1 The basics

1.1 加法规则和乘法规则

$$prob(X|I) + prob(\bar{X}|I) = 1 \tag{1.1}$$

$$prob(X, Y|I) = prob(X|Y, I) \times prob(Y|I) \tag{1.2}$$

成立条件?

1.2 贝叶斯理论和边缘化

$$prob(X|Y, I) = \frac{prob(Y|X, I) \times prob(X|I)}{prob(Y|I)} \tag{1.3}$$

$$prob(X|I) = \int_{-\infty}^{+\infty} prob(X, Y|I)dY \tag{1.4}$$

分母上的 $P(Y|I)$ 要对所有可能的 Y 进行积分，很难计算，所以这个式子只用分子，得到没有归一化的后验 pdf。

$$prob(X|Y, I) \sim prob(Y|X, I) \times prob(X|I) \tag{1.5}$$

边缘化在不关心的参数存在时很有用

1.3 关于 pdf

概率密度函数 f(x) 是一根山峰线，与 x 轴围成的面积为 1，(x 轴代表组距，y 轴代表频率乘以组距，取组距趋于 0 的极限，直方图就变成一根山峰线，即概率密度函数。也可以 x 轴表示频率，y 轴代表取此频率的概率，则山峰线围的面积也代表总频率等于 1）概率密度函数对 dx 进行积分，得到面积，面积的名字叫做分布函数, 分布函数代表的面积不是整个面积，分布函数 $F(x) = \int_{-\infty}^{x} f(t)dt$。
这些是对连续型随机变量的解释, 与概率密度函数相同含义，在离散型随机变量中，这个函数叫做概率质量函数。

用到的函数:

mean(): 分布的均值
median(): 分布的中值
pdf(x): 概率密度函数在 x 点的值
Rvs (size=num_pts): 生成 pdf 的 num_pts 随机值
interval(alpha): 包含 alpha 百分比的分布范围的端点 (置信区间)

几种常见 pdf 的形状:

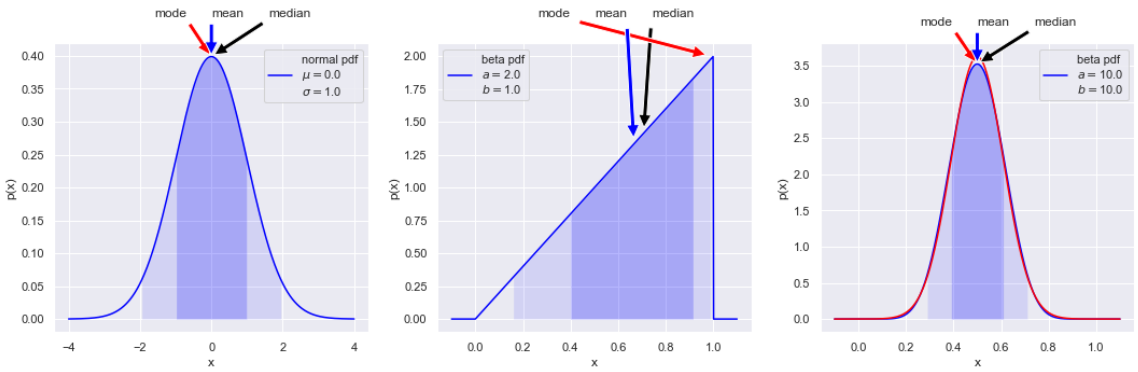

图 1.1: 从左至右分别为正态分布 ($\mu = 0.0, \sigma = 1.0$)、beta 分布 ($\alpha = 2.0, \beta = 1.0$)，第三个图中一起画了 beta 分布 ($\alpha = \beta = 10.0$) 和正态分布 ($\mu = 0.5, \sigma = 0.11$)

在图中用红色箭头指出了众数，蓝色箭头指出平均值，黑色箭头指出中位数。

intro 程序中比较了**频率派**和**贝叶斯方法**，程序是 Sivia 的书中第 2.3 节 Example2 的扩展。**考虑正态分布的均值和方差的估计**问题。正态分布

$$p(x_k|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{1.6}$$

通常用于作为理论模型来描述与实验数据相关的噪声。假设有一系列 M 测量值 $D \equiv X_k = (x_1 \ldots x_M)$, 样本服从一个正态分布 $N(\mu, \sigma^2)$, 现在想知道的是参数 μ 和 σ。频率论的方法：最大似然方法, 贝叶斯方法：计算模型参数的后验分布。这里，bayesTALENT into 中使用 Python 生成一些数据，演示了解决该问题的两种方法。

首先使用按照真实值生成满足高斯分布的数据，然后画了数据的散点图和条形图。高斯参数估计的频率派：

从经典的频率最大似然法开始，一次测量的概率 D_i 的值为 x_i 的概率是由 $p(x_i \mid \mu, \sigma)$ 给出的

$$p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(x_i-\mu)^2}{2\sigma^2}\right], \tag{1.7}$$

通过计算每个数据点的概率乘积来构造似然函数:

$$\mathcal{L}(D|\mu,\sigma)=\prod_{i=1}^{M}p(x_i\ |\mu,\sigma), \tag{1.8}$$

目的是找到 μ_0,σ_0，使 likelihood(或对数 likelihood) 最大化。对于这个简单的问题，最大化可以用解析的方法计算, 即通过 $\left(\frac{\partial\log\mathcal{L}}{\partial\mu}\right)|_{\mu_0\sigma_0}=\left(\frac{\partial\log\mathcal{L}}{\partial\sigma}\right)|_{\mu_0\sigma_0}=0$ 这将导致以下真实参数的最大似然估计:

$$\mu_0=\frac{1}{M}\sum_{i=1}^{M}x_i, \tag{1.9}$$

$$\sigma_0^2=\frac{1}{M}\sum_{i=1}^{M}(x_i-\mu_0)^2, \tag{1.10}$$

程序用这两个表达式计算出了 μ_0 和 σ_0，与真实值非常接近。10，10.06；1，0.89（基于 100 次实验）。

Bayes 方法中，首先定义了先验后验和似然函数，取先验为平坦的 1，则后验

$$p(\mu,\sigma|D,I)\propto\mathcal{L}(D|\mu,\sigma), \tag{1.11}$$

由此画出后验分布的 corner 图，并用 MCMC 进行抽样取点 (MCMC 方法是用来在概率空间，通过随机采样估算兴趣参数的后验分布。)，其中蓝线表示真实值，红线代表频率派最大似然法算出的值，对角线上的图是边缘化后的图（将另一个参数积分掉），左下角的图显示了两个模型参数的联合概率分布，线圈代表等置信区间，在圈上的点的 σ 和 μ 的取值所处的置信区间相等。图中的点就是使用 MCMC 方法进行抽样选出的点，中间最密集的地方所代表的值与频率派最大似然法得出的结果是一致的。数据虽然是按照高斯分布生成的，但是却是随机取样，所以置信区间左右不一定是对称的，数据量不够多的时候会有小的摇摆。但是这个偏差，跟下面非对称的 beta 分布对比可以看出 β 分布的上下误差的绝对值相差了 0.02, 而这里是 0 和 0.01，再下面一个图差的更多。

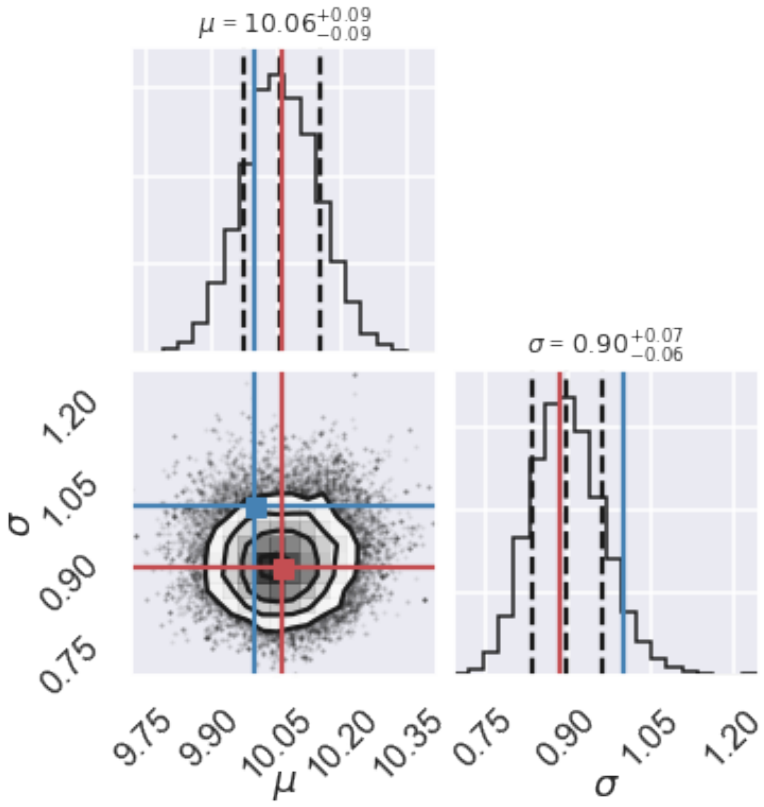

图 1.2:

除此之外，在 exploring pdf 中，还展示了其他几种 pdf 的分布 corner 图，非对称 beta 分布：

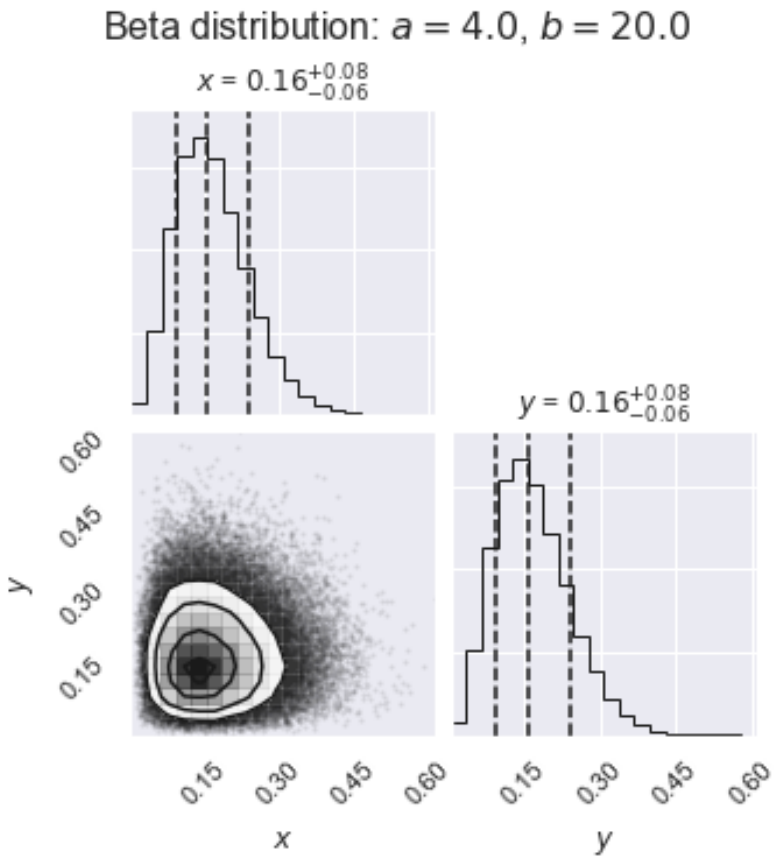

图 1.3: $\alpha = 4, \beta = 20$

多参数的 corner 图:

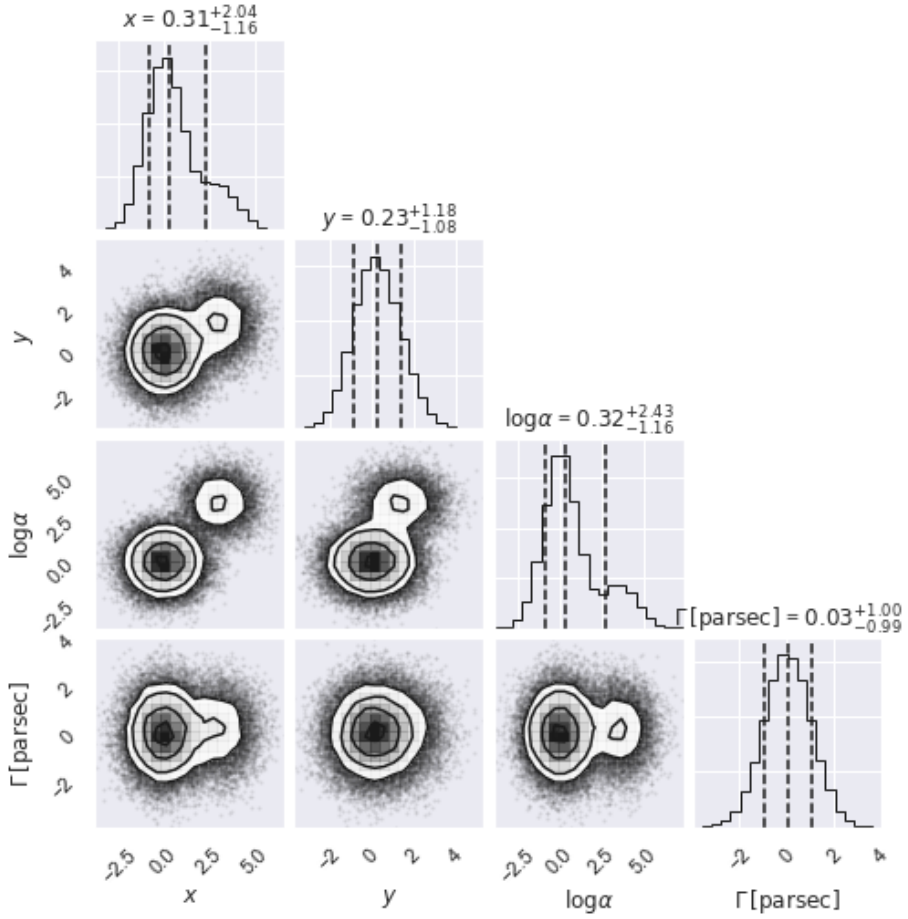

图 1.4: 显示多个参数之间关系的 corner 图

2 Parameter estimation

2.1 研究硬币是否公平

根据经验，对于同一个研究对象也就是同一个事实真理，不论做什么假设，在 likelihood 的修正下，得到的后验概率都是事实，不同的先验概率会收敛于同一个结果，并且在每次实验数据独立情况下，后验概率的结果与数据获取顺序无关。

2.1.1 关于后验和共轭先验 (conjugate prior)

贝叶斯的一个缺点为计算比较麻烦很难保证后验分布的具有解析解，指的是后验分布的密度函数即 $prob(H|\{data\}, I)$ 具有解析解（不过实际中即使 posterior 没有解析解也可以后验分

布进行采样或近似的推断）

为什么 posterior 不一定具有解析解：

针对连续性随机变量的贝叶斯公式：

$$prob(X|Y,I) = \frac{prob(Y|X,I) \times prob(X|I)}{\int_\Omega prob(Y|X,I) \times prob(X|I) dX} \tag{2.1}$$

分母为积分形式，所以 posterior 不一定具有解析解，是否有解析解取决于积分内模型和先验的选取，但并不是选择了可以使后验获得解析解的先验就可以使后验具有解析解。实际上，正如上学期讨论的新冠病毒的粒子，更新计算中的先验概率并不是一直是同一个先验，而是在获取数据之后，用得到的 posterior 当作新的先验代入公式进行更新计算。所以还要保证更新的后验概率代入公式可以使新的后验获得解析解。一个直接的方法就是让后验和先验具有相同的表达式，这样不仅保证了 posterior 具有解析解而且还可以让每轮计算都用同一个公式。

由此定义了共轭先验：在给定模型即似然函数的情况下，如果先验分布和似然函数可以使得先验分布和后验分布有相同的形式，那么就称此先验为这个模型的的共轭先验分布。

（定义：称一个分布族为模型 $Y \sim f_{Y|X}(Y|X) \& X \in \Omega$ 的共轭先验，若只要先验分布 $f_X(X|I)$ 是从该分布族中选取的，那么最终得到的后验 $f_{Y|X}(Y|X)$ 就也属于该分布族）

2.1.2 更新 likelihood 或 pirior 的例子

新冠例子: 实际有病但测得为阳性的概率为 $P(D|\bar{H}) = 2.3\%$, 实际没病缺测得阴性的概率为 $P(\bar{D}|H) = 1.4\%$, 患病概率 $P(H) = 0.1\%$ 现在求实际测得为阳性，患病的概率 $P(H|D) =?$ 已知 $P(D|H) + P(\bar{D}|H) = 1$，所以 $P(D|H) = 98.6\%$

$$\begin{aligned} P(H|D) &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{P(D|H)P(H)}{P(D|\bar{H})P(\bar{H}) + P(D|H)P(H)} \end{aligned} \tag{2.2}$$

一次检测为阳性的患病概率是很小的，所以要进行多次检测，有两种方法。比如，进行三次检测，1. 一种是一次一次的检测不断地更新先验概率密度，2. 一种是直接测三次，改变似然函数。

一次一次测，$P(H) \to P(H')$, 下式中用到 $P(D|H)P(H) + P(D|\bar{H})P(\bar{H}) = P(D)$,$P(D) -$

$P(D|H)P(H) = P(D|\bar{H})P(\bar{H})$

$$\begin{aligned} P'(H|D) &= \frac{P(D|H)P(H')}{P'(D)} \\ &= \frac{P(D|H)[P(D|H)P(H)]/P(D)}{P(D|\bar{H})P'(\bar{H}) + P(D|H)P'(H)} \\ &= \frac{P(D|H)P(D|H)P(H)/P(D)}{P(D|\bar{H})(1-P'(H)) + P(D|H)P(D|H)P(H)/P(D)} \\ &= \frac{[P(D|H)]^2P(H)}{P(D|\bar{H})[P(D) - P(D|H)P(H)] + [P(D|H)]^2P(H)} \\ &= \frac{[P(D|H)]^2P(H)}{P(D|\bar{H})P(D|\bar{H})P(\bar{H}) + [P(D|H)]^2P(H)} \\ &= \frac{[P(D|H)]^2P(H)}{P(D|\bar{H})^2P(\bar{H}) + [P(D|H)]^2P(H)} \end{aligned} \tag{2.3}$$

测量一次与测量两次的后验概率对比,

$$P(H|D) = \frac{P(D|H)P(H)}{P(D|\bar{H})P(\bar{H}) + P(D|H)P(H)} \tag{2.4}$$

$$P'(H|D) = \frac{[P(D|H)]^2P(H)}{P(D|\bar{H})^2P(\bar{H}) + [P(D|H)]^2P(H)} \tag{2.5}$$

这个化简后的结果与直接测量两次的结果一样。只是一次一次测时,先验由 $P(H) \to P(H') = P(H|D)$。直接测量两次，改变的是似然函数 $P(D|H) \to [P(D|H)]^2$

2.1.3 常见的模型及其共轭先验

eg1:beta-伯努利共轭

伯努利分布:

取 1 的概率为 θ，取 0 的概率为 $(1-\theta)$ 的离散型随机变量的分布。伯努利试验, 成功为 1，失败为 0，成功的次数服从伯努利分布, 参数 θ 是试验成功的概率。其 pmf(概率质量函数) 为:

$$P(x, \theta) = \theta^x(1-\theta)^{1-x} \tag{2.6}$$

其中 $x \in \{0, 1\}$

beta 分布:

beta 分布常用作先验 pirior。

beta 分布是由定义在 [0,1] 上的连续性概率分布构成的分布族，具有两个参数，α β ，其

pdf(概率密度函数) 为:

$$p(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{2.7}$$

其中 B 为 beta 函数，是分布函数。先验、后验和模型都有参数，为了区分，先验和后验的参数被称为超参数，模型的参数就叫参数。

实际模型中，beta 分布常用作先验分布，当模型为二项分布或者伯努利分布时，beta 分布都是这个模型的共轭先验。

定理 (**beta-伯努利共轭**): 若 $X \to Bernouli(\Theta)$ 且 $\Theta \to Beta(\alpha,\beta)$ 则观测到 X=x 的后验分布可以选做 $Beta(x+\alpha, \beta-x+1)$。证明如下

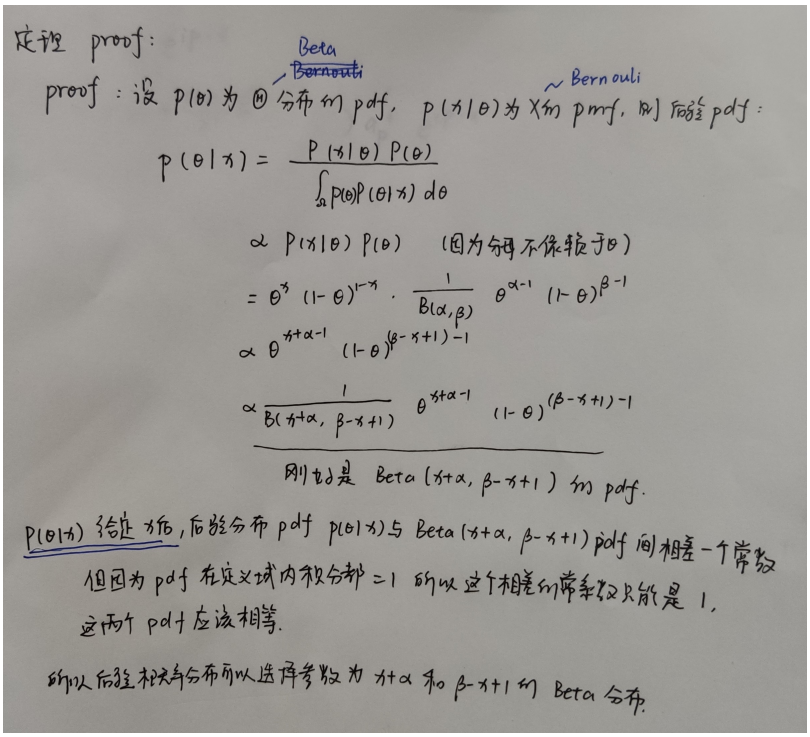

图 2.1:

eg2: 高斯-高斯共轭

高斯分布通常在已知均值或方差二者之一的情况下更容易找到共轭先验，因为在一个参数已知的情况下，只用关心另一个参数就行，高斯高斯共轭就是在已知模型 (likelihood) 的方差的前提下的一个共轭先验。

