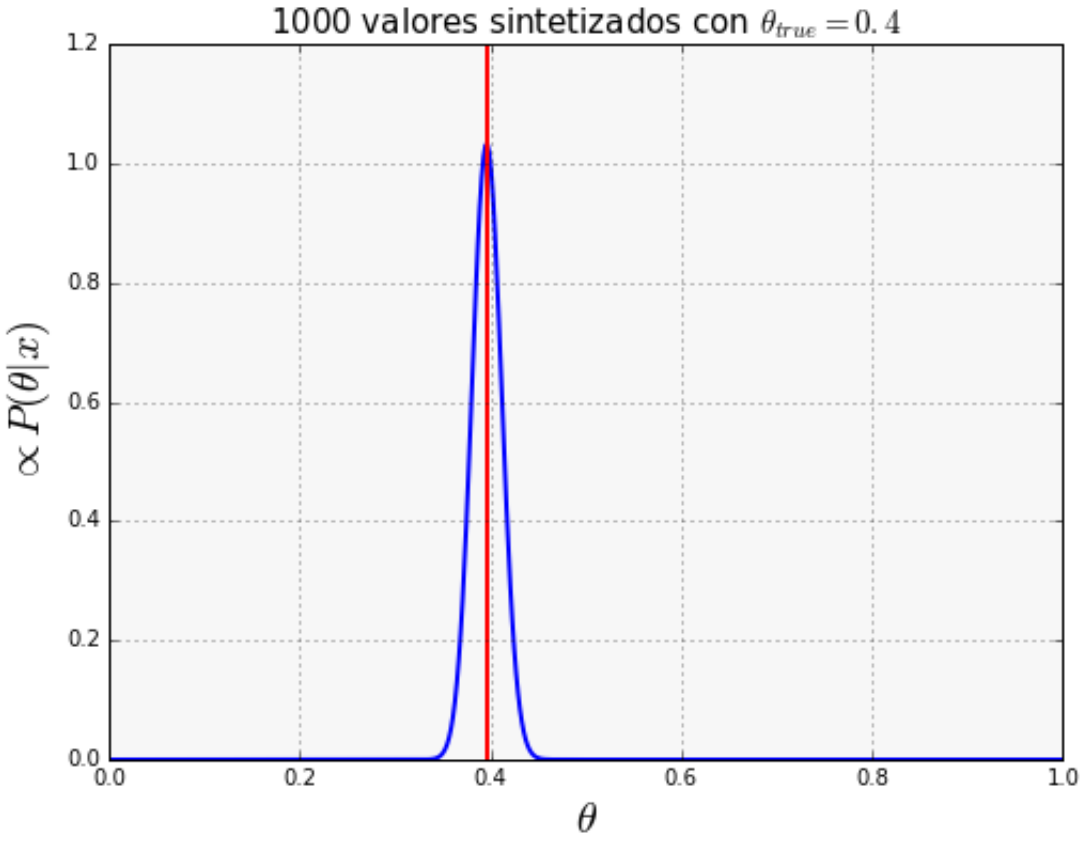


- con dos datos medidos (una cara y una ceca) se hacen cero las probabilidades extremas. Es una parábola centrada en $\theta = 0.5$
- cuanto más mido, más se concentra la probabilidad de θ en el centro.

1.3 Medición al azar

Vamos a sintetizar datos al azar para un $\theta_{true} = 0.4$

Evito explícitamente hacer cuentas analíticas, así me parece más instructivo.



La conclusión es que aún con 1000 valores, no da exactamente 0.4, pero al menos está dentro de lo parece plausible.

2 Aproximación Gaussiana

Como vimos, muchas veces el resultado de la estimación de un parámetro es una forma funcional con un máximo bien localizado. Con excepción de las curvas del experimento donde siempre sale cara (sec. 1.1) todas tienden a ser formas que alcanzaría caracterizarlas con una localización y una dispersión.

Formalmente, podemos hacer la susticución:

$$P(\theta|D) \propto e^{-\chi^2_{(\theta)}}$$

- Sólo estoy sustituyendo la función $P(\theta|D)$ por otra
- La otra se llama $\chi^2(\theta)$, y se le dice *chi-cuadrado* pero no hay que pensarlo como el cuadrado de nada. Es una notación tradicional, que mantenemos.

