

粒子物理与核物理实验中的 数据分析

王喆

清华大学

第四章：统计检验

本讲要点

- ❑ 假设，检验统计量，显著水平，功效
- ❑ 纽曼-皮尔森引理
- ❑ 如何构造一个检验统计量
- ❑ Fisher甄别函数与神经网络
- ❑ 检验拟合优度， P -值定义与应用
- ❑ 信号观测的显著程度
- ❑ 皮尔逊的 χ^2 检验

概率与统计

统计的含义可以通过比较概率理论来理解

概率	统计(参量测定与假设检验)
从理论到数据	从数据到理论
通过计算某些可观测量(例如, 平均值, 分布等)来给出预期的实验分布。 例如: 若宇称守恒, 对一特定衰变分布有什么影响?	进行所谓的假设检验, 比较理论预期的参量值或分布。从观察的实验数据中给出所研究参数的观测值和误差, 并且在某一置信水平上检验理论的正确与否。 例如: 观测到一特定衰变分布, 是否可以断定宇称守恒?

统计分析的目标

假设检验



检验数据是否与某一特定理论相符(注意, 该理论可包含一些自由参数)。



相符的程度由显著水平来表示。

参数拟合



利用数据确定自由参数的大小。

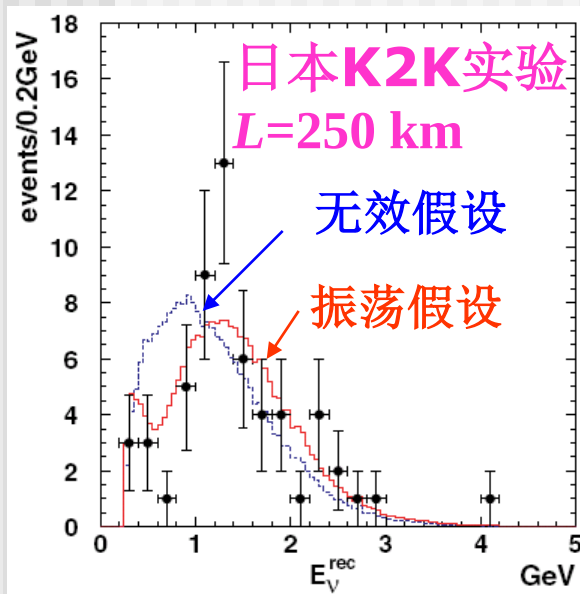


参数的准确程度由对应的误差大小来表示。

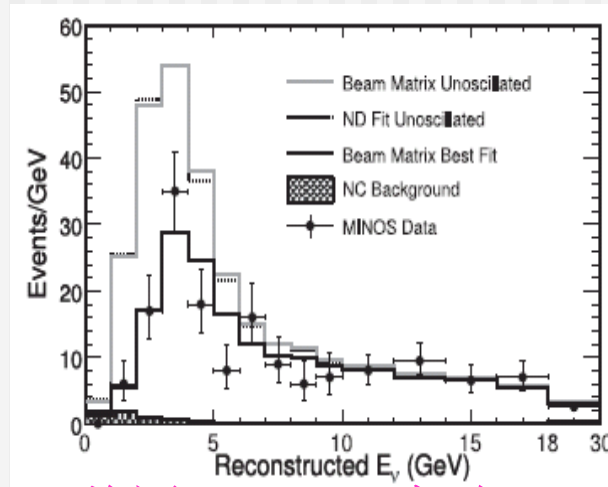
中微子振荡假设检验

利用加速器把中微子射往远处的探测器，观察有多少中微子发生了形态上的改变，即所谓的加速器中微子振荡实验

$$P(\nu_{\mu} \rightarrow \nu_{\mu}) = 1 - \sin^2(2\theta) \sin^2(1.27 \Delta m^2 \frac{L}{E_{\nu_{\mu}}})$$

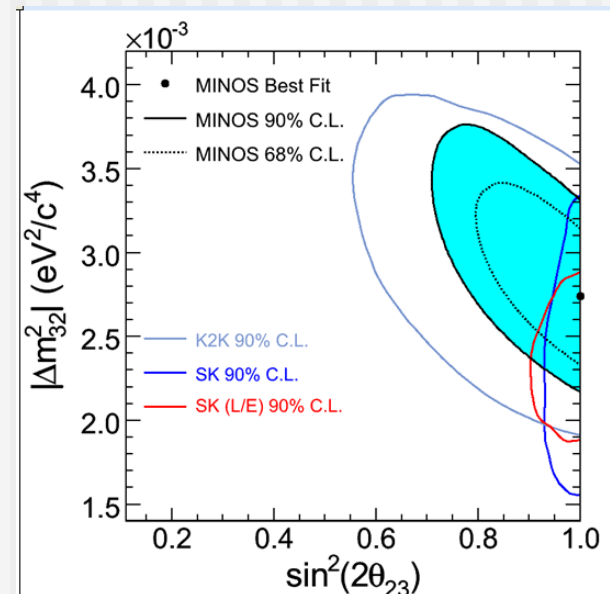


Phys.Rev.D74,072003(2006)
振荡假设符合概率: 37%
无效假设符合概率: 0.07%



美国MINOS实验
 $L=700$ km

Phys.Rev.Lett.97,191801(2006)