高斯分布:

高斯分布（或正态分布）是一个具有两个参数，由连续性分布所构成的参数化分布族，其

参数为均值 μ 和方差 σ^2。高斯分布的 pdf 为：

$$P(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2\sigma^2}(x-\mu^2)\right\} \tag{2.8}$$

高斯-高斯共轭：若 $X \sim \mathcal{N}(\mu_s, \sigma_s^2)$,$\mu_s \sim \mathcal{N}(\mu_P, \sigma_P^2)$，则观测到 X=x 的后验分布可以选做：

$$\mathcal{N}\left(\frac{\sigma_P^2}{\sigma_s^2+\sigma_P^2}x + \frac{\sigma_s^2}{\sigma_s^2+\sigma_P^2}\mu_P, (\frac{1}{\sigma_P^2}+\frac{1}{\sigma_s^2})^{-1}\right) \tag{2.9}$$

其中 σ_s, μ_s 是模型的参数，σ_P, μ_P 是先验的超参数。脚标 s 是信号 signal 的缩写，所以模型参数也叫信号均值和信号方差，同时 P 是 pirior 的缩写，所以超参数也被称为先验均值和先验方差。在**高斯共轭中，信号方差是已知的**。

与高斯分布对应的共轭先验有很多，高斯-高斯共轭只适用于信号方差已知，信号均值未知的情况。

对于抛硬币，由于每次数据都是独立的，likelihood（即模型）由二项分布给出，二项分布是 N 次伯努利试验。
一次伯努利试验

$$prob(\{data\}|H,I) \propto \theta^x(1-\theta)^{1-x} \tag{2.10}$$

因为生成的实际的实验数据，所以 x=p_h 真实值 0.4。与新冠例子中证明的一样，进行 N 次测量，一次次更新实验数据与一次更新完数据得到的测量结果是一样的。所以可以将 N 次实验后得到的 likelihood 变为

$$\begin{aligned} P(data|p_h,I) &= (p_h^x(1-p_h)^{1-x})^N \\ &= p_h^{xN}(1-p_h)^{N-Nx} \end{aligned} \tag{2.11}$$

其中 Nx=R: 头朝上次数，N-Nx=N-R：头朝下次数。这样计算与二项分布也是相同的。即

$$P(data|H,I) = H^{xN}(1-H)^{N-Nx} \tag{2.12}$$

先验取 beta 分布的概率密度函数：

$$P(H) = H^{\alpha-1}(1-H)^{\beta-1} \tag{2.13}$$

由

$$prob(H|\{data\},I) \propto prob(\{data\}|H,I) \times prob\ H|I \tag{2.14}$$

可以得到后验

$$P(H|data,I) = H^{\alpha+R-1}(1-H)^{\beta+N-R-1} \tag{2.15}$$

即参数为 $\alpha+R$ 和 $\beta+N-R$, 按照三行代码：
$y_1 = stats.beta.pdf(x, alpha_1 + heads, beta_1 + N - heads)$.... 画出三条不同 pirior 下的后验概率函数图。

实验数据 $\{data\}$ 由 $generate_data$ 定义的伯努利分布函数生成，超参数为 R，实验数据按照伯努利分布参数 $H = p_h$ 真实值 0.4 生成，产生一组 0 和 1 的数列。

在程序中对先验进行计算时，选用了共轭先验，采用共轭先验的原因是可以使得先验分布和后验分布的形式相同，这样计算起来就较为方便。前面的证明可以看出后验分布可以选择参数为 $\alpha+N$ 和 $\beta-N+1$ 的 Beta 分布。
随着实验数据增加，后验 pdf 的形状如下所示：
蓝色线 $\alpha=\beta=1$，红色线 $\alpha=\beta=30$, 绿色线 $\alpha=\beta=0.2$

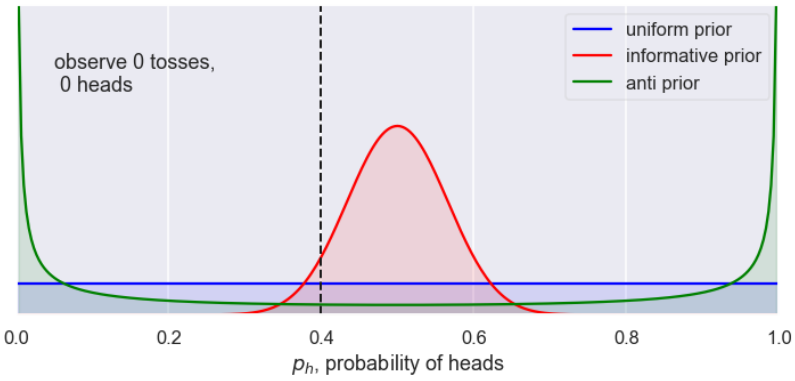

图 2.2: 0 tosse

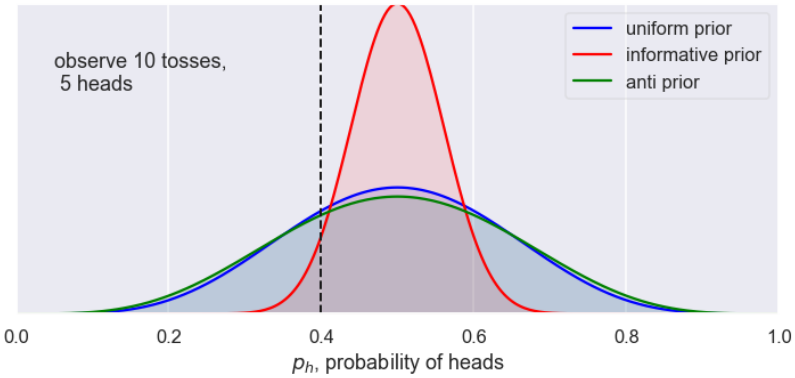

图 2.3: 10 tosses

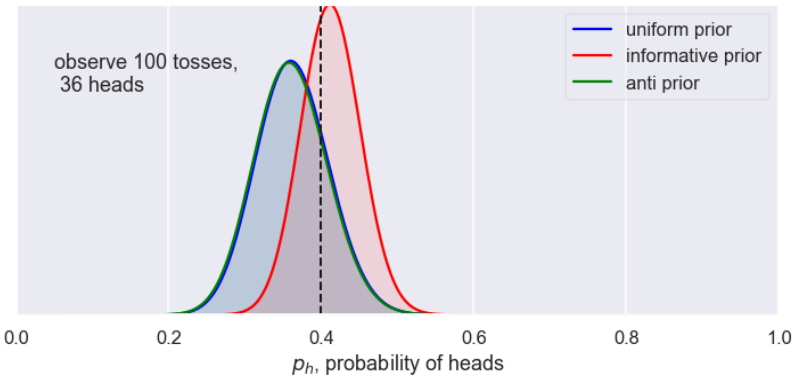

图 2.4: 100 tosses

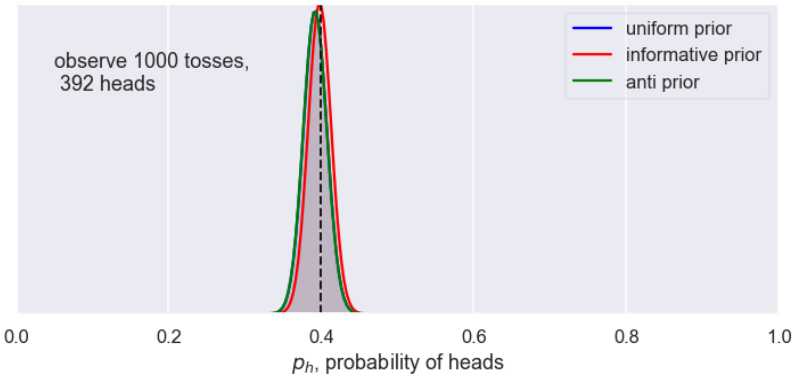

图 2.5: 1000 tosses

并且在程序中单独对平坦先验的结果进行了分析，返回的置信区间中，68% 不包含真实值，但是 98% 置信区间包含真实值。

2.2 最佳估计，误差条，可信度

现在已经知道后验 pdf 如何在给定数据和相关背景信息的情况下，对关于参数值的推断进行概率计算。接下来还需要用两个参数总结这些，即 best estimate 最佳估计和 confidence 对结果可靠性的衡量。

最佳估计由后验 pdf 的最大值给出，如果用 X 表示感兴趣的参数，后验表示为 $P = prob(X|data, I)$, 最佳估计 X_0,

$$\left.\frac{dP}{dX}\right|_{X_0} = 0 \tag{2.16}$$

严格来说，应该检查二阶导数小于零以确保 X_0 代表最大值。

由于进行了微分操作，所以假设了 X 是一个连续参数。如果参数 X 只能取离散值，best estimate 依然是最大后验概率的估计，但是不能用上面的微分表达式。? 那么应该怎么计算呢?

为了获得这一最佳估计的可靠性的度量，需要观察的 pdf 在 X_0 附近的宽度或分布。要考虑函数在特定点附近的行为时，泰勒展开会比较有帮助，泰勒展开是一个简单且标准的工具，用于用低阶多项式逼近一个复杂的函数。

计算 best estimate 时，很难找到解析解，所以可以对后验 pdf 取对数（1. 避免计算机精度引起的误差。2. 可以把乘法化为加减法简化计算减少计算周期。）

$$L = \log_e[prob(X|data, I], \tag{2.17}$$

将 L 在 X_0 处展开，

$$L = L(X_0) + \frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_0}(X - X_0)^2 + ..., \tag{2.18}$$

这里的 $L(X_0)$ 是一个常数，并且一阶导等于 0 了，所以二次项是决定后验 pdf 宽度的主导因素，在可靠性分析中担任中心角色。忽略所有的高阶项，

$$prob(X|\{data\}, I) \approx A\exp[\frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_0}(X - X_0)^2]. \tag{2.19}$$

A 是归一化常数。这么做的目的是用简单的高斯分布（正态分布）来近似后验 pdf，

$$prob(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}exp[-\frac{(x-\mu)^2}{2\sigma^2}], \tag{2.20}$$

这个式子与后验 pdf 近似式相比可以发现，后验 pdf 是最大值在 $\mu = X_0$ 处，由参数 σ 描述的高斯分布，

$$\sigma = \left(-\left.\frac{d^2L}{dX^2}\right|_{X_0}\right)^{-\frac{1}{2}}, \tag{2.21}$$

由高斯积分性质可以看出，X 的真实值落在 $X_0 - \sigma$ 到 $X_0 + \sigma$ 范围内的概率为 67%：

$$prob(X_0 - \sigma \le X \le X_0 + \sigma|data, I) = \int_{X_0-\sigma}^{X_0+\sigma} prob(X, data|I)dX \approx 0.67, \tag{2.22}$$

$$Bestestimate : X_0, Reliability : \sigma$$

参数 σ 也被叫做误差条 error-bar

2.2.1 硬币例子

在硬币例子中 $prob(H|\{data\}, I) \propto H^R(1-H)^{N-R}$，对其取对数，并计算

$$\frac{dL}{dX} = \frac{R}{H}\frac{(N-R)}{1-H}, \quad \frac{d^2L}{dX^2} = -\frac{R}{H^2} - \frac{(N-R)}{1-H}^2, \tag{2.23}$$

根据

$$\left.\frac{dL}{dX}\right|_{H_0} = 0, \tag{2.24}$$

推出最佳估计 $bestesitimate : H_0 = \frac{R}{N}$, 所以

$$\left.\frac{d^2L}{dX^2}\right|_{H_0} = -\frac{N}{H_0(1-H_0)}, \tag{2.25}$$

因此，

$$error - bar : \sigma = \sqrt{\frac{H_0(1-H_0)}{N}}. \tag{2.26}$$

H_0 在经过一定数量的数据分析后变化不大，分子趋于一个定值，因此后验的宽度与数据量的平方根成反比，并且可以看出证明一个硬币不公平比证明他是公平的更简单。

2.2.2 非对称后验 pdfs

在之前的图中可以注意到峰并不是一直对称的，随着数据量的增加，后验 pdf 的形状会越来越像高斯分布，在数据量不够时，error-barσ 的使用就有一些不足，因为误差条隐含了对称的信息，pdf 不对称时，后验最大值 X_0 依然表示最佳估计, 但真实值可能在这个峰的旁边。

一个很好的办法是通过置信区间来推断参数的可靠性，若 pdf 已经归一化，考虑 95% 置信区间。

$$prob(X_1 \leq X \leq X_2|\{data\}, I) = \int_{X_1}^{X_2} prob(X|\{data\}, I)dX \propto 0.95. \tag{2.27}$$

其中 X1，X2 的差值越小越好。同时可以考虑均值 mean 和期望值 expection，他们考虑到了 pdf 的不对称性、偏度。归一化的 pdf 的加权平均

$$< X >= \int prob(X|\{data\}, I)XdX, \tag{2.28}$$

(X 只能取离散值时积分用求和代替), 如果后验 pdf 没有归一化, 那么右边必须除以一个归一化系数 $\int prob(X|\{data\}, I)dX$, 如果是高斯分布, 则均值与最佳估计 X_0 刚好相等 ($X_0 =< X >$)。

2.2.3 多模态后验

如果后验 pdf 一个极大值比其他的都大时，可以简单的忽略附属解，但是如果是几个规模相当的极大值那么 best estimate 是不能做出解释的。因为后验 pdf 给出了完整地描述，可以根据数据和相关的先验知识推断出参数的值。但是企图用最佳估计、误差条、置信区间两三个数字总结后验 pdf 有的时候是没办法做到的。不过后验 pdf 是存在的，可以从中得到适当的结论。

如图

忽略 $X = 20$ 右边的结构，则这个后验传递的 $X = -10$ 或 $+10$，可以写成 $X = -10 \pm 2$

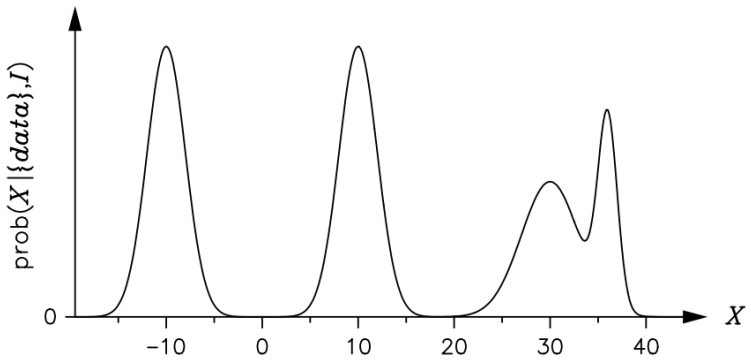

图 2.6:

或 $X = 10 \pm 2$，然而 pdf 的均值仍然是唯一的，所以有时考虑用均值表示最佳估计，但是在图中的 pdf 上，忽略右边的结构后，期望值 $< X >= 0$，在后验 pdf 图画中显示这个取值是很不可能的，即使这样也依然使用均值表示最佳估计的话，需要对其分配一个很大的 error bar 来让置信区间内包含这个值，因此也不能很好的反应后验 pdf 中固有的信息。对于双峰 pdf 可以用几个数据来描述后验 pdf：两个最佳估计，及两个最佳估计分别相关的 error bar，或者不相交的置信区间。一般对于多模态，我们能做的只是诚实的显示后验 pdf 本身。

2.3 高斯噪声和平均值

$$p(x_k|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

第 k 个数据的值为 x_k 的概率由该式给出,μ 是感兴趣的参数，σ 是实验中对测量误差的测量值。给出一组数据 x_k，μ 的最佳估计是什么？对于这一预测的 confident 是多大?

在这个例子中，σ 的值是已知的，因此对 μ 值的推断通过后验 pdf$prob(\mu|x_k,\sigma,I)$ 来计算。

$$prob(\mu|x_k,\sigma,I) \propto prob(x_k|\mu,\sigma,I) \times prob(\mu|\sigma,I), \tag{2.29}$$

如果假设数据之间是独立的，那么

$$prob(\mu|\{x_k\},\sigma,I) = \prod_{k=1}^{N} prob(x_k|\mu,\sigma,I), \tag{2.30}$$

由于高斯峰的宽度没有关于中心值的信息，并选取一个平坦的先验

$$prob(\mu|\sigma, I) = prob(\mu|I) = \begin{cases} A & \mu_{min} \leq \mu \leq \mu_{max}, \\ 0 & otherwise, \end{cases} \tag{2.31}$$

后验 pdf 取对数后

$$L = log_e[prob(\mu|x_k, \sigma, I)] = C(Nln\frac{1}{\sigma\sqrt{2\pi}} + lnA) + \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2}, \tag{2.32}$$

这里的常数项与 μ 无关，后验 pdf 在 $\mu_{min} \leq \mu \leq \mu_{max}$ 范围外等于 0。为了找到最佳估计 μ_0 对 L 求一阶导等于 0，

$$\left.\frac{dL}{d\mu}\right|_{\mu_0} = \sum_{k=1}^{N} \frac{x_k - \mu_0}{\sigma^2} = 0$$

其中 σ 是与 μ 无关的常数，所以可以提到求和外，因此

$$\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} \mu_0 = N\mu_0$$

μ 的最佳估计由 xk 的数值平均 arithmetic average 决定，

$$\mu_0 = \frac{1}{N} \sum_{k=1}^{N} x_k \tag{2.33}$$

由 L 的二阶导数可以求出 σ

$$\left.\frac{d^2L}{d\mu^2}\right|_{\mu_0} = \sum_{k=1}^{N} \frac{1}{\sigma^2} = -\frac{N}{\sigma^2} \tag{2.34}$$

根据公式 (32), 误差条应该由上式的倒数的负的平方根给出，所以

$$\sigma = \sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}}$$

总结对于 μ 值的推断

$$\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}} \tag{2.35}$$

与投硬币实验一样，得到了相似的结果，估计的可靠性与实验数据的数量的平方根成正比。

在前面注意到，误差条的概念依赖于方程的二次展开的有效性，之前忽略了高阶展开，仅仅保留了泰勒展开的二次项，并把它写成了高斯的形式。而在这个高斯噪声的情况，这不是一

个近似形式而是精确的展开式因为高阶导数全为 0。因此后验 pdf 完全可以由误差条的定义来描述函数行为。

唯一的条件是 μ_{min} 和 μ_{max} 的范围。原则上可以通过让 min 和 max 趋于无穷来象征对于先验的无知，但是，如果这个范围大到了一定程度，那么他俩的值对于后验 pd 是没有影响的。如果最佳估计和误差条允许的 μ 值超出了 μ_{min} 和 μ_{max} 这个范围，那么只能去显示后验 pdf 本身及其截断？这种情况告诉我们先验知识也一样重要？

2.3.1 有不同大小 error bar 的数据

在之前的分析中，都是假设对于每个数据，误差条大小都是一样的，如果所有测量都是在一样的实验装置下进行，那么这样是合理的，但是如果实验数据来自不同精密程度的几个实验室获得的，那么应该如何结合不同实验室的数据呢？假设测量误差依然可以通过高斯 pdf 进行建模，也就是 likelihood 依然满足 guass 分布，因此第 k 个数据的值为 xk 的概率分布为

$$prob(x_k|\mu,\sigma_k,I) = \frac{1}{\sigma_k\sqrt{2\pi}} exp[-\frac{(x_k-\mu)}{2\sigma_k^2}]. \tag{2.36}$$

跟之前一样的方法

$$prob(\{x_k\}|\mu,\sigma_k,I) = \prod_{k=1}^{N} prob(x_k|\mu,\sigma_k,I), \tag{2.37}$$

$$L = Constant - \sum_{k=1}^{N} \frac{(x_k-\mu)^2}{2\sigma_k^2}, \tag{2.38}$$

同理，一阶导数等于 0，得到

$$\mu_0 = \frac{\sum_{k=1}^{N}\omega_k x_k}{\sum_{k=1}^{N}\omega_k}, \tag{2.39}$$

其中 $\omega_k = \frac{1}{\sigma_k^2}$ **这里计算 best estimate 使用加权平均 weighted average 而不是算术平均 arithmetic mean，这样不可靠的数据会对应的更大的 error bar，和相应更低的权值**。L 的二阶导数产生最佳估计的误差条，关于 μ，

$$\mu = \mu_0 \pm (\sum_{k=1}^{N}\omega_k)^{-1/2}. \tag{2.40}$$

Note that if all the data were of comparable quality,so that sigmak=sigma,….?