2.1 Una variable

Si $P(\theta|D)$ tiene un máximo en $\hat{\theta}$, podemos encontrarlo con la condición:

$$\frac{\partial \chi^2}{\partial \theta}\Bigg]_{\hat{\theta}} = 0$$

y alrededor de ese punto, desarrollamos en serie de Taylor hasta el segundo orden (que es el primero que

varía):

$$\chi^2(\theta) \approx \chi^2(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \left[\frac{\partial^2 \chi^2}{\partial \theta^2}\right]_{\hat{\theta}}$$

Esto nos dice que la ditribución alrededor del máximo es de tipo gaussiana, con una dispersión de

$$\frac{1}{\sigma^2} = \left[\frac{\partial^2 \chi^2}{\partial \theta^2}\right]_{\hat{\theta}}$$

De esta manera podemos encontrar el valor máximo $\hat{\theta}$ y la dispersión del resultado. Basta calcular las dos primeras derivadas de χ^2

Pero qué pasa si tengo varias variables (si θ tiene muchas dimensiones)?

Primero veamos qué es una gaussiana en muchas dimensiones.

2.2 Distribución Normal Multivariada

Lo más sencillo sería multiplicar las distribuciones gaussianas de dos variables independientes, y así vamos a comenzar.

2.2.1 Variables independientes

Tenemos las variables independientes x e y y queremos expresar la distribución conjunta. Es sencillo, porque son independientes: se multiplican las distribuciones.

$$\begin{aligned} P(x,y) &= P(x)\,P(y) \\ &= G(x,\mu_x,\sigma_x)\,G(y,\mu_y,\sigma_y) \\ &= \frac{1}{\sqrt{2\pi\sigma_x^2}} \;\; \frac{1}{\sqrt{2\pi\sigma_y^2}} \;\; e^{-\frac{1}{2\sigma_x^2}(x-\mu_x)^2 - \frac{1}{2\sigma_y^2}(y-\mu_y)^2} \end{aligned}$$

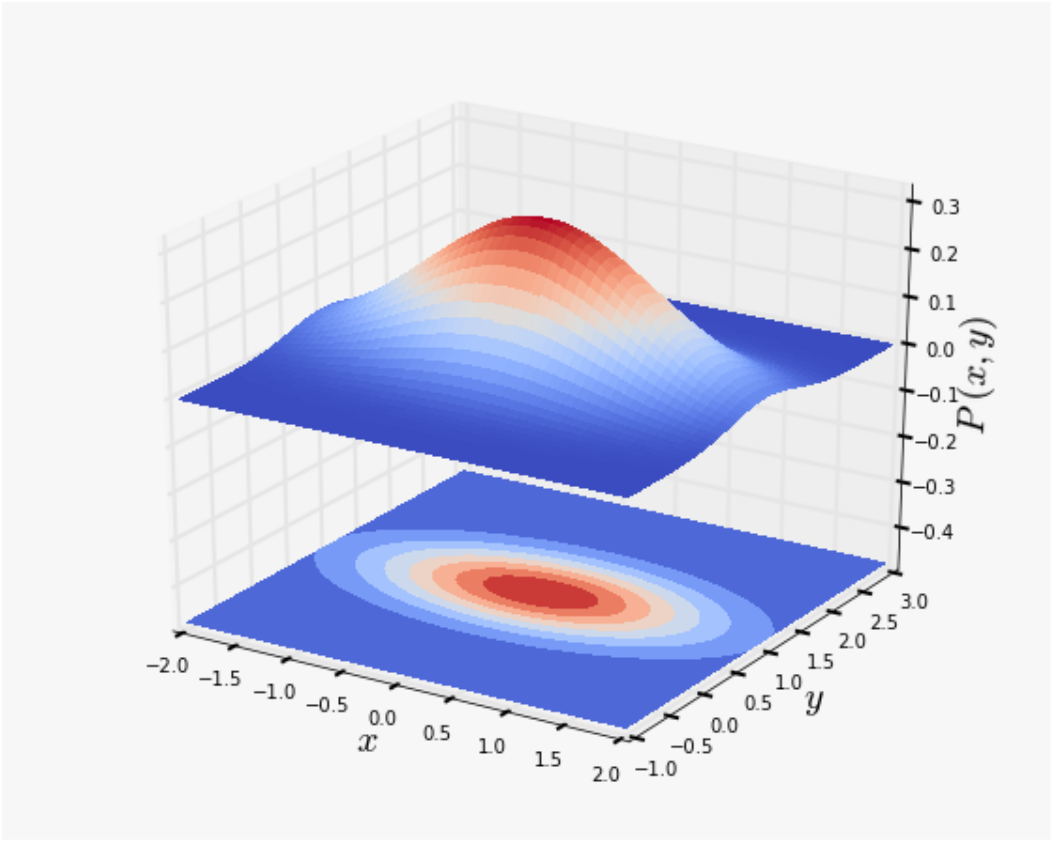
a esta última expresión, la podemos pensar en términos de una matriz diagonal $\mathbf{\Sigma}$ definida como:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