美国实验证实了日本实验而且实验精度更高。

假设检验

假如测量结果为 $\vec{x} = (x_1, x_2, \dots, x_n)$, 例如: 正负电子对撞后所产生的事例中, 对于每个事例, 有下列测量量

x_1 = 产生的带电粒子数; x_2 = 粒子的平均横动量; x_3 = 产生的“喷注”数目;

这里 $\vec{x}$ 服从在 n - 维空间的某些与产生事例类型有关的联合概率密度函数, 例如: 正负电子对撞, 原子核与原子核碰撞等等。那么这些联合的概率密度函数 $f(\vec{x})$ 取决于采取何种假设。

$f(\vec{x}|H_0), f(\vec{x}|H_1),$ 等等

简单假设: $f(\vec{x})$ 无未定参数

复杂假设: $f(\vec{x}; \alpha)$ 含未定参数 α

通常情况下很难处理多维的 $\vec{x}$ 问题, 因此, 常常构造低维的统计检验, 在不失去甄别各种假设能力的条件下, 使得 $t(\vec{x})$ 成为精简后的数据样本。

那么此时的统计量 t 具有概率密度函数 $g(t|H_0), g(t|H_1), \dots$

拒绝域、第一与第二类误差

考虑统计检验量 t 服从 $g(t | H_0), g(t | H_1), \dots$
定义拒绝域, 使得 H_0 假设为真时, t 不大可能发生

例如, 在上述情况下, $t \geq t_{cut}$

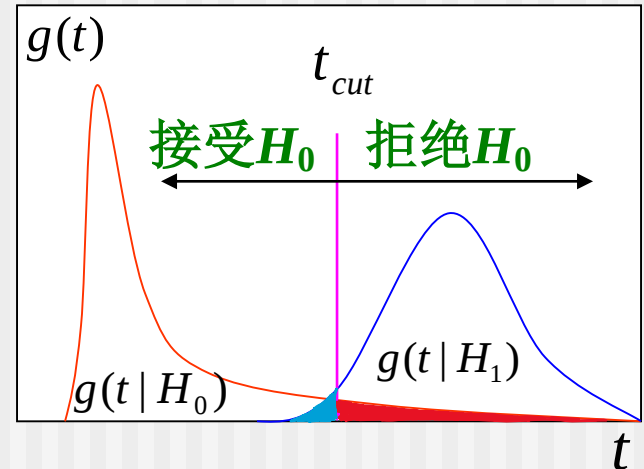
如果观测量 t_{obs} 在拒绝域时, 拒绝 H_0 ,
否则接受 H_0 。

假若 H_0 为真, 但被拒绝的可能性构成第一类错误

$$\alpha = \int_{t_{cut}}^{\infty} g(t | H_0) dt \quad (\text{显著水平})$$

假若接受 H_0 , 但实际情况却是 H_1 为真的可能性构成第二类错误

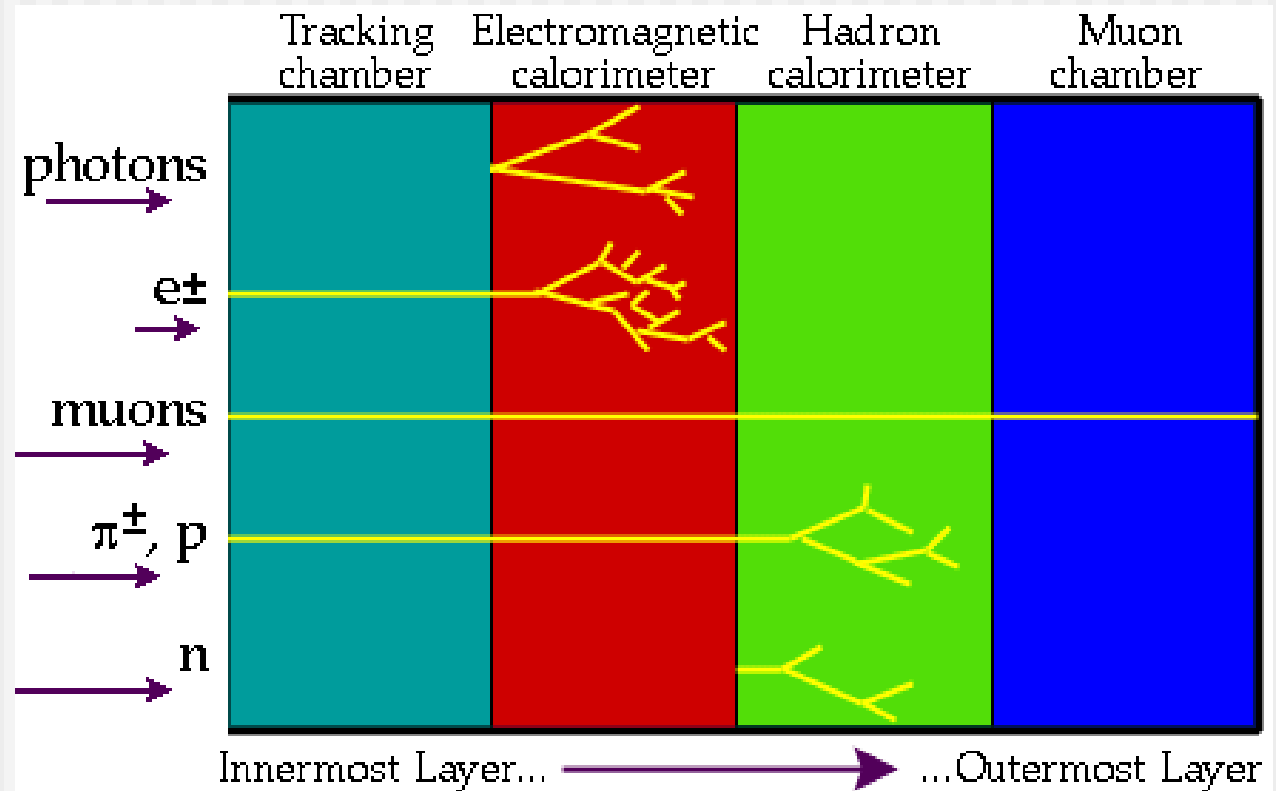
$$\beta = \int_{-\infty}^{t_{cut}} g(t | H_1) dt \quad (1 - \beta = \text{功效})$$