2.4 example3:lighthouse

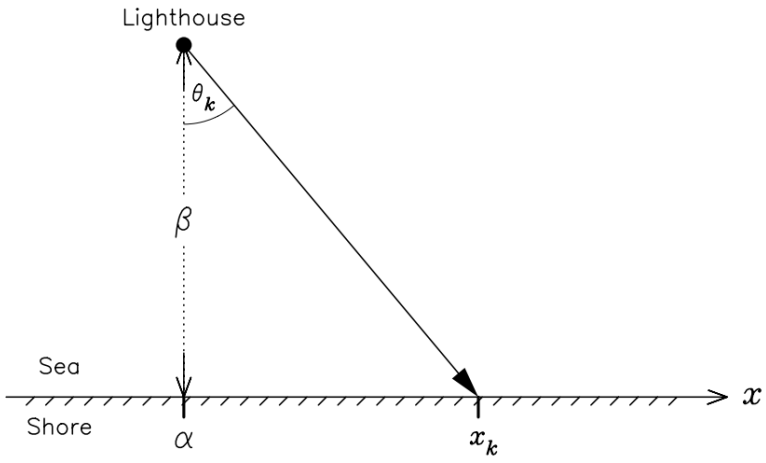

图 2.7:

灯塔的位置对应海岸线上的 α，离海岸的距离为 β，灯塔随机间隔、随即方位角发射一系列闪光照到岸上，岸上记录了 N 次数据x_k, 求灯塔在哪里。考虑灯塔发射的性质，为第 k 个数据的方位角 θ_k 分配一个均匀的 pdf 是合理的，

$$prob(\theta_k|\alpha,\beta,I) = \frac{1}{\pi}, \tag{2.41}$$

θ 的取值范围在 $\pm\pi/2$ 之间。将 θ_k 和 x_k 联系起来

$$\beta tan\theta_k = x_k - \alpha, \tag{2.42}$$

3.6 节可以看到，在处理 changing variables 时，可以用下面的式子重写 2.41 式。

$$prob(x_k|\alpha,\beta,I) = \frac{\beta}{\pi[\beta^2 + (x_k-\alpha)^2]}. \tag{2.43}$$

由此，已知灯塔 (α,β) 的坐标，第 k 次闪光记录在位置 x_k 的概率由柯西分布给出。

这种 pdf 的函数形式在物理学中经常遇到，通常称为洛伦兹函数。它关于最大值 $x_k = \alpha$ 对称，其 FWHM 为 2β(FWHM 是半峰全宽，峰一半高处的峰宽度); 分布如图 2.8 所示。

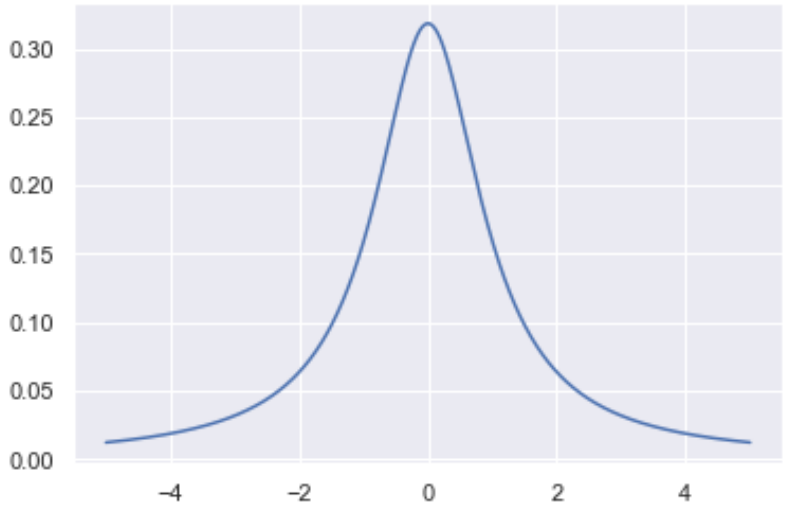

图 2.8:

柯西分布是一个数学期望不存在的连续型概率分布。当随机变量 X 满足它的概率密度函数时，称 X 服从柯西分布。

从数据中推断灯塔位置，需要同时对 α 和 β 进行估计; 这章讨论单参数问题，所以假定灯塔到海面的距离 β 已知，并将其简化为一个单一参数的问题。对于灯塔位置的 inference 由后验 pdf$prob(\alpha|x_k,\beta,I)$ 表示。作为已知数据，β 没有提供任何关于 α 的信息，所以可以将先验取为平坦的:

$$prob(\alpha|\beta,I) = prob(\alpha|I) = \begin{cases} A & \alpha_{min} \le \alpha \le \alpha_{max}, \\ 0 & otherwise, \end{cases} \tag{2.44}$$

α_{min} 和 α_{max} 可以代表海岸线的边界，这些独立数据的概率函数是获得 N 个单独探测的概率的乘积:

$$prob(\{x_k\}|\alpha,\beta,I) = \prod_{k=1}^{N} prob(x_k|\alpha,\beta,I), \tag{2.45}$$

跟之前一样，将由 (55) 得到的 prior，和由 (54)(56) 得到的 likelihood 代入贝叶斯公式，得到 posterior 并对它取对数，得到

$$L = log_e[prob(\alpha|x_k,\beta,I)] = constant - \sum_{k=1}^{N} log_e[\beta^2 + (x_k - \alpha)^2], \tag{2.46}$$

其中常数不包含 α。假设先验的范围无限大，也就是海岸线无限长，这样就不用担心后验函数

的截断。位置的最佳估计由后验函数最大值 α_0 给出，由此

$$\left.\frac{dL}{d\alpha}\right|_{\alpha_0} = 2\sum_{k=1}^{N}\frac{x_k-\alpha_0}{\beta^2+(x_k-\alpha_0)^2} = 0. \tag{2.47}$$

这个方程很难重新排列，所以用 β 和 x_k 来表示 α_0。虽然解析解不好搞，但是可以**用数值法解**这个式子: 从 (57) 式，为 α 的一系列不同的可能值计算 L。得到最大 L 的 α 就是最佳估计。如果我们在纵轴上画出 L 的指数 exp(L)，横轴为 α，就得到了灯塔位置的后验 pdf，提供了推断的图像，优点是不需要担心后验 pdf 是否是不对称的或多模态的。

对于 (57)，倾向于从总和中忽略常数项，因为它既不影响最佳估计，也不影响误差条。它的值对于后验 pdf 来说只是乘了一个常数，只是将函数整体放大。

生成随机数：根据 (52) 生成均匀分布的方位角样本 θ_k，用 (53) 将其转换为位置数据 x_k，假定已知灯塔距离海岸 1 公里 ($\beta = 1$)。如图 12：在前几个图中，闪光的位置由图形顶部的小圆圈标记，数据的数量显示在右上角。

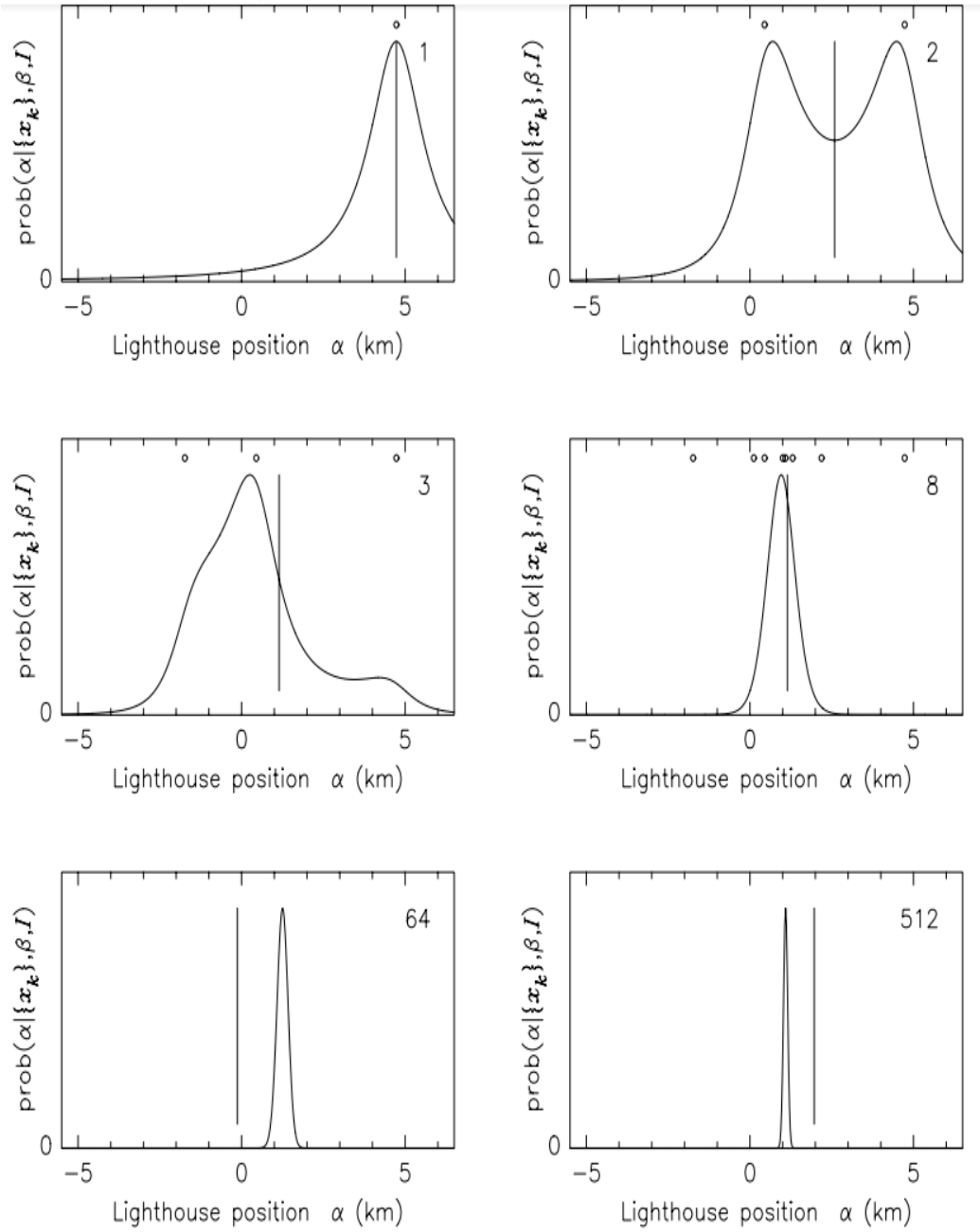

图 2.9: 随着海岸上探测到的闪光数目的增加，灯塔位置的后验 pdf 如何演变，直线表示平均值

数据量很少时很容易成为多模态，经过大约十几次的测量，后验接近高斯 pdf 形式。随着实验证据的增加，后 pdf 变窄，峰值收敛到 α= 1。与预期一致，因为数据是在沿海岸 1 公里

处生成的。

2.4.1 对应的代码

首先 $num\_pts = 512$ 用 $x\_pts = np.arange(num\_pts)$ 生成 512 个坐标点数组，然后用 $dist = cauchy(x0\_true, y0\_true), dist\_pts = dist.rvs(num\_pts)$ 生成了满足柯西分布的数据数组 () 柯西分布的两个参数选用了真实值 $x0\_true = 1.y0\_true = 1.$ ，并从中随机抽取 512 个随机样本，并与坐标点一一对应，绘图，以下三个图：

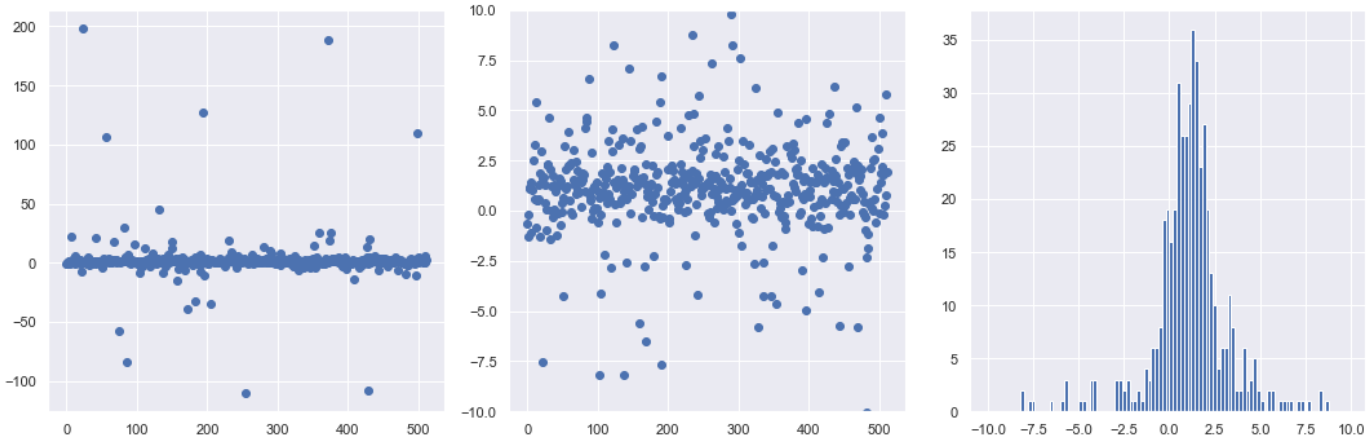

图 2.10:

第一个图：以 x_pts512 个坐标点为横轴，以 $dist\_pts$ 从柯西分布中抽出的 512 个样本值为纵轴，可以看出样本值在 0 附近的点显著的多。

第二个图：是把第一个图放大了，显示样本值在-10 到 10 附近的数据点。第三个图：横坐标是样本取值，纵坐标是取该值的样本的个数。

上面的例子展示了柯西分布，接下来对于灯塔的例子。

同样首先 $num\_pts = 512$ 用 $x\_pts = np.arange(num\_pts)$ 生成 512 个坐标点数组，已知 $prob(\theta_k|\alpha, \beta, I) = 1/\pi$，所以用 $theta\_dist = uniform(-np.pi/2., np.pi/2.)$，均匀的在 $-\pi/2$ 到 $\pi/2$ 之间取点，再用 rvs 对其随机取样取了 512 个；也可以用 np.random.uniform($-\pi/2, \pi/2$,512) 直接随机取出 512 个样本，画出来的图是一样的，只是两个 uniform 来自于不同的库。

对于取出的 θ_k 的值，需要转化为坐标的样本，利用公式

$$\beta tan\theta_k = x_k - \alpha.$$

并且，似然函数的表达式为

$$prob(x_k|\alpha,\beta,I)=\frac{\beta}{\pi[\beta^2+(x_k-\alpha)^2]},$$

将 $x_k-\alpha$ 整体带入表达式右侧，并且已经假设 β 已知，所以将 $\beta=1$ 代入似然函数，得到似然函数的表示式

$$prob(x_k|\alpha,\beta,I)=\frac{1}{\pi[1+(tan\theta_k)^2]},$$

$prob(x_k|\alpha,\beta,I)=dist\_pts\_alt=1/(np.pi*(1+(np.tan(theta\_dist\_pts).T)*(np.tan(theta\_dist\_pts))))$

直接对似然函数绘图

画出来三个图：

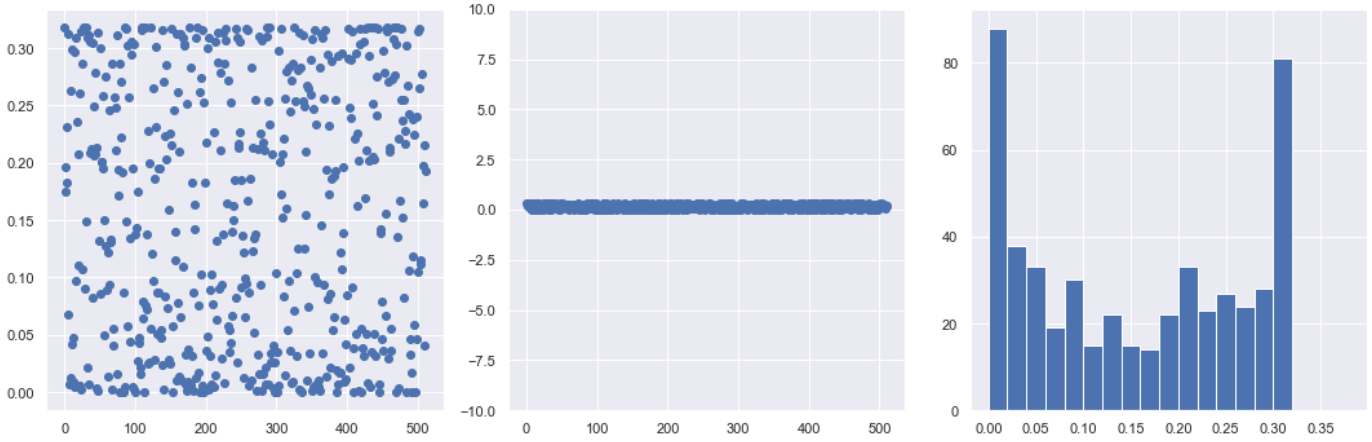

图 2.11:

这三个图可以看出似然函数的值分布在，0 到 0.3 多，并且从第三个图中看出，似然函数取值再 0 附近和右边界附近的数量明显比中间区域更多。

这样分布是合理的。$\tan\theta_k$ 的图像如下，

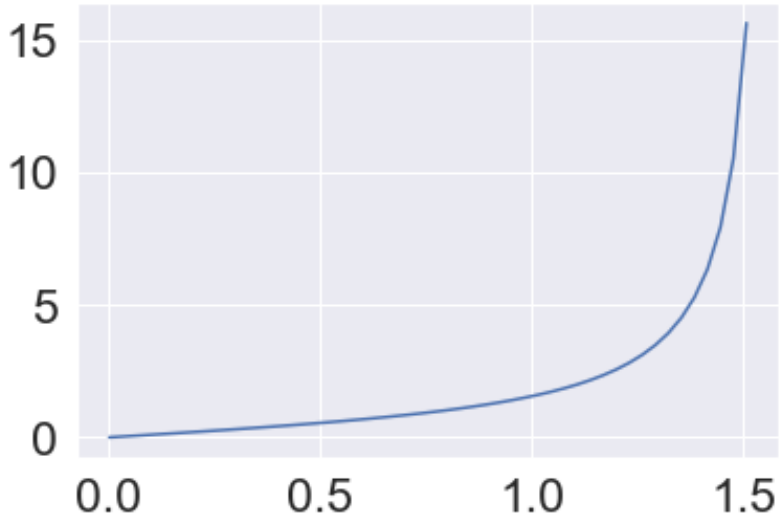

图 2.12:

根据计算，tan 值处在 1.128 到 1.524 的时候，似然函数的值才会洒在 0.05 到 0.25 之间，下面的 tan 函数形状中可以看出，tan 处在 1.128 到 1.524 的区间范围很小，对应的上面第三个图中，生成的数据点大多在 0 和 0.3 附近。

根据 $x_k = \beta tan\theta_k = x_k + \alpha$，画出 $prob(x_k|\alpha, \beta, I) - x_k$ 图

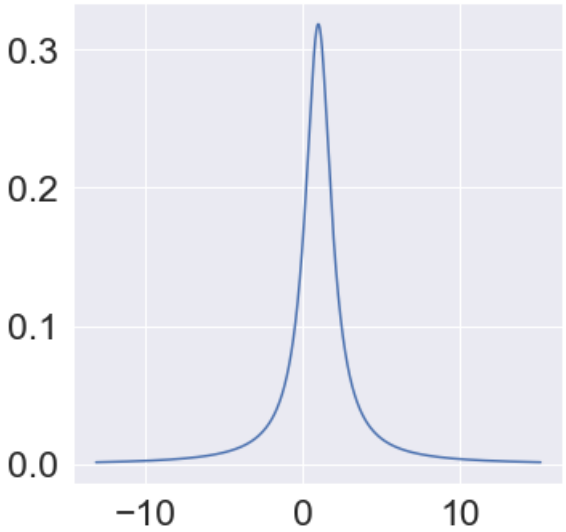

图 2.13:

在 α 和 β 已知的情况下，生成的样本 x_k 落在 $x_k = 1$ 的概率最大。

接下来绘制了计算并绘制了不同数据量下的 xo 后验 $prob(x_k|\alpha, \beta, I)$

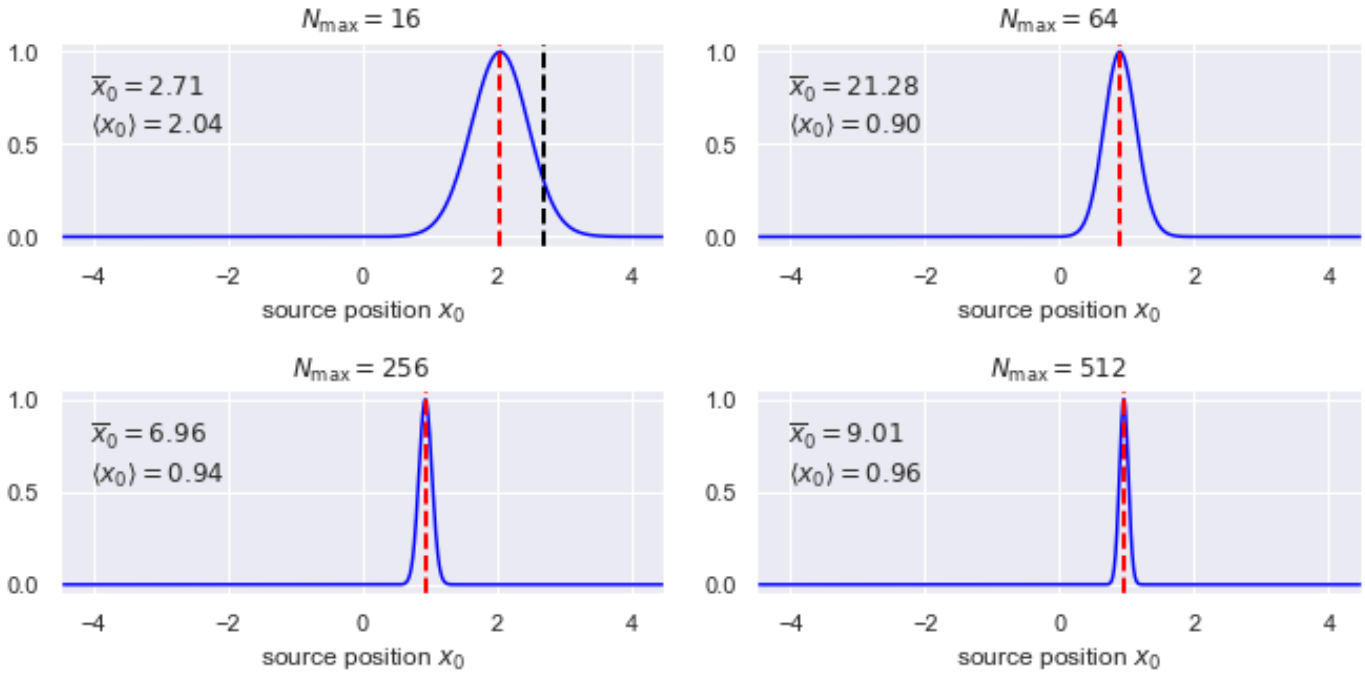

图 2.14:

红线是 $< X_0 >$

$$< X_0 >= \frac{\sum X_0 * posterior}{\sum posterior},$$

黑线是 $\bar{X}_0$

$$\bar{X}_0 = mean(X_0[0 : N_max]).$$

数据的均值并不能很好的描述参数的最佳估计，图上也可以看出黑线与最佳估计之间有一定的差距。与之前不同，在高斯分布中，参数 μ 的最佳估计 $\mu_0 = \frac{1}{N}\sum_{k=1}^{N} x_k$，由样本简单的算数平均给出。在灯塔例子中，数据来自柯西 pdf，由于这个分布也是关于我们感兴趣的参数 α 对称的，所可能会认为，数据的平均值也可以提供对灯塔位置的良好估计。实际上样本均值并不是 X 的最佳估计。图中也可以看到，数据的平均值由黑色长竖线表示。可以看到对于这个问题，样本均值不是一个很好的估计。

中心极限定理: 如果从 (几乎) 任何 pdf 中随机抽取样本，pdf 的均值为 μ，那么在数据充足的情况下，数据的平均值将趋向于这个值 μ;μ 和样本均值之间的差值的误差条会以 $\sqrt{N}$ 下降，随着数据量增加，在除以 N 来计算平均值时，平均值上的误差条会以 $\sqrt{N}$ 的量级减小。但

对于柯西分布，由于在 $-\infty$ 到 ∞ 是不可积的，所以柯西分布的期望值是没有定义的，σ^2 无限大，并且 μ 不确定，所以数据平均值的可变性并不会随着测量次数的增加而减少，而且在测量了一千个或一百万个数据后，其“错误”可能和测量了一个数据后一样。

反而后验均值对于最佳估计有一个很好的描述，也就是红线。

$$< X_0 > \frac{\sum X_0 * posterior}{\sum posterior}.$$

3 参数估计

之前的内容只涉及一个未知变量，接下来考虑有好几个参数的情况，对其中一些感兴趣。对error-bar、边缘化进行推广。还将看到某些近似如何自然地产生一些最常用的分析过程，并讨论所谓的误差传播 (propagation of errors)。

3.1 exampli 4: 存在信号本底时的信号振幅

最简单的情况，如图

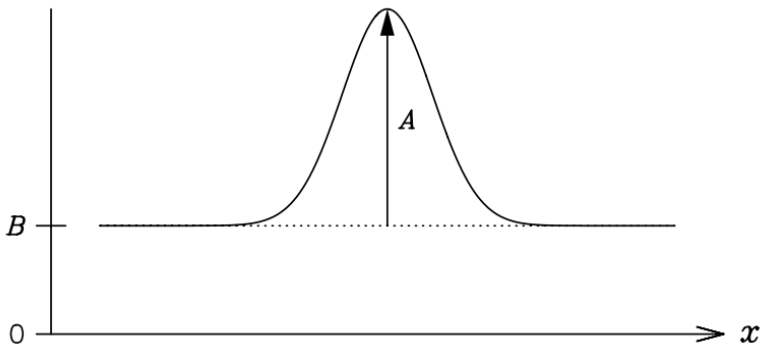

图 3.1:

横轴 c 为测量的变量。背景 (比如系统误差) 可以被认为是平坦的，幅度是 B 是未知的，而感兴趣的信号是已知形状和位置峰，幅值为 A。问题的数据通常都是整数值。给定一组实验装置测量的计数 N_k ，实验探测装置位于 X_k，**对信号峰值和背景的振幅的最佳估计是什么?**

先对数据的性质进行判断，应该期望第 k 个数据通道中的计数与 x_k 处的信号和背景之和成正比；取峰值的形状为 gauss 分布，宽度为 ω，中心值在 x_0 处**理想数据**由 D_k 给出：

$$D_k = n_0[Ae^{-(x_k-x_0)^2/2\omega^2} + B], \tag{3.1}$$

n_0 是与测量时间有关的常数。然而，与计数 N_k 的数量不同，上式中的 D_k 通常不是一个整数。因此，实际的数据将是这种理想数据附近的一个大于等于 0 的整数。**泊松分布**是一个满足这种属性的 pdf，通常在**这样的计数实验**中被调用。

泊松分布：

日常生活中，大量事件是有固定频率的。比如某医院平均每小时出生 3 个婴儿，这种事件的特点是可以预估这些事件的总数，但是没法知道具体的发生时间。已知平均每小时出生 3 个婴儿，那么两个小时中的第二个小时，会出生几个? 有可能一下子出生 6 个，也有可能一个都不出生。

泊松分布就是描述某段时间内，事件具体的发生概率，

$$P(N(t)=n)=\frac{(\lambda t)^n e^{-\lambda}}{n!}, \tag{3.2}$$

等号的左边，P 表示概率，N 表示某种函数关系，t 表示时间，n 表示数量，λ1 小时内出生 3 个婴儿的概率，就表示为 P(N(1) = 3) 。等号的右边，λ **表示事件的频次**。接下来两个小时，一个婴儿都不出生的概率是 0.25%，基本不可能发生。与这个例子相对应，单位时间 t=1 内，已知理想数据数据为 D_k 周围的整数，所以 D 对应频率 λ，$N_k = N$ 的概率分布函数为

$$prob(N|D)=\frac{D^N e^{-D}}{N!}, \tag{3.3}$$

与之前的定义 <X> 相似，将上式代入离散形式定义，期望值 =D ：

$$< N >= \sum_{N=0}^{\infty} N prob(N|D) = D, \tag{3.4}$$

根据式子 (60)，数据为 N_k 的似然函数:

$$prob(N_k|A,B,I)=\frac{D_k^{N_k} e^{-D_k}}{N_k!}, \tag{3.5}$$

其中，背景信息 I 包括：计数的期望数量 D_k 与感兴趣的参数 A 和 B 之间的关系的知识; 对于 (59) 的高斯峰形模型，这意味着 x_0、ω 和 n_0 取给定 (以及 x_k)。如果数据是独立的，那么，当 A，B 给定时，在一个通道内观察到的 N_k 不影响在另一个通道内发现的粒子数，所以 likelihood 是单个测量的 prob 的乘积：

$$prob(\{N_k\}|A,B,I)=\prod_{k=1}^{M} prob(N_k|A,B,I), \tag{3.6}$$

对信号振幅和信号本底的推断体现在 posterior 中，

$$prob(A,B|\{N_k\},I) \propto prob(\{N_k\}|A,B,I) \times prob(A,B|I) \tag{3.7}$$

最简单的是均匀的 pdf，由于振幅和信号本底都不能是负的，所以：

$$prob(A,B|I)\begin{cases} Constant \quad for A \geq 0 and B \geq 0, \\ 0 \quad otherwise, \end{cases} \tag{3.8}$$

将公式代入 (65), 并取对数，可以得到

$$L = log_e[prob(A, B|N_k, I)] = constant + \sum_{k=1}^{M}[N_k log_e(D_k) - D_k], \tag{3.9}$$

常数项不包含 AB 系数并且求和项大于 0，对信号峰和背景幅度的最佳估计是由 L 的最大值给出的，可靠性由 posterior 在最佳点附近的宽度给出。

数据是根据方程 (63)，用泊松分布随机数发生器从方程 (59) “平坦背景上的高斯模型” 生成的，以下是四组数据以及得到的后验，绘制为直方图因为观测通道是离散的，并且统一观测装置的宽度。基础信号以原点为中心，所以 $x_0 = 0$，并且半高宽为 5 个单位，假设这些对于分析是已知的

后验现在是二维的因为它同时是 A 和 B 的函数，可以用等高线表示，也就是等置信区间线。最外圈到最内圈，分别是 $10\%, 30\%, 50\%, 70\%, 90\%$ 等置信区间线。

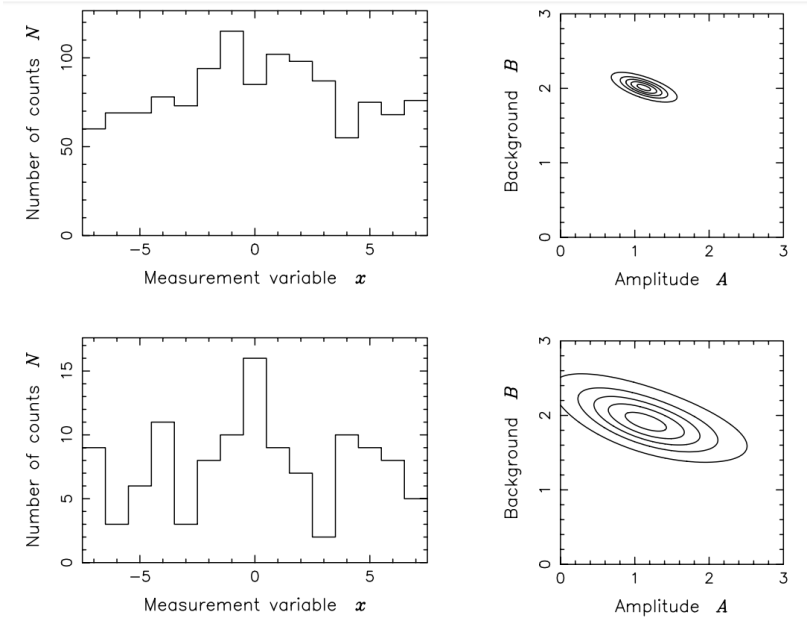

图 3.2:

第一张图显示在 15 个数据箱中检测到的计数数，参数 n_0 是给定的，n_0 **是与时间有关的常数**，理想数据 D_k 的最大值给出 n_0，$n_0 = D_{kmax}/A + B$，最大 expect data 的值为 100，由此而得到 n_0，由方程 (67) 的指数得到对应的后验 pdf，画在第一行的第二个图中；表明信号幅度的最佳估计大约 1，大约是背景幅度一半。

第二张图是相同的实验设置下，但是实验只进行了十分之一的时间，理想计数数量是之前的 1/10 倍即 $D'_{kmax} = 0.1D_{kmax}$，从而得到新的 n_0，实验数据减少，数据看起来更嘈杂。右

边的图在两个方向上的后验 pdf 大约比之前多了三倍宽，与之前 $\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}}$ 类似，数据量减少，置信区间宽度以 $\sqrt{N}$ 的比例增大，大约是三倍多宽。并且可以看到第二行右边的图中 A 小于 0 的后验被截断了体现了当数据不够好时，先验的重要性。

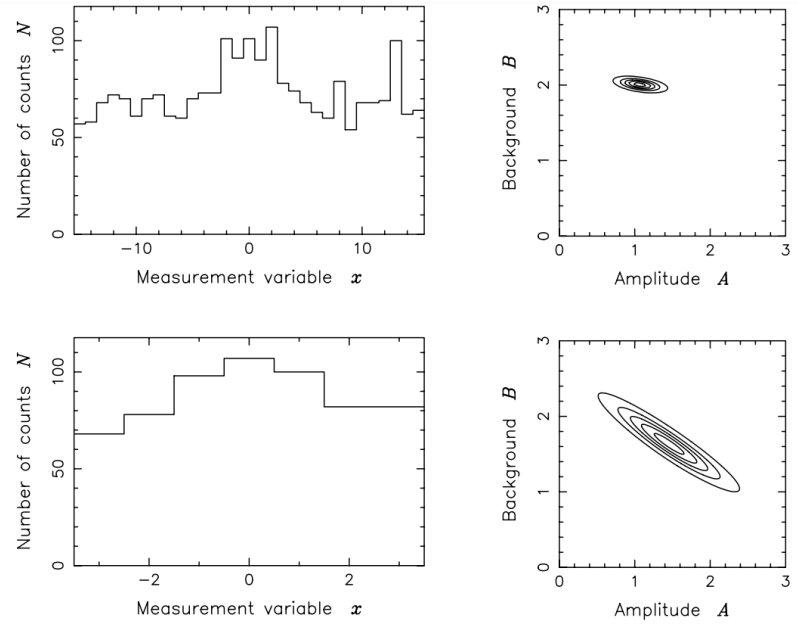

图 3.3:

这张图的第一行与上一张图第一行的计数率相同，有着相同的 D_kmax 和 n_0 但是粒子探测通道增加，总探测范围变为原来两倍，探测器间距不变，可以看到 A 和 B 的置信区间都缩小了，等置信区间图变得更加平稳。

第二行探测器数量减少，探测范围减少，数据量是是之前的图第一行的数据的一半，但是置信区间的宽度明显大于 $\sqrt{2}$ 倍，这是因为只有**在** $x_0 = 0$ **附近的数据**才能体现信号峰的信息，而**远离** x_0 **的数据**会提供背景幅度的信息，所以虽然计数率提高，但是得到的结果依然不好，并且图像明显倾斜。这些特征表明了对 A 和 B 的估计之间有很强的**相关性**。由于收集数据的 x 的范围受到了严重的限制，因此很**难将信号与背景区分**开。

3.1.1 Marginal distributions

二维后验 pdf 很好的描述了对 A 和 B 的值的联合推断。但是实际上通常对背景信息不感兴趣，只想计算对 A 的估计，也就是需要后验 $prob(A|\{N_k\}, I)$。根据边缘化规则可以通过积分得到：

$$prob(A|\{N_k\}, I) = \int_0^\infty prob(A, B|\{N_k\}, I)dB. \tag{3.10}$$

也可以对 A 积分得到关于背景振幅 B 的后验 pdf:

$$prob(B|\{N_k\},I) = \int_0^\infty prob(A,B|\{N_k\},I)dA. \tag{3.11}$$

四组边缘化后的数据如图所示，实验设置与是一致的，图 3.2 和图 3.3 中数据集对应的四组边缘分布和后验 pdf 画在图 3.4 和图 3.5 中，图 3.4、3.5 中更容易看到不同的实验设置对推断 A 和 B 的值的可靠性的影响。

这里应该注意到 $prob(A|\{N_k\},I)$ 与 $prob(A|\{N_k\},B,I)$ 是不同的，第一个 prob 表示了对于 B 的值的无知，但是第二个 prob 表示已知 B，在图 3.4、3.5 中用虚线表示已知 B=2 的后验。

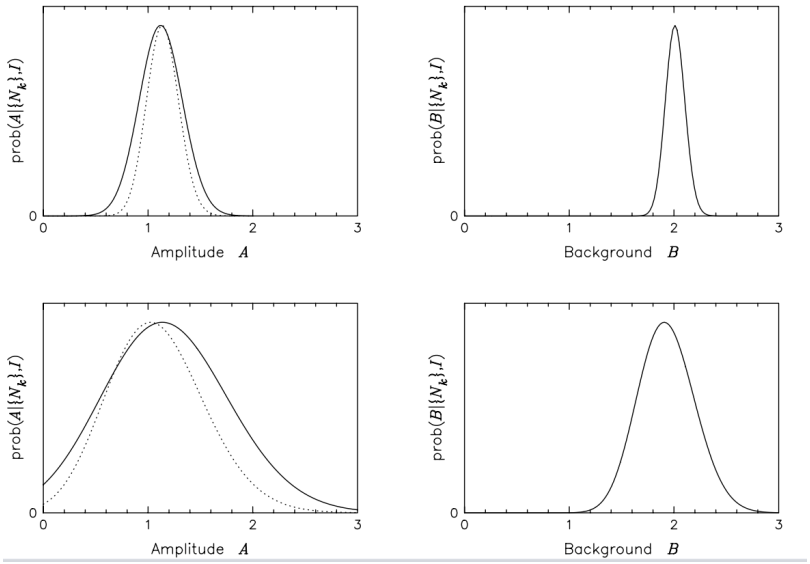

图 3.4: 一、二组数据

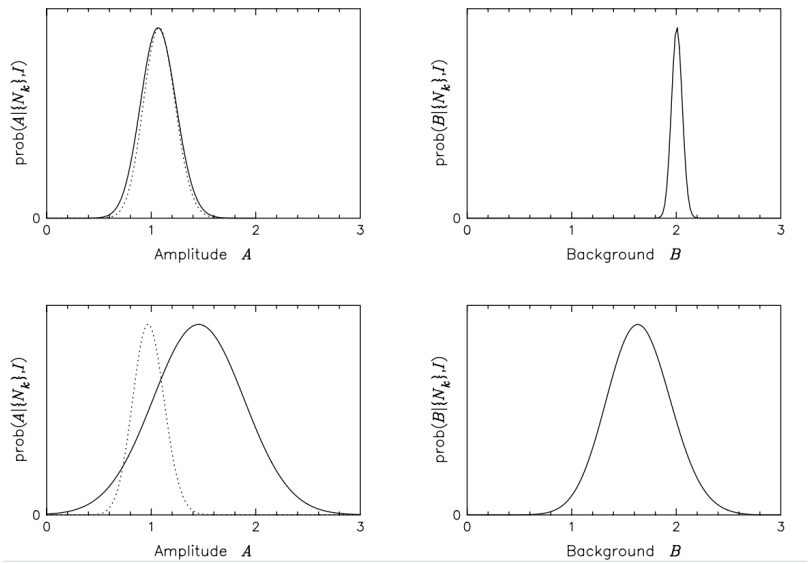

图 3.5: 三、四组数据

与边缘化后验对比：

最后一个数据集，相比边缘化 pdf，虚线的后验概率显著缩小，因为第四组实验并不能很好的区分信号振幅和背景，所以已知 B 可以更好地对 A 进行分析。

第三组数据集，测量范围远超过信号峰范围的情况下，实线和虚线差异最小。
当能够很好地区分背景和信号峰时，对 B 进行单独的校准实验所得到的东西很少, 如果数据集被严重截断 (比如第四组数据)，或者背景是高度结构化的，这些额外的信息 (比如已知 B 的值) 非常有益。

假设信号峰的形状和位置已知，在高斯模型方程中

$$D_k = n_0[Ae^{-(x_k-x_0)^2/2\omega^2} + B],$$

隐含了背景信息 I 中的 ω 和 x_0 的值。如果不知道这些值，从边缘化的讨论中可以看出，面对妨害参数应该整合相关变量：

$$prob(A, B|\{N_k\}, I) = \int\int prob(A, B, \omega, x_0|\{N_k\}, I)d\omega dx_0, \tag{3.12}$$

假设 I 包含用高斯模型模型提供理想数据，但不一定包含对高斯峰的宽度和位置的知识。二重积分下的四参数后验 pdf 本身可以展开：

$$prob(A, B, \omega, x_0|\{N_k\}, I) \propto prob(\{N_k\}|A, B, \omega, x_0, I) \times prob(A, B, \omega, x_0|I). \tag{3.13}$$

右边第二项是 A, B, ω, x_0 的 prior pdf，可以展开为

$$prob(A, B, \omega, x_0|I) = prob(A, B|I) \times prob(\omega, x_0|I). \tag{3.14}$$

如果已知高斯峰的宽度和位置，那么 ω, x_0 的 prior 是非常尖锐的。在完全确定这两个参数的极限下，有

$$prob(\omega, x_0|I) = \delta(\omega - 2.12)\delta(x_0), \tag{3.15}$$

非零点在 $\omega = 2.12$(即 FWHM=5) 和 $x_0 = 0$，在这种情况下，积分方程 (3.12) 非常容易计算

$$prob(A, B|\{N_k\}, I) = \int\int prob(A, B, \omega, x_0|\{N_k\}, I)d\omega dx_0,$$