$$P(x,y) = Multivar\,(x,y,\mu,\mathbf{\Sigma}) = \frac{1}{\left(\sqrt{2\pi}\right)^2 \sqrt{|\mathbf{\Sigma}|}} \;\; e^{-\frac{1}{2}\,(x-\mu_x\,;\,y-\mu_y)\cdot\mathbf{\Sigma}^{-1}\cdot\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}}$$

y así queda una forma muy compacta de una distribución de dos variables. Esta distribución se llama *normal multivariada* y se puede generalizar fácilmente a muchas variables.

El hecho de que sean independientes está reflejado en que la distribución conjunta se construye multiplicando la distribución de cada variable por separado. Visualmente, esto se ve en que la forma resultante está alineada según los ejes.



2.2.2 Variables correlacionadas

El ejemplo anterior es fácilmente generalizable, si liberamos la restricción que la matriz $\mathbf{\Sigma}$ sea diagonal. Lo vamos a relajar, pero no tanto: vamos a necesitar que la matriz sea simétrica y (semi) definida positiva.

Eso nos permitirá escribir en el exponente expresiones de segundo orden en x e y.

Se puede demostrar que los elementos de nuestra matriz representan la covariancia entre las variables:

$$\mathbf{\Sigma}_{ij} = Cov(x_i, x_j) = \overline{(x_i - \overline{x_i})\,(x_j - \overline{x_j})}$$

La covarianza es una medida de cuánto cambian a la vez dos variables. Si las variables son independientes, la covarianza es nula. La inversa no es estrictamente cierta, pero para los casos con los que nos vamos a encontrar en el barrio, vamos a usarlo como si así lo fuera.

Los elementos de la diagonal, son las correlaciones de una variable consigo misma y la definición coincide con la varianza.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{pmatrix}$$