代入 delta 函数

$$prob(A, B|\{N_k\}, I) \propto prob(\{N_k\}|A, B, \omega = 2.12, x_0 = 0) \times prob(A, B|I) \tag{3.16}$$

这个表达式回到了方程 (3.7)

$$prob(A, B|\{N_k\}, I) \propto prob(\{N_k\}|A, B, I) \times prob(A, B|I)$$

如果不知道 ω 和 x_0 的值，就必须为这些参数 (包括 A 和 B) 分配一个较宽的先验。边缘化积分相比于已知参数时的 delta 函数，需要做更多的计算，既可以数值计算，也可以解析近似。

3.1.2 绑定数据 Binning the data

用直方图将数据绘制在图 3.2、3.3 中时，提到用直方图是因为实验测量通常在有限宽度的通道中检测计数。这意味着，对于理想数据 D_k，eqn(3.1)，实际上应该被写成第 k 个数据箱上的一个积分：

$$D_k = \int_{x_k-\Delta/2}^{x_k+\Delta/2} n_0[Ae^{-(x_k-x_0)^2/2\omega^2} + B]dx, \tag{3.17}$$

这里假设所有的测量通道都有相同的宽度 。只要箱的宽度不太大，方程式（3.12）的积分可以近似为长矩形的面积：

$$D_k = n_0[Ae^{-(x_k-x_0)^2/2\omega^2} + B]\Delta. \tag{3.18}$$

因此，方程 (3.1) 是合理的，因为固定大小的 Δ 可以被吸收进 n_0。新的 n_0 反映了进行实验测量的时间和 “收集区域” 的大小。然而，容器宽度 Δ 并不总是由检测器的物理大小决定的，但通常被选择为足够大，以便在由此产生的复合数据通道中有合理数量的计数。

对图 3.2 中第一个面板的实验设置对应的数据进行分析，但箱子变窄了 4 倍。

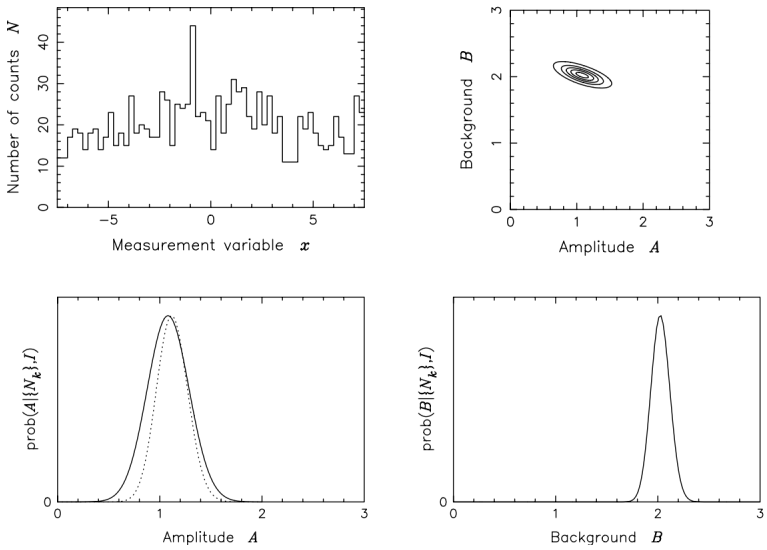

图 3.6:

这个数据集看起来更嘈杂，因为平均看，每个通道只有之前计数的四分之一。这张图还显示了 A 和 B 的后验 pdf，及其边缘分布，假设 $\omega = 2.12$ 和 $x_0 = 0$。. 与图 3.2 和图 3.4 中相应的 pdf 相比，所推断参数的可靠性几乎是相同的。但是这种分箱在处理测量数据较少时有优势。过于粗糙的装箱 (Δ 过大) 会破坏数据中的有用信息。比如用一个很宽的大箱子计数，这样会把所有的计数加成一个数字，这样会完全没办法从区分背景和信号的信息，同时 (3.18) 的公式近似也会出现问题。

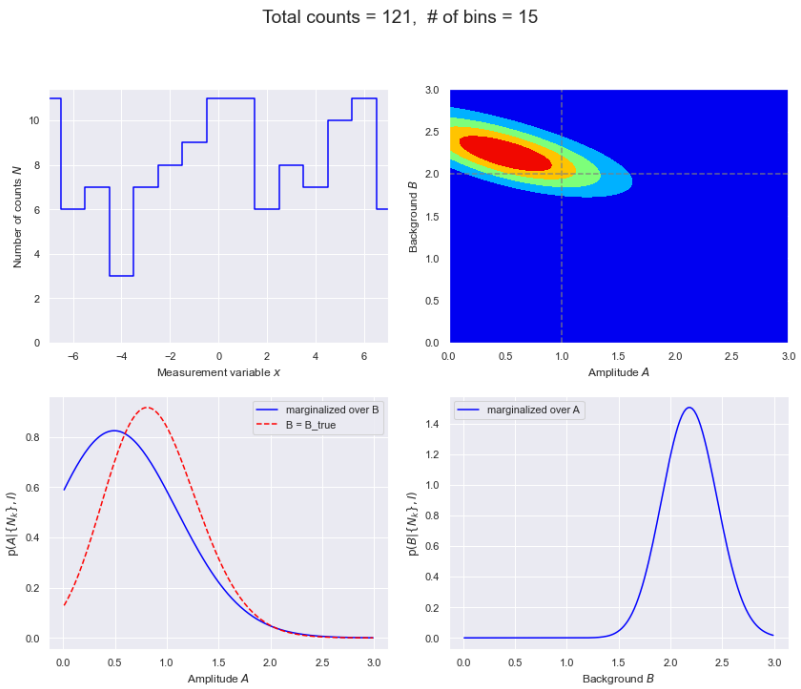

图 3.7: D max = 10,databins = 15,delta x = 1

在数据不充足时，将箱子间隔缩小得到的结果不稳定（见程序）。

3.2 Reliabiilties:best estimates,correlations and error-bars

对于多参数问题，如果用 $\{X_j\}$ 表示感兴趣的参数集，那么对于这些参数的最佳估计 $\{X_{Oj}\}$ 由后验 pdf:$prob(\{X_j\}|\{data\}, I)$ 的一阶导数给出

$$\left.\frac{\partial P}{\partial X_i}\right|_{\{X_{Oj}\}} = 0, \tag{3.19}$$

严格来说，还需要二阶导数小于 0，写出这个微分时，就隐含了假设后验关于这些参数是连续的。使用 P 的对数还是更加方便，所以

$$L = log_e[prob(\{X_j\}|\{data\}, I)]. \tag{3.20}$$

3.19 式也由 L 代替 P。首先考虑**两个变量**的具体情况：用 X 和 Y 表示。求解联立方程得到最佳估计 X_0, Y_0：

$$\left.\frac{\partial L}{\partial X}\right|_{X_0,Y_0} = 0 and \left.\frac{\partial L}{\partial Y}\right|_{X_0,Y_0} = 0 \tag{3.21}$$

这里

$$L = log_e[prob(X, Y|\{data\}, I)]$$

为了对该最佳估计的可靠性进行度量，需要要看二维后验 pdf 关于点 (X_0, Y_0) 附近的函数行为。如第 2.2 节所述，可以**通过泰勒展开**分析 pdf 的局部行为:

$$\begin{aligned} L = L\left(X_{\mathrm{O}}, Y_{\mathrm{O}}\right) + \frac{1}{2}\Bigg[& \left.\frac{\partial^2 L}{\partial X^2}\right|_{X_{\mathrm{o}}, Y_{\mathrm{O}}} \left(X - X_{\mathrm{O}}\right)^2 + \left.\frac{\partial^2 L}{\partial Y^2}\right|_{X_{\mathrm{o}}, Y_{\mathrm{O}}} \left(Y - Y_{\mathrm{O}}\right)^2 \\ & + 2\left.\frac{\partial^2 L}{\partial X \partial Y}\right|_{X_{\mathrm{o}}, Y_{\mathrm{o}}} \left(X - X_{\mathrm{O}}\right)\left(Y - Y_{\mathrm{O}}\right)\Bigg] + \cdots, \end{aligned} \tag{3.22}$$

其中 X_0, Y_0 由一阶导数 (3.21) 的条件给出, 并通过 $\partial^2 L/\partial X \partial Y = \partial^2 L/\partial Y \partial X$ 将等式化为三项。式中第一项 $L(X_0, Y_0)$ 是常数，对 posterior 的形状没有影响，由于 $\frac{\partial L}{\partial X}\big|_{X_0, Y_0} = 0$, $\frac{\partial L}{\partial Y}\big|_{X_0, Y_0} = 0$ 所以 $(X - X_0), (Y - Y_0)$ 一次方项均为 0，所以三个二次项是决定后验 pdf 宽度的主要因素，在对可信度进行分析中起着关键作用。为了可以扩展到更多系数的情况，所以写成矩阵形式，三个二次项用 Q 来表示 (不包含系数 1/2)，用矩阵形式表示可以写为

$$Q = \begin{pmatrix} X - X_{\mathrm{O}} & Y - Y_{\mathrm{O}} \end{pmatrix} \begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} X - X_{\mathrm{O}} \\ Y - Y_{\mathrm{O}} \end{pmatrix}, \tag{3.23}$$

其中，中间的 2×2 对称矩阵的分量由 L 的二阶导数给出，

$$A = \left.\frac{\partial^2 L}{\partial X^2}\right|_{X_0, Y_0}, \quad B = \left.\frac{\partial^2 L}{\partial Y^2}\right|_{X_0, Y_0}, \quad C = \left.\frac{\partial^2 L}{\partial X \partial Y}\right|_{X_0, Y_0}. \tag{3.24}$$

$$\begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}. \tag{3.25}$$

Q 在 X-Y 平面中的轮廓为画在图 3.7 中，在二次近似中是一条后验 pdf 不变的线，形状是是一个椭圆，以 $(X - X_0), (Y - Y_0)$ 为中心，方向和偏心都由 ABC 的值决定，对于给定的等高级 Q=k，k 是一个常数，ABC 也控制着椭圆的大小。椭圆性质由 (3.23) – (3.20) 中定义的二阶导数矩阵的特征值　和特征向量 e 决定。特征值 λ_1 和 λ_2，与椭圆沿其主方向的宽度的平方成反比，并且 λ_1 和 λ_2 都是负值才能保证点 $(X - X_0), (Y - Y_0)$ 是最大值而不是最小值或鞍点。

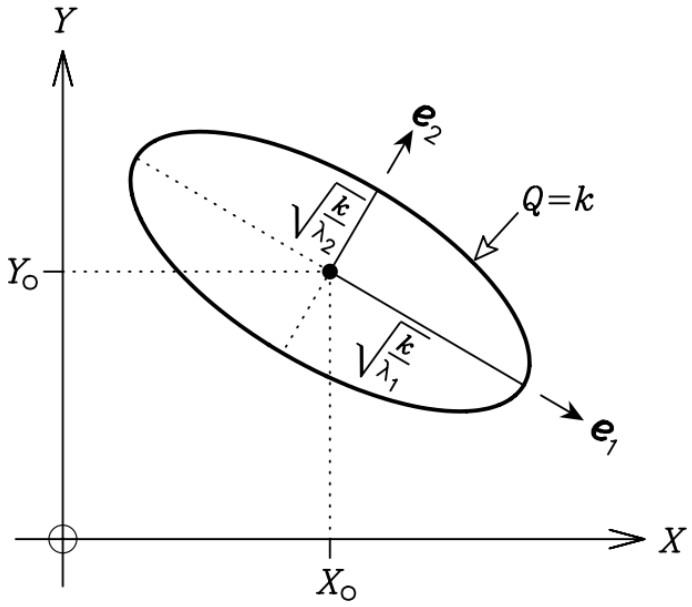

图 3.8:

并且 λ_1 和 λ_2 都是负值才能保证点 $(X-X_0),(Y-Y_0)$ 是最大值而不是最小值或鞍点，这点对于 ABC 的要求为

$$A<0,\quad B<0,\quad AB>C^2.$$

为什么 λ_1 和 λ_2 都是负值:

求最大估计不仅要求 L 一阶导等于 0，而且要求二阶导小于 0，而 Q 正代表 L 的二阶导项，Q:

$$Q=\begin{pmatrix} X-X_{\mathrm{O}} & Y-Y_{\mathrm{O}} \end{pmatrix}\begin{pmatrix} A & C \\ C & B \end{pmatrix}\begin{pmatrix} X-X_{\mathrm{O}} \\ Y-Y_{\mathrm{O}} \end{pmatrix}, \tag{3.26}$$

D 是对称矩阵，对角化后可以写作

$$Q=\begin{pmatrix} X-X_{\mathrm{O}} & Y-Y_{\mathrm{O}} \end{pmatrix}\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}\begin{pmatrix} X-X_{\mathrm{O}} \\ Y-Y_{\mathrm{O}} \end{pmatrix}, \tag{3.27}$$

行列式的值为 $\lambda_1(X-X_0)^2+\lambda_1(Y-Y_0)^2<0$，所以 λ_1 和 λ_2 都小于 0。同样的 D 的行列式可以写为 $AB-C^2$

分析可靠性

1. 如果椭圆不倾斜

那么 error bar 就只与 λ 的根号分之一有关。

2. 如果椭圆倾斜

只对一个参数感兴趣，边缘化

$$prob(X|\{data\},I)=\int_{-\infty}^{\infty}prob(X,Y|\{data\},I)dY.$$

$prob(X,Y|\{data\},I)=exp(L)\propto exp(Q/2)$，假设后验 pdf 没有明显的被先验的边界截断，则可以利用高斯积分将上式化为

$$prob(X|\{data\},I)\propto exp(\frac{1}{2}[\frac{AB-C^2}{B}](X-X_0)^2)$$

其中省略了归一化常数，并与以下两式进行对比，

$$prob(X|\{data\},I)\approx A\exp[\frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_0}(X-X_0)^2],$$

$$prob(x|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}exp[-\frac{(x-\mu)^2}{2\sigma^2}],$$

可以看出，X 的边缘化分布是一维高斯 pdf，最佳估计仍然是 X_0，其相关的误差条 σ_X 如下

$$\sigma_X=\sqrt{\frac{-B}{AB-C^2}}A. \tag{3.28}$$

同理可得 Y 的边缘化误差条 σ_Y，

$$\sigma_Y=\sqrt{\frac{-A}{AB-C^2}}. \tag{3.29}$$

虽然上面的 σ_X 和 σ_Y 的表达式提供了对最佳估计的可靠性的一个度量方法，但它们描绘的画面并不完整。为了更好地理解这一点，必须对 σ_X 和 σ_Y 的分母进行更深入的了解。

$AB-C^2$ 实际上是实对称矩阵 D 的行列式，对于这样一个实对称矩阵是可以正交化的，所以 D 对应的行列式的值等于矩阵 D 的特征值的乘积。因此，如果 λ_1 或 λ_2 非常小，使得图 3.7 中的椭圆在其主方向之一上被极度拉长并且，相应的，$AB-C^2\to 0$。这时，除了 C=0 时的特殊情况外，σ_X,σ_Y 都会很大。当然 pdf 也可能在一个方向非常尖锐而在另一个方向非常宽。

到目前为止，从方程（2.20）开始，认为误差条表示高斯 pdf 的宽度：$FWHM\approx 2.35\sigma$。另一种 r 认识是根据后验的方差，这也给出了 pdf 宽度的度量。

定义如下

$$\left\langle(X-\mu)^2\right\rangle=\int(X-\mu)^2prob(X|\{data\},I)dX, \tag{3.30}$$

其中 μ 是平均值 $< X >$，$< X >$ 是后验均值，表达式为

$$< X >= \int X prob(X|\{data\}, I) dX.$$

对于一维正态分布，

$$\left\langle (X-\mu)^2 \right\rangle = \sigma^2 \tag{3.31}$$

误差条的定义可以扩展到多个参数的情况，比如刚才谈到的两个参数，即

$$\sigma_X^2 = \left\langle (X - X_0)^2 \right\rangle = \int\int (X - X_0)^2 prob(X, Y|\{data\}, I) dX dY, \tag{3.32}$$

当在方程 ns（3.22）–（3.24）的二次近似下计算二重积分时，σ_X 与 eqn（3.26）中相同。σ_Y 的对应表达式是相似的。同时描述 X 和 Y 的是协方差 σ_{XY}^2

$$\begin{aligned} \sigma_{XY}^2 &= \langle (X - X_{\mathrm{O}})(Y - Y_{\mathrm{O}}) \rangle \\ &= \iint (X - X_{\mathrm{O}})(Y - Y_{\mathrm{O}}) \,\mathrm{prob}(X, Y \mid \{data\}, I) \mathrm{d}X \; \mathrm{d}Y, \end{aligned} \tag{3.33}$$

协方差 σ_{XY}^2 可以描述参数 X 和 Y 之间的相关性。对一个参数的估计对另一个参数的推断值影响很小或没有影响时，那么协方差的大小与方差项相比可以忽略不计（$|\sigma_{XY}^2| \ll \sqrt{\sigma_X^2 \sigma_Y^2}$）。如果对一个参数的估计大了，导致对其他的参数估计平均值也大了，那么 $X - X_0$ 与 $Y - Y_0$ 是正相关的，如果低估也是如此，所以当 $X - X_0$ 是负时，$Y - Y_0$ 通常是负的，那么偏差的乘积的期望值将是正的：协方差将大于零。如果 $X - X_0$ 与 $Y - Y_0$ 是负相关的。

当 pdf 取和前面一样的近似的时候，可以得到

$$\sigma_{XY}^2 = \frac{C}{AB - C^2}. \tag{3.34}$$

将 $\sigma_X^2, \sigma_Y^2, \sigma_{XY}^2$ 的表达式进行对比可以发现

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \sigma_{XY}^2 & \sigma_Y^2 \end{pmatrix} = \frac{1}{AB - C^2} \begin{pmatrix} -B & C \\ C & -A \end{pmatrix} = - \begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1} \tag{3.35}$$

这个被称为协方差矩阵，（a）当 C=0 时，σ_{XY}^2 也等于 0，意味着参数 X 和 Y 的估计是不相关的，对应的，椭圆的两个主轴分别于 X 轴和 Y 轴平行。所示随着 C 的增大，后 pdf 变得越来越倾斜和拉长；这反映了参数的估计之间的相关性的增加。（b）当 $C < 0$ 时，也意味着 $\sigma_{XY}^2 < 0$ 所以参数估计之间呈负相关。（c）当 $C > 0$ 时，也意味着 $\sigma_{XY}^2 > 0$ 所以参数估计之间呈正相关。

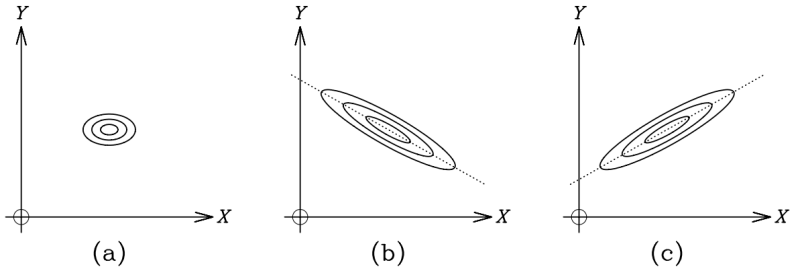

图 3.9:

在 $C = \pm\sqrt{AB}$ 的极限下, 椭圆轮廓将在一个方向上无限宽，这时也可以体现出先验的重要性，先验会对椭圆进行截断。与 x 轴的夹角的正切值，即斜率为 $\sqrt{\frac{A}{B}}$?

3.2.1 推广

把双变量分析推广到有多个参数时的情况，对有 M 个参数的 $\{X_j\}$ 进行分析，根据其对数 L

$$L = log_e[prob(\{X_j\}|\{data\}, I)].$$

对最优值 $\{X_{0j}\}$（或表示为 X_0）的要求，可以用联立方程来表示:

$$\left.\frac{\partial L}{\partial X_i}\right|_{\mathbf{X_0}} = 0 \tag{3.36}$$

对于多元的情况，泰勒展开式写为

$$L = L\left(\boldsymbol{X}_{\mathrm{O}}\right) + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left.\frac{\partial^2 L}{\partial X_i \partial X_j}\right|_{\boldsymbol{X}_{\circ}}\left(X_i - X_{\mathrm{O}i}\right)\left(X_j - X_{\mathrm{O}j}\right) + \cdots \tag{3.37}$$

忽略高阶项后，取 e 指数，后验 pdf 可以写为下面的形式，

$$\mathrm{prob}(\boldsymbol{X} \mid \{data\}, I) \propto \exp\left[\frac{1}{2}\left(\boldsymbol{X} - \boldsymbol{X}_{\mathrm{O}}\right)^{\mathrm{T}} \boldsymbol{\nabla}\boldsymbol{\nabla} L\left(\boldsymbol{X}_{\mathrm{O}}\right)\left(\boldsymbol{X} - \boldsymbol{X}_{\mathrm{O}}\right)\right] \tag{3.38}$$

这里的 $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ 表示二阶微分项，是 $M \times M$ 阶矩阵，第 ij 个矩阵元为

$$\frac{\partial^2 L}{\partial X_i \partial X_j}$$

可以看作是之前的矩阵 Q 的推广。方程（3.38）的 pdf 被称为多元高斯分布。这个 pdf 的归一化系数为

$$C = \frac{\sqrt{det(\boldsymbol{\nabla}\boldsymbol{\nabla}L)}}{(2\pi)^{\frac{M}{2}}}$$

最佳估计由 $\boldsymbol{\nabla}L(X_0)$ 给出，通过与高斯分布表达式进行对比，可以看出 $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ 与 $-1/\sigma^2$ 类似，说明后验的宽度应该与二阶导数矩阵的倒数有关。实际上，

$$\boldsymbol{\sigma}_{ij}^2 = \langle (X_i - X_{\mathrm{O}i})(X_j - X_{\mathrm{O}j})\rangle = \boldsymbol{H}^{-1}. \tag{3.39}$$

是方程 3.35 和 3.33 的概括。对角线元素的平方根 (i=j) 对应于相关参数的（边缘）误差条；非对角线分量 (i=j) 对应于 X_i, X_j 的推断值之间的相关性。

3.2.2 不对称和多模态的后验 pdfs

上面的分析依赖于方程 3.38 的有效性，

$$\mathrm{prob}(\boldsymbol{X} \mid \{data\}, I) \propto \exp\left[\frac{1}{2}(\boldsymbol{X} - \boldsymbol{X}_{\mathrm{O}})^{\mathrm{T}}\, \boldsymbol{\nabla}\boldsymbol{\nabla}L(\boldsymbol{X}_{\mathrm{O}})(\boldsymbol{X} - \boldsymbol{X}_{\mathrm{O}})\right]$$

与之前所讨论的类似，当后验 pdf 不对称时，采用二次近似的方法就不太合适了，而且协方差的积分定义与微分矩阵的联系之间会出现问题（3.33 和 3.39）。并且想要通过边缘化而对参数的后验 $prob(X_j|\{data\}, I)$ 进行研究的话，会发现很难对 pdf 进行多元积分。

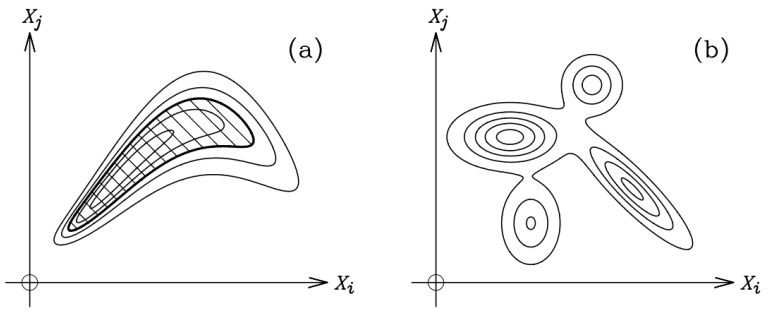

图 3.10:

后验还可能有多个最大值即多模态，这样的 pdf 如图 3.9(b) 所示，对于某些类型的问题，无论数据的质量如何，这种多模态性质都可以持续存在。当其中一个概率凸点比其他概率凸点大得多时，跟第二章提到的一样，通过对全局最大值做二次展开，忽略其他的小值，还是可以定义最佳估计并表达它可靠性。. 如果有几个最大值大小相似，可以对所有的凸点附近进行二次展开，并把这些最佳估计和它相关的协方差矩阵列一个表，但是如果重要解的数量非常大，这

样解起来就比较麻烦。实际上即使多模态有一个明显的单一最佳估计，在解的过程中也会遇到一些麻烦，会在 3.4 中提到。

椭圆方程的一般形式为

$$ax^2 + bxy + cy^2 = 1$$

，写作矩阵形式为

$$X^T \begin{pmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{pmatrix} X = 1, \text{其中} X = \begin{pmatrix} x \\ y \end{pmatrix} \text{设} A = \begin{pmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{pmatrix}$$

这种形式下无法看出椭圆的特征，所以利用正交变换将矩阵进行化简由于矩阵 A 为实对称矩阵，所以其一定能实现对角化，即一定存在一个正交矩阵 Q，有

$$Q^T \begin{pmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{pmatrix} Q = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} \text{，其中} Q = (\alpha_1, \alpha_2)$$

其中 α_1, α_2 是矩阵 A 的两个正交且长度为一的特征向量，λ_1, λ_2 是他们所对应的特征值。设

$$X = QY, \text{其中:} \quad Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

则有

$$X^T AX \to Y^T Q^T AQY = Y^T \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} Y = \lambda_1 y_1^2 + \lambda_2 y_2^2 = 1$$

椭圆方程的一般形式化为

$$\lambda_1 y_1^2 + \lambda_2 y_2^2 = 1 \Rightarrow \frac{y_1^2}{\left(\frac{1}{\sqrt{\lambda_1}}\right)^2} + \frac{y_2^2}{\left(\frac{1}{\sqrt{\lambda_2}}\right)^2} = 1$$

$\frac{1}{\sqrt{\lambda_1}}$ 和$\frac{1}{\sqrt{\lambda_2}}$ 一个为长半轴长一个为短半轴长。这个就是特征值和椭圆长短轴的关系。

接下来看特征向量与椭圆的关系

已知

$$X = QY \to Y = Q^T X = \begin{pmatrix} \alpha_1^T \\ \alpha_2^T \end{pmatrix} X$$

所以

$$Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \alpha_1^T \\ \alpha_2^T \end{pmatrix} X$$

可以看出

$$y_1 = \alpha_1^T X, \quad y_2 = \alpha_2^T X$$

根据椭圆的性质，$y_1 = 0$ 或 $y_2 = 0$ 对应的分别是椭圆长轴或短轴所在的两条线。并且注意到 α_1, α_2 是正交的，所以有

$$y_1 = 0 \Rightarrow y_1 = \alpha_1^T X = 0 \Rightarrow X = k_1\alpha_2 y_2 = 0 = y_2 = \alpha_2^T X = 0 \Rightarrow X = k_2\alpha_1$$

由于

$$X = \begin{pmatrix} x \\ y \end{pmatrix}.$$

所以，比如设

$$\alpha_2 = \begin{pmatrix} a \\ b \end{pmatrix} \text{ ，则} y_1 = 0 \Rightarrow X = k_1\alpha_2 \Rightarrow X = \begin{pmatrix} k_1 a \\ k2b \end{pmatrix}$$

即此时 $x = k_1 a, y = k_1 b$，$\frac{y}{x} = \frac{b}{a}$，也就是 y_2 轴与特征向量 α_2 平行

3.3 回顾高斯噪声

第二章讨论高斯噪声时，由于只讨论单参数，所以假设 σ 已知，这里继续对高斯噪声进行讨论。实际上我们要找的后验是 $prob(\mu|\{x_k\},I)$ 而不是 $prob(\mu|\{x_k\},\sigma,I)$，所以需要对高斯分布进行边缘化积分

$$\text{prob}\left(\mu \mid \{x_k\}, I\right) = \int_0^\infty \text{prob}\left(\mu, \sigma \mid \{x_k\}, I\right) \mathrm{d}\sigma \tag{3.40}$$

积分项可以写为似然函数与先验的乘积

$$\text{prob}\left(\mu, \sigma \mid \{x_k\}, I\right) \propto \text{prob}\left(\{x_k\} \mid \mu, \sigma, I\right) \times \text{prob}(\mu, \sigma \mid I) \tag{3.41}$$

数据独立时，N 组数据下，似然（模型）跟之前一样可以写成高斯连乘的形式，

$$\text{prob}\left(\{x_k\} \mid \mu, \sigma, I\right) = (\sigma\sqrt{2\pi})^{-N} \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{N}\left(x_k-\mu\right)^2\right] \tag{3.42}$$

先验函数 $\text{prob}(\mu, \sigma \mid I)$ 由于对于参数 μ, σ 的一无所知，为了简单使用一个平坦 pirior：

$$\text{prob}(\mu, \sigma \mid I) = \begin{cases} \text{constant} & \text{for } \sigma > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.43}$$

因此，

$$\text{prob}\left(\mu \mid \{x_k\}, I\right) \propto \int_0^\infty t^{N-2} \exp\left[-\frac{t^2}{2}\sum_{k=1}^{N}\left(x_k-\mu\right)^2\right] \mathrm{d}t, \tag{3.44}$$

使用了变量代换，这里的 $t = 1/t$，$(d\sigma = -dt/t^2)$，利用高斯积分公式积分结果近似为

$$\text{prob}\left(\mu \mid \{x_k\}, I\right) \propto \left[\sum_{k=1}^{N}\left(x_k-\mu\right)^2\right]^{-(N-1)/2}. \tag{3.45}$$

可以从这个边缘后验的对数的一阶和二阶导数中得到最佳估计 μ_0 和其可靠性的度量。因此，

$$\left.\frac{\mathrm{d}L}{\mathrm{d}\mu}\right|_{\mu_\circ} = \frac{(N-1)\sum\left(x_k-\mu_\circ\right)}{\sum\left(x_k-\mu_\circ\right)^2} = 0 \tag{3.46}$$

只有当分子等于 0 时才满足，由此得到 μ_0，结果与第二章得到的结果是一样的，

$$\mu_0 = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{3.47}$$

μ_0 仍然由 x_k 的算术平均给出，对于二阶导数，

$$\left.\frac{\mathrm{d}^2 L}{\mathrm{d}\mu^2}\right|_{\mu_\mathrm{o}} = -\frac{N(N-1)}{\sum\left(x_k-\mu_\mathrm{o}\right)^2}. \tag{3.48}$$

第二章讨论得到，最佳估计的 error bar 是由负 [二阶导数的平方根的倒数] 给出的，由此进行总结，

$$\mu = \mu_\mathrm{o} \pm \frac{S}{\sqrt{N}}, \quad \text{where } S^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(x_k-\mu_0\right)^2. \tag{3.49}$$

与 $\mu = \mu_\mathrm{o} \pm \frac{\sigma}{\sqrt{N}}$（其中 σ 已知）进行对比，可以看到与第二章的唯一区别是，以前假设已知的 σ 值被由数据得到的估计所取代。

3.3.1 t 和 χ^2 分布

第二章的高斯噪声问题中，后验 $prob(\{x_k\}|\mu,\sigma,I)$ 可以完全由最佳估计和 error bar 描述后验函数，因为当时对于高斯分布来说二次近似是精确的，但是当 σ 事先不知道时，这样仅用最佳估计和误差条来描述后验 pdf 就不正确了：eqn（3.49）的总结只是代表了对（边缘）后验问题 $prob(\mu|\{x_k\},I)$ 的一个有用的近似，实际的 pdf 在 eqn（3.45）中给出（不是高斯形式），如果根据数据得到的两个参数算术平均和 V 来重写 pdf，pdf 的形状更容易确定：

$$\sum_{k=1}^{N}\left(x_k-\mu\right)^2 = N(\bar{x}-\mu)^2 + V \tag{3.50}$$

这里的 V，

$$V = \sum_{k=1}^{N}\left(x_k-\bar{x}\right)^2 \tag{3.51}$$

后验函数（3.45）可以写作

$$\mathrm{prob}\left(\mu \mid \{x_k\}, I\right) \propto \left[N(\bar{x}-\mu)^2 + V\right]^{-(N-1)/2} \tag{3.52}$$

这个分布被称为 T 分布。N=3 时，此分布等于与灯塔例子中遇到的柯西分布

$$\mathrm{prob}\left(x_k \mid \alpha, \beta, I\right) = \frac{\beta}{\pi\left[\beta^2 + \left(x_k-\alpha\right)^2\right]} \tag{3.53}$$

在 $\mu = \bar{x}$ 处有一个最大值，一个 FWHM 与 $\sqrt{V}$ 成正比，随着 N 的增加，T 分布会的形状会越来越接近高斯分布，最终以 $\bar{x}$ 为中心。因此 μ_0 等于样本均值，随着 N 增大，误差条变得更有意义，通过 eqn（3.49）与 V 相关。

（t 分布: 研究 t 分布是为了可以根据有限数据的样本来估计呈正态分布且方差未知的总体的均值，如果方差已知或者样本数量足够多时应该使用正态分布进行估计，相比于正态分布来说，t 分布的容错率更高。T 分布的形状与 n 有关，也就是与自由度有关，样本数量为 N 时，自由度为 N-1，因为 pdf 中含有均值项 $\bar{x} = 1/N \sum x_k$，因此自由度减一，为 N-1）

之前提到将先验取为常数，现在考虑先验 **if we had assigned a pdf which was uniform with respect to log .**（把先验取为 σ 会得到下面的结果）这导致后验函数变为

$$\text{prob}\left(\mu \mid \{x_k\}, I\right) \propto \left[N(\bar{x}-\mu)^2 + V\right]^{-N/2} \tag{3.54}$$

与之前对比，指数上的 N-1 变为了 N，这也就是说得到了一个具有 N−1 自由度的 t 分布，而不是 N−2。这两个 pdf 形状非常相似，都在 $\bar{x}$ 处有一个最大值，但 eqn（3.54）的 pdf 峰比 eqn（3.52）稍窄一些。eqn（3.49）中，S^2 将由 $V/\sqrt{N}$ 而不是 $V/\sqrt{N-1}$ 给出；如果 N 足够大，这个差异可以忽略不计，因为当 N 大于等于 10 以后，t 分布与高斯分布就基本没有差异了。因此，当数据较少时，对先验进行均匀的分配会给出对最优值 μ_0 的可靠性稍微保守一点的估计，但实际上这个结果基本保持不变。

除了这些，关于测量中预期误差的大小，也就是对参数 σ 的推断是由（边缘）后验 pdf$prob(\sigma|\{data\}\ I)$ 描述的：

$$\text{prob}\left(\sigma \mid \{x_k\}, I\right) = \int_{-\infty}^{+\infty} \text{prob}\left(\mu, \sigma \mid \{x_k\}, I\right) \mathrm{d}\mu. \tag{3.55}$$

使用 3.42 和 3.43 的似然函数和平坦先验，再用

$$\sum_{k=1}^{N} \left(x_k - \mu\right)^2 = N(\bar{x}-\mu)^2 + V,$$

进行替换，可以得到

$$\text{prob}\left(\sigma \mid \{x_k\}, I\right) \propto \sigma^{-N} \exp\left(-\frac{V}{2\sigma^2}\right) \int_{-\infty}^{+\infty} \exp\left[-\frac{N(\bar{x}-\mu)^2}{2\sigma^2}\right] \mathrm{d}\mu. \tag{3.56}$$

这里的 $\sigma > 0$，由于上面的积分项结果与 σ 成正比，所以得到结果

$$\text{prob}\left(\sigma \mid \{x_k\}, I\right) \propto \sigma^{1-N} \exp\left(-\frac{V}{2\sigma^2}\right). \tag{3.57}$$

通过替换 $X = V/\sigma^2$ 可以看出这个结果与 χ^2：卡方分布有关系。根据第二章参数估计的分析，可以从 eqn（3.57）的对数的一阶和二阶导数中推导出最佳估计及其误差条，可以通过 σ 对参

数进行总结：

$$\sigma = \sigma_{\rm o} \pm \frac{\sigma_{\rm o}}{\sqrt{2(N-1)}}, \tag{3.58}$$

其中最优值是 $\sigma_0 = V/(\sqrt{N-1})$，虽然 eqn（3.58）允许 σ 的负值存在，但这只是表明当数据数量较小时，eqn（2.18）的二次展开近似对非对称后验 pdf 的近似较差，给出 $log\sigma$ 的估计和误差条会更好，将在第 3.6 节中学习。

3.4 算法

之前已经有了很多例子，展示了如何使用概率论的方法来计算感兴趣的量的后验 pdf，为了总结这种推断结果，通常需要找到它的最大值。有的时候可以通过分析来完成，但更通常要使计算机解决。计算机的优点在于它能够进行数值计算（和图形绘图），比手动操作要快得多。这节只是对可能出现的问题，给出可能用于处理它们的程序类型指示。第 9 章和第 10 章中给出了一种更先进的贝叶斯计算的蒙特卡罗方法说明。

这节中的内容材料与算法有关，而不是数据分析的基本原理。也就是说，用于寻找（而不是定义）最优解的技术方法纯粹是基于实际考虑，使用它们来解决概率问题的事实本质上是巧合。

3.4.1 直接法

对于一个只有单参数的问题，找到后验 pdf 的最大值的最直接的方法是将它绘制出来，这种方法需要将横轴划分为有限数量的网格点，表示参数的可能值，并评估每个参数的后验概率，沿着纵轴绘制后验，可以简单地找到最大概率的值。这种基本算法的最大优点是，它可以给出关于参数值的推断的完整图像（通常几百个网格点就足够了），后 pdf 是不对称的、多模态、甚至可微性的问题也无关紧要。这种蛮力方法很容易推广到双参数问题：只是将后验概率垂直地绘制在一个二维的点网格上

这就是在图 3.2（椭圆等 pdf 图）中所做的事情。除了上面提到的所有优点外，还很容易获得边缘分布：简单地将 X 或 Y 方向上的概率相加。即使是由先验 pdf 施加的截止也不会对这种数值积分造成困难，这可以从图 3.2 中的低计数示例中看到。

但是这种直接的方法，在双参数之后的多参数问题中，很快就变得不切实际。除了如何在一张纸上显示多个变量的函数的问题外，如果每个轴被分成 10 个离散点，一个 M 参数问题就需要 10^M 的评估。单参数的计算时间可能只需要一秒，但是七个参数就需要一天，对于多参量的问题需要找到更高效的算法。

3.4.2 线性

在第 3.2 节中，用一个（列）向量 $\mathbf{X}$ 的分量来表示一组 M 个参数 X_j，对它们值的最佳估计的条件，用 $\boldsymbol{X_0}$ 表示，可以写成

$$\boldsymbol{\nabla} L\left(\boldsymbol{X}_{\mathrm{O}}\right)=0, \tag{3.59}$$

其中，$\boldsymbol{\nabla} L$ 的第 j 个元素由后验 pdf 的对数的偏导数 $\partial L/\partial X_j$ 给出 (在 $\mathbf{X}$=$\mathbf{X}$o 处计算)。方程（3.59）是一组 M 个联立方程的非常紧凑的符号，一般来说，除非它们是线性的，否则难以解决。也就是说，**如果能够**将 $\boldsymbol{\nabla} L$ 重新排列成一种类似于**直线**的形式

$$\boldsymbol{\nabla} L=\boldsymbol{H}\boldsymbol{X}+\boldsymbol{C}. \tag{3.60}$$

其中 C 和 H 矩阵的分量都是常数，(3.59) 的解可以写出

$$\boldsymbol{X}_0=-\boldsymbol{H}^{-1}\boldsymbol{C}. \tag{3.61}$$

对式子 3.60 取微分，可以发现，矩阵 $\boldsymbol{\nabla}\boldsymbol{\nabla} L$ 是不变的

$$\boldsymbol{\nabla}\boldsymbol{\nabla} L=\boldsymbol{H} \tag{3.62}$$

所以所有的高阶导数都为零。协方差矩阵 σ^2，由负的 $\boldsymbol{\nabla}\boldsymbol{\nabla} L$ 的倒数给出，因此提供了对后验 pdf 形状的完整描述：

$$\left[\boldsymbol{\sigma}^2\right]_{ij}=\left\langle\left(X_i-X_{\mathrm{O}i}\right)\left(X_j-X_{\mathrm{O}j}\right)\right\rangle=-\left[\mathbf{H}^{-1}\right]_{ij} \tag{3.63}$$

其中，X_{0j} 是 eqn（3.61）中的向量 $\boldsymbol{X}_0$ 的分量。或者使用 $\boldsymbol{X}_0\boldsymbol{H}=-\boldsymbol{C}$，这样可以避免计算矩阵的逆，简化运算。

这种矩阵计算所花费的计算时间趋于变量数 M 的三次方的量级，与之前 10^M 相比小了很多，个人计算机足已解决合理的参数估计问题。

如果矩阵 $\boldsymbol{H}$ 的行列式为 0 或者极小，则也很难求解最佳估计的方程。这样的情况下，结果对数据的微小变化导致的 $\boldsymbol{H}$ 的微小变化会引起结果的明显变化，并且相应的（边缘）误差条将很大。

与之前类似，如果 Q=k 的椭球体，其中 $Q=(\boldsymbol{X}-\boldsymbol{X_0})^T\boldsymbol{\Delta}\boldsymbol{\Delta} L(\boldsymbol{X_0})(\boldsymbol{X}-\boldsymbol{X_0})$，在其任何主方向上（几乎）无限长，对应着，就会发生 $\boldsymbol{H}$ 的行列式为 0 或者极小这种情况。在这种情况下，分析 H 的特征值和特征向量是有用的，主轴表示哪些参数的线性组合可以相互独立地推断，特征值给出了它们可以估计的可靠性。解决这种问题唯一真正方法是改善后 pdf 的特征，这可以通过获得更多的相关数据，或者通过补充有说服力的先验信息来实现。