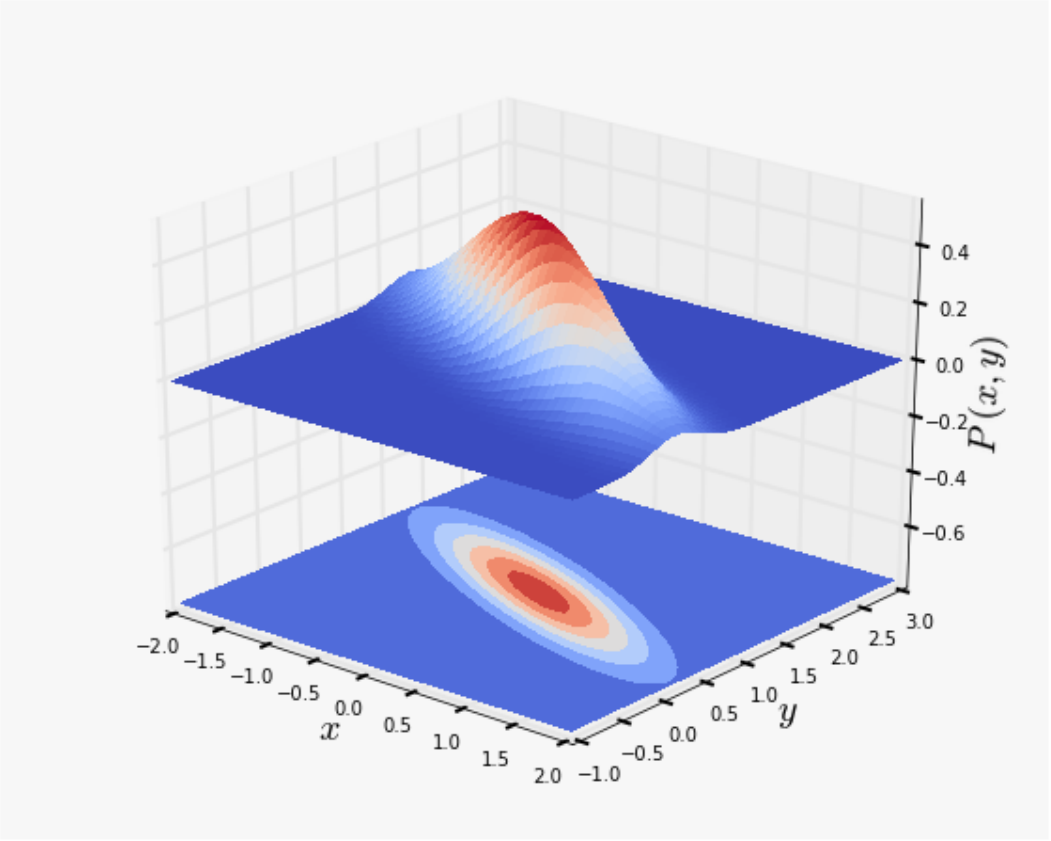
Aquí escribimos una versión *normalizada* de la covariancia, que es el coeficiente de correlación.

$$\rho_{ij} = \frac{Cov(x_i, x_j)}{\sigma_i \sigma_j}$$

- Gracias a la normalización, ρ varía entre -1 y $+1$.
- En los extremos, las variables están tan correlacionadas que podemos saber el valor de una de ellas, conociendo la otra.
- Si $\rho = 0$ las variables son independientes (salvo casos *muy* raros).
- Si la correlación es positiva, ambas variables aumentan o disminuyen juntas, si la correlación es negativa, cuando una variable aumenta, la otra disminuye.

En el caso general, la distribución normal multivariada no tiene nada de raro, salvo que el máximo no se encuentra orientado según los ejes. Se puede demostrar fácilmente que basta una rotación en el espacio x, y para tener nuevas variables independientes.

Para ilustración, nada mejor que un dibujito. Reproducimos el caso anterior pero con una correlación.



La expresión general que usaremos es:

$$P(\mathbf{x}) = Multivar\,(\mathbf{x}, \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^N\ |\mathbf{\Sigma}|}}\ e^{-\frac{1}{2}\,(\mathbf{x}-\mu)^T \cdot \mathbf{\Sigma}^{-1} \cdot (\mathbf{x}-\mu)}$$

2.3 Desarrollo en el máximo

Ya entendemos cómo se expresa la gaussiana multivariada. Podemos generalizar el resultado de una variable.

Primero encontramos el máximo pidiendo que todas las derivadas parciales sean nulas:

$$\nabla\chi^2(\theta)\big|_{\hat{\theta}} = 0$$

y el desarrollo en serie alrededor de ese punto es:

$$\chi^2(\theta) \approx \chi^2(\hat{\theta}) + \frac{1}{2}\,(\theta-\hat{\theta})^T\,\mathbf{H}\,(\theta-\hat{\theta})$$

donde $\mathbf{H}$ es el *Hessiano*: la matriz de las derivadas segundas. Y la analogía con la distribución multivariada, es inmediata

$$\mathbf{H} = \left[\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right]_{\hat{\theta}} = \mathbf{\Sigma}^{-1}$$

De esta manera, conociendo la función $\chi^2(\theta)$, puedo calcular el centro de la distribución y la matriz de correlaciones que resulta de aproximarla por una distribución normal.