3.4.3 迭代线性化

线性问题在解析和计算上都很方便，即使 $\boldsymbol{\nabla} L$ 不能完全写成 eqn（3.60）的形式，也值得尝试使用它。为了了解如何做到这一点，考虑 L 关于参数空间 $\boldsymbol{X}_1$ 中**任意点**的泰勒级数展开：

$$L = L\left(\boldsymbol{X}_1\right) + \left(\boldsymbol{X} - \boldsymbol{X}_1\right)^{\mathrm{T}} \boldsymbol{\nabla} L\left(\boldsymbol{X}_1\right) + \frac{1}{2}\left(\boldsymbol{X} - \boldsymbol{X}_1\right)^{\mathrm{T}} \boldsymbol{\nabla}\nabla L\left(\boldsymbol{X}_1\right)\left(\boldsymbol{X} - \boldsymbol{X}_1\right) + \cdots$$

之前一阶导项都没有保留因为总是在最优解处展开。对 $\boldsymbol{\nabla} L$ 进行展开，

$$\boldsymbol{\nabla} L = \boldsymbol{\nabla} L\left(\boldsymbol{X}_1\right) + \boldsymbol{\nabla}\boldsymbol{\nabla} L\left(\boldsymbol{X}_1\right)\left(\boldsymbol{X} - \boldsymbol{X}_1\right) + \cdots, \tag{3.64}$$

如果我们忽略右边的高阶项，使 eqn（3.63）可以重新排列为 eqn（3.60）的线性形式，则 eqn（3.59）的解为

$$\boldsymbol{X}_{\mathrm{O}} \approx \boldsymbol{X}_1 - \left[\boldsymbol{\nabla}\boldsymbol{\nabla} L\left(\boldsymbol{X}_1\right)\right]^{-1} \boldsymbol{\nabla} L\left(\boldsymbol{X}_1\right), \tag{3.65}$$

当 $\boldsymbol{X_1} = \boldsymbol{X_0}$，或者 $\boldsymbol{\nabla} L$ 真的是线性的时，这种关系将是精确的，只要 $\boldsymbol{X_1}$ 足够接近最优估计，这都将是一个合理的近似。因此提出了一个迭代算法：(i) 猜测一个 $\boldsymbol{X_1}$(ii) 在 $\boldsymbol{X} = \boldsymbol{X_1}$ 处计算梯度向量 $\boldsymbol{\nabla} L$ 和二阶导数矩阵 $\boldsymbol{\nabla}\boldsymbol{\nabla} L$，(iii) 通过将 eqn（3.65）的右侧等同于 $\boldsymbol{X_2}$，计算改进的估计 $\boldsymbol{X_2}$，(iv) 重复这个过程，直到 $\boldsymbol{\nabla} L = 0$。

上述过程被称为牛顿-拉夫逊算法。它是寻找函数 $f(x_0) = 0$ 的根的数值方法的推广。在这种情况下，函数是多元的：$\boldsymbol{\nabla} L(\boldsymbol{X}_0) = 0$，可以用递归关系来总结它

$$\boldsymbol{X}_{N+1} = \boldsymbol{X}_N - \left[\boldsymbol{\nabla}\boldsymbol{\nabla} L\left(\boldsymbol{X}_N\right)\right]^{-1} \nabla L\left(\boldsymbol{X}_N\right). \tag{3.66}$$

其中 $\boldsymbol{X_N}$ 是对 N−1 迭代后解的估计，只要一开始的猜测接近最优解，则迭代会很快收敛至最优解。

迭代过程的稳定性通常可以通过降低迭代速度来提高，减小迭代的步长，这意味着从 $\boldsymbol{X_N}$ 到 $\boldsymbol{X_{N+1}}$ 可以做一个比 eqn（3.66）稍小一点的变化是有利的。虽然可以很容易地通过将 eqn（3.66）右边的矩阵向量乘一个分数常数来实现，但还是选择了通过在 $\boldsymbol{\nabla}\boldsymbol{\nabla} L$ 的所有对角元素中添加一个小的的数字 c 来达到类似的效果：

$$\boldsymbol{X}_{N+1} = \boldsymbol{X}_N - \left[\boldsymbol{\nabla}\nabla L\left(\boldsymbol{X}_N\right) + c\mathbf{I}\right]^{-1} \nabla L\left(\boldsymbol{X}_N\right), \tag{3.67}$$

其中 I 是单位矩阵。这一奇怪选择的产生的原因最好从图 3.6 中的二次形式中理解，其中矩阵的性质由其特征值 $\{\lambda_j\}$ 和特征向量 $\{e_j\}$ 来描述；明确地，对于 $\boldsymbol{\nabla}\boldsymbol{\nabla} L$，这些是方程的解

$$[\boldsymbol{\nabla}\boldsymbol{\nabla} L]e_j = \lambda_j e_j, \tag{3.68}$$

如果把单位矩阵的倍数加到这个矩阵上，则会发现，特征向量不变，特征值的大小发生改变，对比图 3.7，可以看出椭圆的方向不变，主轴的宽度发生变化，相比于没有加 c 倍单位矩阵时，收敛步长确实变小。

$$[\boldsymbol{\nabla}\boldsymbol{\nabla}L + c\boldsymbol{I}]e_j = [\lambda_j + c]e_j, \tag{3.69}$$

对角线的增强没有改变椭圆的方向，也就是没有影响参数之间的相关性. 由于较小的特征值与较大的不确定性相关联，eqn (3.69) 通过选择性地减少它们的影响来稳定迭代算法。矩阵的逆与行列式的倒数有关，由于这是由特征值的乘积给出的，所以特征值很小时会导致 $(\boldsymbol{\nabla}\boldsymbol{\nabla}L)_{-1}$ 特别大，从而使得矩阵逆和一阶导乘积项很大，导致迭代步长很大，迭代不稳定。通过向 $(\boldsymbol{\nabla}\boldsymbol{\nabla}L)_{-1}$ 添加单位矩阵的一个小的（负的）倍数，在 $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ 很小时，通过常数 c 来稳定迭代，可以确保行列式的大小大于零，这导致混合二阶导数矩阵的逆的有限值大于 0，可以根据一阶导部分进行小步迭代。

除此之外，如果参数的数量太大，导致 $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ 矩阵难以存储或反转，那么通过共轭-梯度算法可以实现牛顿-拉夫逊过程。

要求 $\boldsymbol{\nabla}L(X_0) = 0$ 实际上是寻找一个平稳点的一个条件，有时可以在后验概率的极端尾部得到满足。因此，如果初始猜测 X_1 不够接近最优解，牛顿-拉夫逊过程将会发散（接近无穷大）。具体情况如图 3.10 所示

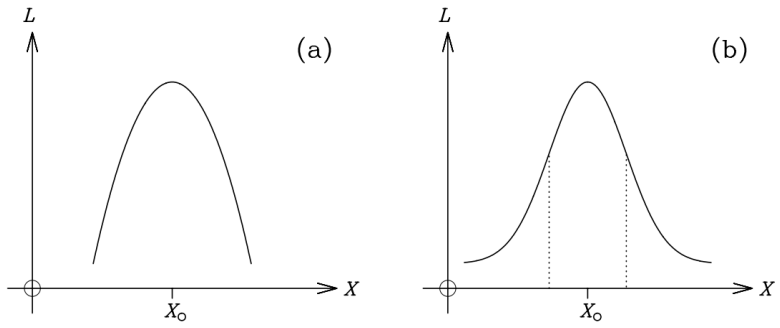

图 3.11:

两个后验 pdf 的对数说明：(a) 一个表现良好（类线性）问题，牛顿-拉夫逊程序将从任何起点收敛到 X_0；(b) 一个单峰 pdf，其中牛顿-拉夫逊算法将发散，除非初始猜测 (X_1) 在两条虚线内。对于图 3.10(b) 的情况，确保计算机程序向 X_0 收敛的一种方法是使用 Nelder-Mead 算法，可以在收敛开始时进行使用保证正确的收敛方向。

3.4.4 困难

在多模态 pdf 后验的情况下是最困难的优化任务。对于一两个个参数，最好的方法是画出整个 pdf，并寻找最大的值。但是正如前面提到的，这对于变量很多的情况来说是非常不切实际的。虽然基于梯度的算法是多元分析中最有效的算法，但它对多模态情况没有帮助：对局部斜率的了解并不能揭示全局最大值的位置。

介绍了几个算法

3.5 近似：最大似然和最小二乘

用向量 $\boldsymbol{D}$ 表示 N 组数据，用向量 $\boldsymbol{X}$ 表示 M 个参数，贝叶斯定理可以写为

$$\text{prob}(\boldsymbol{X} \mid \boldsymbol{D}, I) \propto \text{prob}(\boldsymbol{D} \mid \boldsymbol{X}, I) \times \text{prob}(\boldsymbol{X} \mid I), \tag{3.70}$$

令先验为常数，则有

$$\text{prob}(\boldsymbol{X} \mid \boldsymbol{D}, I) \propto \text{prob}(\boldsymbol{D} \mid \boldsymbol{X}, I), \tag{3.71}$$

后验直接与似然函数相关，因此最佳估计由后验最大值给出也就对应着数据概率最大，这被称为最大似然法。可以通过对似然函数本身进行化简简化过程，比如，假设数据是独立的，那么它们的联合 pdf 问题 prob($\boldsymbol{D}|\boldsymbol{X}$，I) 是由单个测量的概率的乘积给出的：

$$\text{prob}(\boldsymbol{D} \mid \boldsymbol{X}, I) = \prod_{k=1}^{N} \text{prob}\left(D_k \mid \boldsymbol{X}, I\right). \tag{3.72}$$

虽然之前已经使用过这个结果，但在这里强调，它遵循乘积规则，

$$\text{prob}\left(D_k, D_l \mid \boldsymbol{X}, I\right) = \text{prob}\left(D_k \mid D_l, \boldsymbol{X}, I\right) \times \text{prob}\left(D_l \mid \boldsymbol{X}, I\right). \tag{3.73}$$

并且，如果对一个数据的研究，不会影响另一个数据的取值，即

$$prob(D_k|D_l, \boldsymbol{X}, I) = prob(D_k|\boldsymbol{X}, I), \tag{3.74}$$

假设与实验测量相关的噪声可以合理的表示为一个高斯过程，

$$\text{prob}\left(D_k \mid \boldsymbol{X}, I\right) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left[-\frac{\left(F_k - D_k\right)^2}{2\sigma_k^2}\right], \tag{3.75}$$

这里的 I 隐含了：

1. 理想数据（没有噪声的）与参数之间的函数模型

$$F_k = f(\boldsymbol{X}, k), \tag{3.76}$$

2. 已知误差条 $\{\sigma_k\}$ 的预期大小。
$\text{prob}\,(x_k \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]$ 是第二章的高斯噪声表达式：高斯分布经常被用作描述与实验数据相关的噪声（或缺陷）的理论模型，该式表示第 k 个数据具有值 xk 的概率。似然函数可以表示为

$$\text{prob}(\boldsymbol{D} \mid \boldsymbol{X}, I) \propto \exp\left(-\frac{\chi^2}{2}\right), \tag{3.77}$$

其中，χ^2 为归一化残差的平方和（归一化后的残差叫做标准残差（$R_k = (F_k - D_k)/\sigma_k$），满足的是 N（0，1）的标准正态分布，在进行线性回归时，可以更简单的判断数据中的可疑值）：

$$\chi^2 = \sum_{k=1}^{N} \left(\frac{F_k - D_k}{\sigma_k}\right)^2. \tag{3.78}$$

由于取用均匀先验，后验与似然直接相关，根据（3.71）和（3.72），后验对数可以写为

$$L = \log_e[\text{prob}(\boldsymbol{X} \mid \boldsymbol{D}, I)] = \text{constant} - \frac{\chi^2}{2}. \tag{3.79}$$

由于后验的最大值会出现在 χ^2 最小时，因此相应的最优解 X_0 通常称为最小二乘估计。但是最大似然和最小二乘是基于前面的假设，再进行化简才成立的，比如数据彼此独立，噪声可以看作高斯分布等假设。如果没有这些假设，那么需要再从贝叶斯定理出发找一个新的方法。

最小二乘流行的原因之一是很容易应用，比如，如果函数关系（3.76）是线性的，则 $\boldsymbol{\nabla} L$ 可以写为 $\boldsymbol{\nabla} L = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{C}$ 的形式。
证明：第 k 个理想数据可以写为

$$F_k = \sum_{j=1}^{M} T_{kj} X_j + C_k, \tag{3.80}$$

其中 T_{kj}, C_k 的值都独立于 X_j，在矩阵向量符号中 $\boldsymbol{F} = \boldsymbol{T}\boldsymbol{X} + \boldsymbol{C}$。$\boldsymbol{\nabla} L$ 的第 j 个分量由下给出

$$\chi^2 = \sum_{k=1}^{N} \left(\frac{F_k - D_k}{\sigma_k}\right)^2,$$

$$\frac{\partial L}{\partial X_j} = -\frac{1}{2}\frac{\partial \chi^2}{\partial X_j} = -\sum_{k=1}^{N} \frac{(F_k - D_k)}{\sigma_k^2}\frac{\partial F_k}{\partial X_j}. \tag{3.81}$$

虽然可以通过（3.80）将 $\frac{\partial F_k}{\partial X_j} = T_{kj}$ 代入（3.81），并将其重新排列为 $\boldsymbol{\nabla} L = \boldsymbol{HX} + \boldsymbol{C}$ 的形式，但是在代数上这样做有些混乱。验证 $\boldsymbol{\nabla} L$ 的线性的一个简单方法是式子（3.81）对 X_i 微分，并注意二阶导数矩阵元都是常数。

$$\frac{\partial^2 L}{\partial X_i \partial X_j} = -\sum_{k=1}^{N} \frac{T_{ki} T_{kj}}{\sigma_k^2}, \tag{3.82}$$

由于 L 的所有高阶导数都同为零，后验 pdf 完全由最优解 $\boldsymbol{X_0}$ 及其协方差矩阵定义，协方差矩阵的分量与 $\boldsymbol{\nabla\nabla}\chi^2$ 或 χ^2 矩阵的两倍逆有关：

$$\langle (X_i - X_{\mathrm{O}i})(X_j - X_{\mathrm{O}j}) \rangle = -\left[(\boldsymbol{\nabla\nabla} L)^{-1} \right]_{ij} = 2\left[\left(\boldsymbol{\nabla\nabla}\chi^2 \right)^{-1} \right]_{ij}. \tag{3.83}$$

如果没有 eqn（3.79），参数 $\mathbf{X}$ 与数据线性相关，矩阵 $\boldsymbol{\nabla\nabla} L$ 是不变的这一有用性质通常不成立；这就是为什么最小二乘近似非常方便。例如
eqn（3.1）的函数模型

$$D_k = n_0[Ae^{-(x_k - x_0)^2/2\omega^2} + B],$$

其中给出了信号峰的形状和位置，D_k 相对于振幅 A 和背景 B 是线性的。然而，由于 eqn（3.9）中后验 pdf 的梯度向量不能重新排列成 eqn（3.50）的线性形式，

$$L = log_e[prob(A, B|N_k, I)] = constant + \sum_{k=1}^{M} [N_k log_e(D_k) - D_k],$$

因此最优解 (A_0, B_0) 很难解析出来，但是就像之前一样用直接的数值解法也可以找出最优点。

其实即使是泊松似然的情况，也可以通过使用最小二乘近似得到一个相当好的估计，这是因为，泊松分布 D>10 后分布就接近正态分布了，eqn（3.3）就开始具有高斯特征

$$prob(N|D) = \frac{D^N e^{-D}}{N!},$$

在大数的极限下，可以正式地证明 eqn（3.3）可以用一个正态分布很好地表示：

$$\mathrm{prob}(N \mid D) = \frac{D^N \mathrm{e}^{-D}}{N!} \propto \exp\left[-\frac{(N-D)^2}{2D} \right]. \tag{3.84}$$

根据之前对高斯分布的总结（$x_0 = \mu + \frac{\sigma}{\sqrt{N}}$），上面的概率分布可以总结为 $N = D + \sqrt{D}$，由于测量的计数大致等于期望值 N，用 $\sqrt{N}$ 代替 $\sqrt{D}$ 误差条，使指数的分母独立于参数 A 和 B，这样有助于线性化。在均匀先验下，后验 pdf 的对数可以很好地近似为 eqn（3.79），根据目前

的符号，相应的 χ^2 统计量由下给出

$$\chi^2 = \frac{(N-D)^2}{N},$$

$$\chi^2 = \sum_{k=1}^{N} \frac{(F_k - D_k)^2}{D_k}, \tag{3.85}$$

这里的 D_k 是在第 k 个数据通道中测量的计数数，F_k 是基于 eqn（3.1）的线性关系对它们的期望数的估计。这个结果，其中 eqn（3.78）中的误差条被数据的平方根 ($\sigma_k^2 = D_k$，或更常见的是 F_k) 所取代，与许多书中的 χ^2 的定义一致。

最小二乘有实际适合使用的情况是高斯似然函数和均匀先验。

3.5.1 拟合直线（最小二乘法的例子）

在数据分析中最常见的遇到的问题之一是直线与图形数据的拟合。假设给定一组 N 组数据 $\{Y_k\}$，在位置 $\{x_k\}$ 上测量，带有相关的误差条 $\{\sigma_k\}$。对描述它们的直线的两个参数的最佳估计是什么？这个作为最小二乘法的例子，具体情况如图

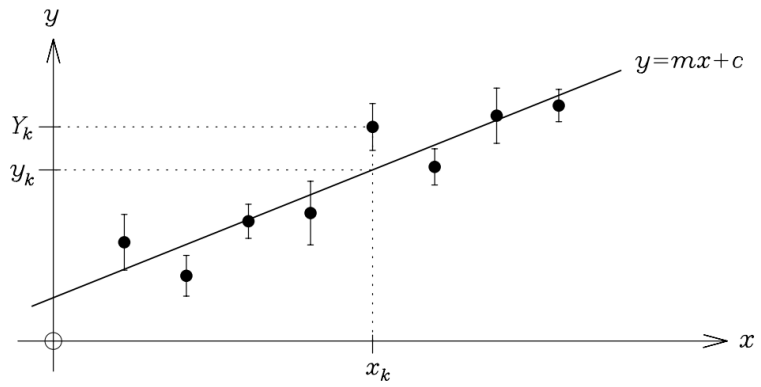

图 3.12:

对于直线模型第 k 个理想数据的值 y_k 为

$$y_k = mx_k + c$$

将 $F_k = y_k, D_k = Y_k$ 代入公式（3.78）$\chi^2 = \sum_{k=1}^{N} \left(\frac{F_k - D_k}{\sigma_k}\right)^2$，可以得到

$$\chi^2 = \sum_{k=1}^{N} \frac{(mx_k + c - Y_k)^2}{\sigma_k^2}, \tag{3.86}$$

根据 $\frac{\partial L}{\partial X_j} = -\frac{1}{2}\frac{\partial \chi^2}{\partial X_j} = -\sum_{k=1}^{N} \frac{(F_k - D_k)}{\sigma_k^2}\frac{\partial F_k}{\partial X_j}$.（噪声满足高斯分布，理想数据呈线性）可以算出 $\boldsymbol{\Delta} L$ 的分量有 $-\frac{1}{2}$ 倍的 χ^2 的偏导数给出

$$\frac{\partial \chi^2}{\partial m} = \sum_{k=1}^{N} \frac{2\,(mx_k + c - Y_k)\,x_k}{\sigma_k^2} \quad \text{和} \quad \frac{\partial \chi^2}{\partial c} = \sum_{k=1}^{N} \frac{2\,(mx_k + c - Y_k)}{\sigma_k^2} \tag{3.87}$$

为了方便，用之前写的 $\boldsymbol{\nabla} L = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{C}$ 的形式来写 $\boldsymbol{\nabla}\chi^2$,

$$\boldsymbol{\nabla}\chi^2 = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} \begin{pmatrix} m \\ c \end{pmatrix} - \begin{pmatrix} p \\ q \end{pmatrix}, \tag{3.88}$$

其中常数 $\alpha, \beta, \gamma, p, q$ 与数据有关

$$\alpha = \sum w_k x_k^2, \quad \beta = \sum w_k, \quad \gamma = \sum w_k x_k, \quad p = \sum w_k x_k Y_k, \quad q = \sum w_k Y_k, \tag{3.89}$$

由 $\boldsymbol{\nabla}\chi^2 = 0$ 通过 χ^2 的最小值得到最佳估计 m_0 c_0，3.88 所示的一对线性联立方程可以直接求解，或通过 3.61 式 $\boldsymbol{X}_0 = -\boldsymbol{H}^{-1}\boldsymbol{C}$ 的矩阵反演，得到

$$m_0 = \frac{\beta p - \gamma q}{\alpha\beta - \gamma^2} \quad \text{和} \quad c_0 = \frac{\alpha q - \gamma p}{\alpha\beta - \gamma^2}. \tag{3.90}$$

根据式 3.83

$$\langle (X_i - X_{\mathrm{O}i})\,(X_j - X_{\mathrm{O}j}) \rangle = -\left[(\boldsymbol{\nabla}\boldsymbol{\nabla} L)^{-1}\right]_{ij} = 2\left[\left(\boldsymbol{\nabla}\boldsymbol{\nabla}\chi^2\right)^{-1}\right]_{ij},$$