3. Ajuste de curvas - Cuadrados mínimos

Mi problema ahora se puede plantear así:

- tengo una variable de control x que la puedo colocar en varios valores x_i
  (La variable puede ser una medición, pero para el formalismo necesitamos que no tenga error asociado)
- en cada una de las mediciones mido una variable y_i que responde a un modelo donde y es una función de x con un ruido gaussiano de dispersión σ_i

$$y_i = f(x_i,\theta) + \epsilon_i$$

Si colocamos una prior uniforme, resulta que

$$\begin{aligned} P(\theta|D) &\propto P(D|\theta) \\ &\propto \prod_i P(y_i|\,x_i\;\theta) \\ &\propto \prod_i \exp -\frac{\epsilon_i^2}{2\sigma^2} \\ &\propto exp - \sum_i \frac{\left(y_i - f(x_i,\theta)\right)^2}{2\sigma_i^2} \end{aligned}$$

de donde resulta que

$$\boxed{\chi^2(\theta) = \textstyle\sum_i \frac{\left(y_i - f(x_i,\theta)\right)^2}{2\sigma_i^2}}$$

Encontrar los calores óptimos de θ equivale a encontrar los máximos de $P(\theta|D)$ y eso equivale a minimizar la función $\chi^2(\theta)$.

Este método se llama de los **cuadrados mínimos** y es el que se usa habitualmente para ajustar funciones a los datos.

Calculando el Hessiano, podemos obtener los errores y correlaciones del ajuste.

La derivación que usamos evidencia las condiciones de validez del método:

- la variable x no tiene error
- los errores de y son gaussianos
- la prior es uniforme
- la distribución resultante $P(\theta)$ puede aproximarse por una gaussiana

Hay muchos casos en los que alguna de estas hipótesis no es válida, pero mucha gente pasa una exitosa vida científica sin necesitar salir de estas condiciones. Por eso es bueno registrar ahora, para saber volver a los principios cuando las cosas fallan.

Si el espacio de parámetros tiene muchas dimensiones, encontrar el mínimo es todo un trabajo. Hay muchos métodos de optimización entre los que vale la pena mencionar el de particiones, el de Newton-Raphson y el de Levenberg-Marquardt. En esta etapa de la vida no nos vamos a preocupar por cómo funcionan esos métodos y vamos a utilizar los programas que otros escribieron con tanto arte.

4. Cuidado ! : incluir el error de los datos

Si miramos la expresión del χ^2, resulta que en el caso en que todos los y_i tengan la misma dispersión ($\sigma_i \equiv \sigma$), esa dispersión sale como factor común.

El valor de esa dispersión no es necesario para encontrar los valores óptimos $\hat{\theta}$, porque al igualar a cero las derivadas, desaparece.

Pero cuando calculo la dispersión de los parámetros, es fácil ver que todas quedan proporcionales a la dispersión de los datos.

Conclusión:
No necesito conocer σ para calcular los parámetros óptimos. Los errores y correlaciones son todos proporcionales. Puedo calcular suponiendo que $\sigma = 1$ y recién al final del cálculo, multiplico las dispersiones por la dispersión de los datos.

Eso es lo que hacen muchos programas en el cálculo de errores. A veces al resultado del programa hay que multiplicarlo por σ para tener la dispersión correcta de los parámetros. No hay que creerles (o hay que hacer unas pruebas) al resultado de los programas si no tengo acceso a la documentación detallada.

5. Recta

Muchos ajustes que hacemos se hacen modelando los datos por una recta.

$$f(x, \theta) = a + bx$$

Es un caso muy lindo porque los resultados son analíticos y son muy interesantes para interpretar. Para que sean más elegantes todavía, vamos a suponer que todos los datos y_i tienen la misma dispersión σ.