这些参数对应的协方差矩阵 $\boldsymbol{\nabla}\boldsymbol{\nabla}\chi^2$ 倒数的两倍给出。通过显式微分，或与 $\boldsymbol{\nabla} L = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{C}$ 和 $\boldsymbol{\nabla}\boldsymbol{\nabla} L = \boldsymbol{H}$ 进行比较，后者与 3.88 式中的 2×2 矩阵相同，因此与之前相同，有

$$\begin{pmatrix} \sigma_{mm}^2 & \sigma_{mc}^2 \\ \sigma_{mc}^2 & \sigma_{cc}^2 \end{pmatrix} = 2\begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}^{-1} = \frac{2}{\alpha\beta - \gamma^2}\begin{pmatrix} \beta & -\gamma \\ -\gamma & \alpha \end{pmatrix}, \tag{3.91}$$

其中，对角线元素的平方根给出了 m 和 c 的推断值的（边缘）误差条，而 γ 项描述了它们是如何关联的。如果理想数据的误差条是未知的，那么可以通过假设它们都是相同的大小来进行分析：$\sigma_k = \sigma$。通过在上述 $\alpha, \beta, \gamma, p, q$ 的定义中设置 $\omega_k = 1/\sigma^2$，发现 m_0 和 c_0 的值与 σ 的

大小无关，但是这样对最优解的可靠性进行估计时，参考的参数 α, β, γ 都是与 σ_k 有关的，所以把 σ_k 当作同一常数时对参数可靠性的估计是不正确的。与第 3.3 节一样，必须考虑到 σ 的不确定性，将其作为一个讨厌的参数进行边缘化积分：

$$\begin{aligned} \text{prob}\left(m, c \mid \{Y_k\}, I\right) &= \int_0^\infty \text{prob}\left(m, c, \sigma \mid \{Y_k\}, I\right) \mathrm{d}\sigma \\ &\propto \int_0^\infty \text{prob}\left(\{Y_k\} \mid m, c, \sigma, I\right) \times \text{prob}(m, c, \sigma \mid I) \mathrm{d}\sigma, \end{aligned} \tag{3.92}$$

在高斯噪声的假设下，似然函数的形式仍然为 $exp(-\chi^2/2)$，

$$\text{prob}\left(\{Y_k\} \mid m, c, \sigma, I\right) \propto \sigma^{-N} \exp\left[-\frac{1}{2\sigma^2} \sum_{k=1}^{N} \left(m x_k + c - Y_k\right)^2\right], \tag{3.93}$$

如果将其与均匀先验相结合，那么 σ 上的积分可以用与第 3.3 节中相同的计算，结果是一个 T 分布，非常类似于 3.45 式：

$$\text{prob}\left(m, c \mid \{Y_k\}, I\right) \propto \left[\sum_{k=1}^{N} \left(m x_k + c - Y_k\right)^2\right]^{-(N-1)/2}. \tag{3.94}$$

与之前一样，m 和 c 的最佳估计，及其相关的协方差矩阵，是由这个边缘后验 pdf 的对数的偏导数给出的。这导致方程 3.90 和 3.91 公式的恢复，误差条 σ 的未知值被由数据得到的估计 S 所取代：

$$S^2 = \frac{1}{N-1} \sum_{k=1}^{N} \left(m_{\mathrm{o}} x_k + c_{\mathrm{o}} - Y_k\right)^2. \tag{3.95}$$

3.6 Error-propagation: changing variables

例如，假设已知 $X = 10 \pm 3$ 和 $Y = 7 \pm 2$，对于 X−Y，或者比率 X/Y，或者它们的平方和 $X^2 + Y^2$ 等等可以做出怎样的推断？就是讨论变量的变化：给定 pdf $prob(X, Y|I)$，相关信息 I 包括: 数据处理需要相应的 pdf $prob(Z|I)$，Z=X−Y 或 Z=X/Y，或视情况而定。从最简单的变量转化开始，即涉及单个变量和它的函数。探究 $Y = f(X)$，pdf$prob(X|I)$ 与 $prob(Y|I)$ 之间的关系，假设对任意点 $X = X^*$ 取一个非常小的区间 δX，X 在范围 $X^* - \delta X$ 到 $X^* + \delta X$ 内的概率由下给出

$$\text{prob}\left(X^* - \frac{\delta X}{2} \leqslant X < X^* + \frac{\delta X}{2} \mid I\right) \approx \text{prob}\left(X = X^* \mid I\right) \delta X, \tag{3.96}$$

现在假设把这个 pdf 看作另一个量 Y 的函数，Y=f(X) 关于 X（单调）。然后，f 将点 $X = X^*$（单值的）映射到 $Y = Y^* = f(X^*)$，将区间 δX 映射到相应的区域 δY，具体情况如图 3.13 所

示。由于 $Y^* \pm \delta Y/2$ 所跨越的 Y 值的范围相当于 $X^* \pm \delta X/2$ 之间的 X 的变化，因此 pdf 面积应该等于 3.96 式所表示的概率。

$$\text{prob}\,(X = X^* \mid I)\,\delta X = \text{prob}\,(Y = Y^* \mid I)\,\delta Y,$$

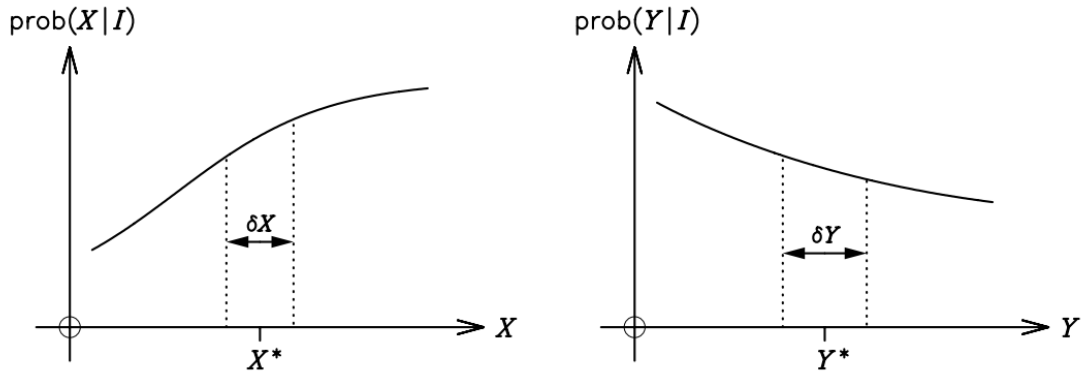

图 3.13: 函数 f 将点 X^* 映射到 $Y^* = f(X^*)$，将小区间 X 映射到对应的区域 Y

由于这对于 x 空间中的任何一点都必须是正确的，所以在无穷小区间的极限下，得到下面的关系

$$\text{prob}(X \mid I) = \text{prob}(Y \mid I) \times \left|\frac{\mathrm{d}Y}{\mathrm{d}X}\right|, \tag{3.97}$$

其中，最右边的项，由 df/dX 的模量表示，称为雅可比矩阵。取模的原因是：即使 X 的正增量与 Y 的负增量之比，也能确保 dy/dx 代表的是长度之比。

作为这个程序的一个具体例子，还看第 2.4 节中的灯塔问题。当时令 $prob(\theta|\alpha,\beta,I) = 1/\pi$ 为方位角 θ 分配了一个统一的 pdf，其中角度 θ 必须位于 $\pm\pi/2$ 弧度之间。还有方位角和沿海岸 x 的位置之间的联系：$\beta tan\theta = x - \alpha$，对这个式子两边分别对 x 求导，可以得到

$$\beta \sec^2\theta \times \frac{\mathrm{d}\theta}{\mathrm{d}x} = 1,$$

利用三角恒等式 $tan^2\theta + 1 = sec^2\theta$，并替换 θ, x 关系式中的 $tan\theta$，雅可比矩阵可以写为

$$\frac{\mathrm{d}\theta}{\mathrm{d}x} = \left(\beta\left[1 + \tan^2\theta\right]\right)^{-1} = \left(\beta\left[1 + \left(\frac{x-\alpha}{\beta}\right)^2\right]\right)^{-1}. \tag{3.98}$$

最后，我们可以利用 3.97 式将关于 θ 的 pdf 转换为关于 x 的等价形式，得到如之前 eqn (2.34)

$prob(x_k|\alpha,\beta,I) = \frac{\beta}{\pi[\beta^2+(x_k-\alpha)^2]}$. 的柯西分布：

$$\text{prob}(x \mid \alpha,\beta,I) = \text{prob}(\theta \mid \alpha,\beta,I) \times \left|\frac{\mathrm{d}\theta}{\mathrm{d}x}\right| = \frac{\beta}{\pi\left[\beta^2+(x-\alpha)^2\right]}. \tag{3.99}$$

3.97 式中的结果可以推广到几个变量的情况，多变量时会很难用类似图 3.13 这样的一个简单的图形表示，但讨论的中心问题是一样的：如果想写与 M 个参数 $\{X_j\}$ 相关的与参数数量相同的 $\{Y_j\}$ 的 pdf，那么必须确保

$$\text{prob}\left(\{X_j\} \mid I\right)\delta X_1\delta X_2\cdots\delta X_M = \text{prob}\left(\{Y_j\} \mid I\right)\delta^M \,\text{Vol}\left(\{Y_j\}\right).$$

其中，$\delta_M Vol(Y_j)$ 是由 X 空间中的小超立方体区域 $\delta X_1\delta X_2\delta X_3 ...\delta X_M$ 绘制出来的 Y 空间中的 M 维体积。大多数关于数学方法的教科书都推导出了这个公式

$$\delta^M \,\text{Vol}\left(\{Y_j\}\right) = \left|\frac{\partial\left(Y_1,Y_2,\cdots,Y_M\right)}{\partial\left(X_1,X_2,\cdots,X_M\right)}\right|\delta X_1\delta X_2\cdots\delta X_M.$$

其中模符号中量是多元雅可比矩阵，它由偏导数 $\partial Y_i/\partial X_j$ 的 M×M 矩阵的行列式给出。因此，3.97 的一般形式可以写成

$$\text{prob}\left(\{X_j\} \mid I\right) = \text{prob}\left(\{Y_j\} \mid I\right) \times \left|\frac{\partial\left(Y_1,Y_2,\cdots,Y_M\right)}{\partial\left(X_1,X_2,\cdots,X_M\right)}\right|. \tag{3.100}$$

3.100 的用法：

考虑定义极坐标 (R，θ) 到在二维笛卡尔网格 (x，y) 上的 pdf 中的等价形式的转换，两组参数之间的函数关系为

$$x = R\cos\theta, \quad y = R\cos\theta,$$

取 X 和 Y 对 R 和 θ 的偏导数，可以很容易地计算所得到的雅可比矩阵的 2×2 矩阵的行列式：

$$\left|\frac{\partial(x,y)}{\partial(R,\theta)}\right| = \begin{vmatrix} \cos\theta & -R\sin\theta \\ \sin\theta & R\cos\theta \end{vmatrix} = R\left[\cos^2\theta+\sin^2\theta\right] = R, \tag{3.101}$$

（多元雅各比矩阵：若在 n 维欧式空间中的一个向量映射成 m 维欧式空间中的另一个向量的对应法则为 f，f 由 m 个实函数组成，即：

$$\begin{cases} y_1 & = f_1(x_1, \ldots, x_n) \\ y_2 & = f_2(x_1, \ldots, x_n) \\ \ldots & \\ y_m & = f_m(x_1, \ldots, x_n) \end{cases}$$

那么雅可比矩阵是一个 $m \times n$ 矩阵:

$$J = \left[\frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_n} \right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

图 3.14:

根据 3.100 式，pdf $prob(R, \theta|I)$ 与 $prob(X, Y|I)$ 是通过下式联系的

$$\text{prob}(R, \theta \mid I) = \text{prob}(x, y \mid I) \times R, \tag{3.102}$$

因此，如果 x 和 y 的 pdf 是各向同性的双变量高斯分布，

$$\text{prob}(x, y \mid I) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(x^2 + y^2)}{2\sigma^2} \right], \tag{3.103}$$

则 R 和 θ 对应的 pdf 将采用以下形式，

$$\text{prob}(R, \theta \mid I) = \frac{R}{2\pi\sigma^2} \exp\left(-\frac{R^2}{2\sigma^2} \right). \tag{3.104}$$

不通过雅可比矩阵分析，还可以直接从一个简单的几何论证中得到 3.102 式

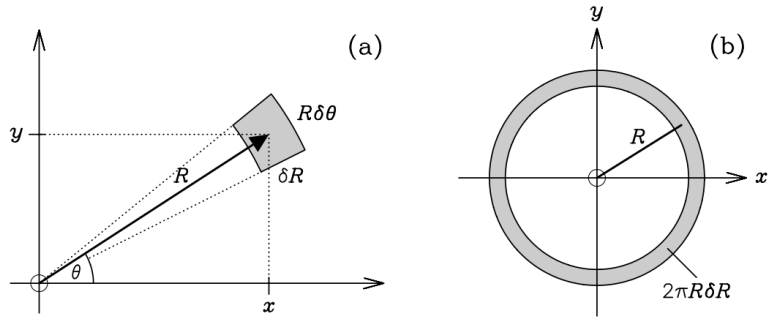

图 3.15: 将变量从笛卡尔坐标转换为极坐标。

极坐标参数落在小范围 $R\pm\delta R/2$ 和 $\theta\pm\theta/2$ 的概率，由：与 (R,θ) 处对应的由 $prob(x,y|I)$ 的大小与图(a)中灰色部分 $R\delta R\delta\theta$ 的乘积给出。通过对 R,θ 的 pdf 的定义，这等价于 $prob(R,\theta|I)\delta R\delta\theta$。因此，

$$prob(R,\theta|I)\delta R\delta\theta = prob(x,y|I)R\delta R\delta\theta.$$

将 prob(x,y|I) 转换为极坐标后，获得半径 $R=\sqrt{x^2+y^2}$ 的 pdf 成为一项简单的任务：只需要在 θ 上边缘化联合 pdf $prob(R,\theta|I)$。对于 3.103 二元高斯分布的情况下，3.104 式很容易对 θ 进行积分

$$\text{prob}(R\mid I)=\int_0^{2\pi}\text{prob}(R,\theta\mid I)\mathrm{d}\theta=\frac{R}{\sigma^2}\exp\left(-\frac{R^2}{2\sigma^2}\right). \tag{3.105}$$

同样，也可以从图片上得出这个结果：因为 3.103 中 pdf 的值只取决于与原点的距离，与角度 θ 无关，所以 R 位于狭窄范围内 δR 的概率是由该半径上高斯 pdf 的大小和阴影环的相应面积决定的，因此，$prob(R|I)\delta R = prob(x,y|I)2\pi R\delta R$ 并且用 R^2 替换 x^2+y^2。

最后一个讨论的多维推广使我们能够推导出在第 3.3 节末尾提到的 χ^2 分布。

2. 已知误差条 $\{\sigma_k\}$ 的预期大小。
$\text{prob}\left(x_k\mid\mu,\sigma\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]$ 是第二章的高斯噪声表达式：高斯分布经常被用作描述与实验数据相关的噪声（或缺陷）的理论模型，该式表示第 k 个数据具有值 xk 的概率。似然函数可以表示为

$$\text{prob}(\boldsymbol{D}\mid\boldsymbol{X},I)\propto\exp\left(-\frac{\chi^2}{2}\right), \tag{3.77}$$

其中，χ^2 为归一化残差的平方和（归一化后的残差叫做标准残差（$R_k=(F_k-D_k)/\sigma_k$），满足的是 N（0，1）的标准正态分布，在进行线性回归时，可以更简单的判断数据中的可疑值）：

$$\chi^2=\sum_{k=1}^{N}\left(\frac{F_k-D_k}{\sigma_k}\right)^2. \tag{3.78}$$

由于取用均匀先验，后验与似然直接相关，根据（3.71）和（3.72），后验对数可以写为

$$L=\log_e[\text{prob}(\boldsymbol{X}\mid\boldsymbol{D},I)]=\text{constant}-\frac{\chi^2}{2}. \tag{3.79}$$

图 3.16:

3.77 和 3.78 的似然函数是一个 n 维的、各向同性的高斯函数：

$$\text{prob}(\boldsymbol{D}\mid\boldsymbol{X},I)\propto\exp\left[-\frac{r_1^2+r_2^2+\cdots+r_N^2}{2}\right], \tag{3.106}$$

这里 $r_k=(F_k-D_k)/\sigma_k$。一如刚才的讨论，pdf 的值只取决于到原点的距离，这个半径 $R=\sqrt{\sum r_k^2}$，这也是 χ^2 的平方根。R 位于一个狭窄范围 δR 的概率，$prob(R|X,I)\delta R$，将等于在该半径上的似然函数的大小与相关球壳的超体积的乘积，由于体积与 R^{N-1} 成正比，

$$\text{prob}(R\mid\boldsymbol{X},I)\propto R^{N-1}\exp\left(-R^2/2\right). \tag{3.107}$$

这样看出来边缘分布只是 3.105 式的一个多维扩展。为了将其转换为 χ^2 的 pdf，只需要根据 3.97 式 $\text{prob}(X \mid I) = \text{prob}(Y \mid I) \times \left|\frac{\mathrm{d}Y}{\mathrm{d}X}\right|$ 进行一对一的转换，利用函数关系 $\chi^2 = R^2$，得到

$$\text{prob}\left(\chi^2 \mid \boldsymbol{X}, I\right) \propto \left(\chi^2\right)^{N/2-1} \exp\left(-\chi^2/2\right). \tag{3.108}$$

这被称为具有 N 个自由度的 χ^2 分布。

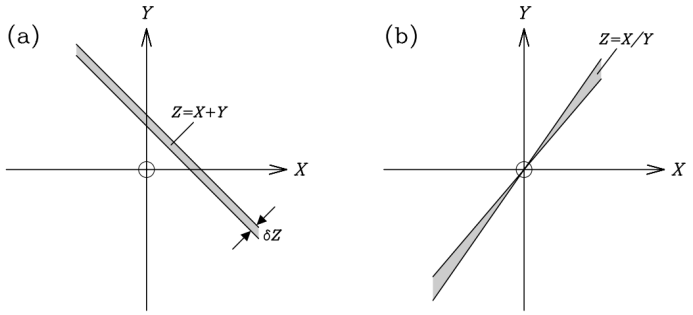

图 3.17:

现在已经看到了变量之间传递所需的基本步骤：它要么涉及 3.100 意义上的转换，要么是像 3.105 这样的积分，或者是这两种的结合。对于想要估计 Z=X+Y 或比值 Z=X/Y 的常见问题，只需要沿着图 3.15 中的阴影条对联合 pdf$prob(X, Y|I)$ 进行积分。如果直观上看的不明显，可以通过使用边缘化和乘积规则来证明：

$$prob(Z, X, Y|I) = prob(Z|X, Y, I) \times prob(X, Y|I),$$

$$\begin{aligned}
\text{prob}(Z \mid I) &= \iint \text{prob}(Z, X, Y \mid I) dX dY \\
&= \iint \text{prob}(Z \mid X, Y, I)\, \text{prob}(X, Y \mid I) \mathrm{d}X\ \mathrm{d}Y \\
&= \iint \delta(Z - f(X, Y))\, \text{prob}(X, Y \mid I) \mathrm{d}X\ \mathrm{d}Y,
\end{aligned} \tag{3.109}$$

其中，第二行中的 δ 函数等于零，除非 Z=f(X，Y)，并且表示当 X 和 Y 已知时，Z 是明确确定的 (由函数 f)。对于 Z=X+Y，图中对角阴影线的积分可以写为

$$\text{prob}(Z \mid I) = \iint \delta(Z - (X + Y))\, \text{prob}(X, Y \mid I) \mathrm{d}X\ \mathrm{d}Y, \tag{3.110}$$

如果信息 I 只告诉了 $X = x_0 \pm \sigma_x$ 和 $Y = y_0 \pm \sigma_y$，那么可以合理地假设这些参数是不相关的，

prob(X，Y|I) 可以写为 X 和 Y 的各自的 pdf 的乘积，因此 3.110 写作

$$\text{prob}(Z \mid I) = \int \mathrm{d}X \,\text{prob}(X \mid I) \int \text{prob}(Y \mid I)\delta(Z - X - Y)\mathrm{d}Y.$$

根据 δ 函数的性质，积分化简为

$$\text{prob}(Z \mid I) = \int \text{prob}(X \mid I)\,\text{prob}(Y = Z - X \mid I)\mathrm{d}X \tag{3.111}$$

根据这些信息的性质，书里建议应该为这两个参数分配高斯 pdfs，最大值分别为 $x_0.y_0$，宽度为 $\sigma_x\sigma_y$，将它们代入 3.111，

$$\text{prob}(Z \mid I) = \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{+\infty} \exp\left[-\frac{(X - x_0)^2}{2\sigma_x^2}\right] \exp\left[-\frac{(Z - X - y_0)^2}{2\sigma_y^2}\right] \mathrm{d}X. \tag{3.112}$$

经过化简，

$$\text{prob}(Z \mid I) = \frac{1}{\sigma_z\sqrt{2\pi}} \exp\left[-\frac{(Z - z_{\text{o}})^2}{2\sigma_z^2}\right]. \tag{3.113}$$

这里

$$z_0 = x_0 + y_0 \quad \text{和} \quad \sigma_z^2 = \sigma_x^2 + \sigma_y^2,$$

因此，和的 pdf 是也是由高斯分布给出的，最大值在 z_0 处，宽度为 σ_z，事实上差（Z=X-Y）的 pdf 也是由高斯分布给出的，z_0 由 $x_0 - y_0$ 给出。