En ese caso resulta que los valores óptimos son:

$$\hat{b} = \frac{Cov(x,y)}{Var(x)}$$
$$\hat{a} = \overline{y} - \hat{b}\,\overline{x}$$

y los errores:

$$\sigma_b^2 = \frac{1}{NVar(x)}\ \sigma^2$$
$$\sigma_a^2 = \frac{\overline{x^2}}{NVar(x)}\ \sigma^2 \qquad Cov(a,b) = -\frac{\overline{x}}{NVar(x)}\ \sigma^2$$

Lo más interesante es el análisis de los errores:

- sólo dependen de la variable x, es decir, de la estrategia de medición, no del resultado de las mediciones
- todos los errores se achican según la $\sqrt{N}$. Medir más siempre conviene.
- para achicar el error de la pendiente, necesito que los valores de x estén extendidos, así tienen variancia grande.
- para achicar el error en la ordenada al origen, conviene medir cerca del origen que es donde achico los valores de $\overline{x^2}$.
- los valores tienen una correlación si no están centrados en el origen. Es fácil de ver: si los datos no están cerca del origen, una disminución de la pendiente, va a hacer que la ordenada al origen sea más grande.

La correlación entre los parámetros es fundamental a la hora de utilizar el ajuste en una calibración.

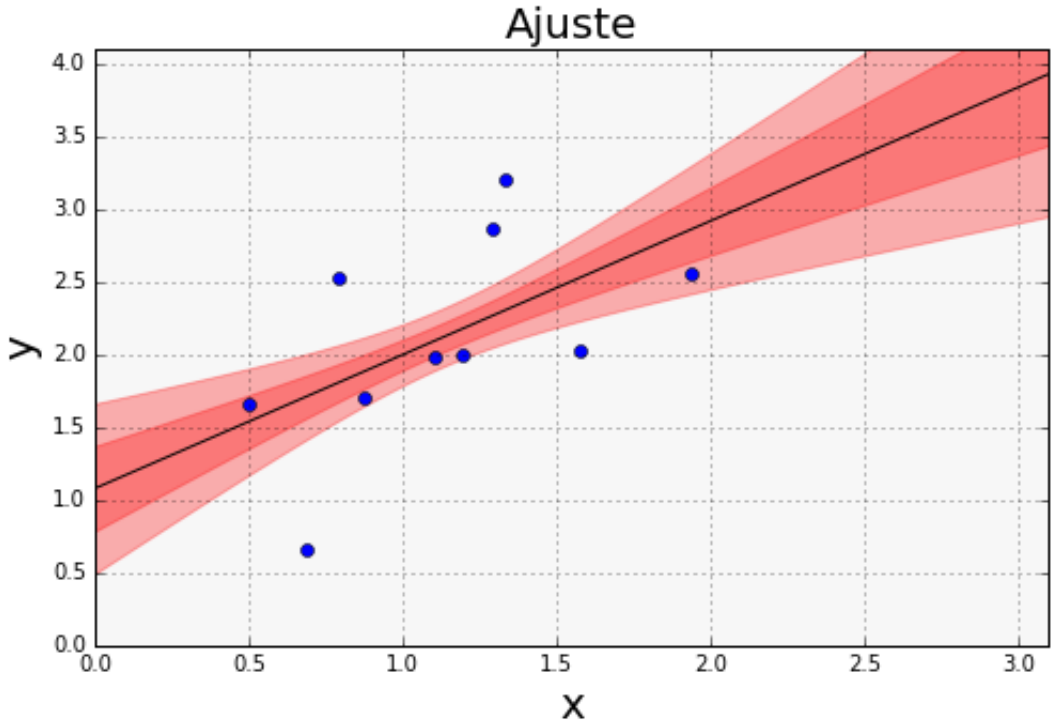
Supongamos que ya obtuve los valores óptimos $\hat{a}$ y $\hat{b}$ y quiero calcular el valor de y dado x. Si uso la fórmula de propagación de errores:

$$y = \hat{a} + \hat{b}x$$
$$\sigma_y^2 = \sigma_a^2 + x^2\sigma_b^2 \leftarrow Mal!$$

va a dar todo muy feo. La manera correcta es usar la versión completa, que incluye la covarianza de los parámetros:

$$\sigma_y^2 = \sigma_a^2 + x^2\sigma_b^2 + 2x\ Cov(a,b)$$

y esto me da los clásicos gráficos de *trompetitas*



Aquí graficamos la región de uno y dos sigmas alrededor del ajuste. Cool.

(Fin de la segunda parte de las clases de errores. Willy 2017.)