例子：粒子鉴别

探测器中的稳定粒子($\tau \geq 10^{-9}\text{s}$)

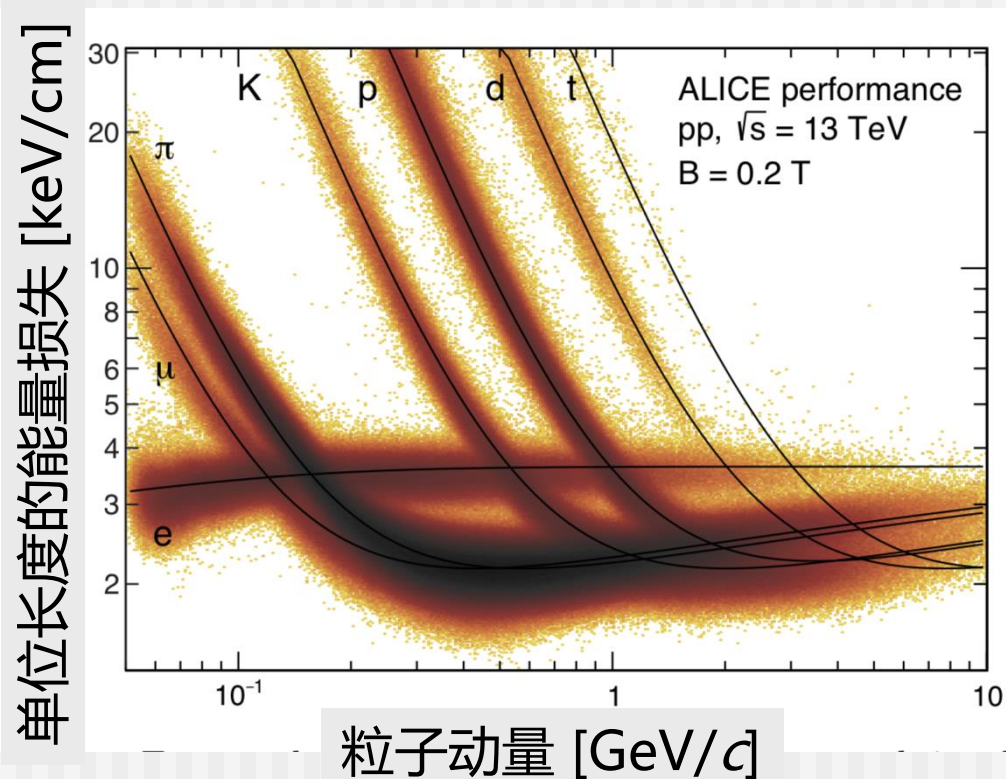
$\gamma, e^{\pm}, \mu^{\pm},$
 $\pi^{\pm}, K^{\pm}, K_L$
 p, n, Λ



例子：粒子鉴别

粒子在物质中的任何相互作用原则上都可用来作粒子鉴别，区别在于效果好坏。

带电粒子在物质中电离造成能量损失，可用来鉴别带电粒子：



这种方法的局限是什么？

例子:选择不同粒子

一束包含 K/π 粒子的束流穿过**2**厘米厚的闪烁体, 根据电离能损的大小可以用来进行粒子鉴别。构造能量沉积测量量 t , 并假设只有**两种**可能

$H_0 = \pi$ (信号)

$H_1 = K$ (本底)

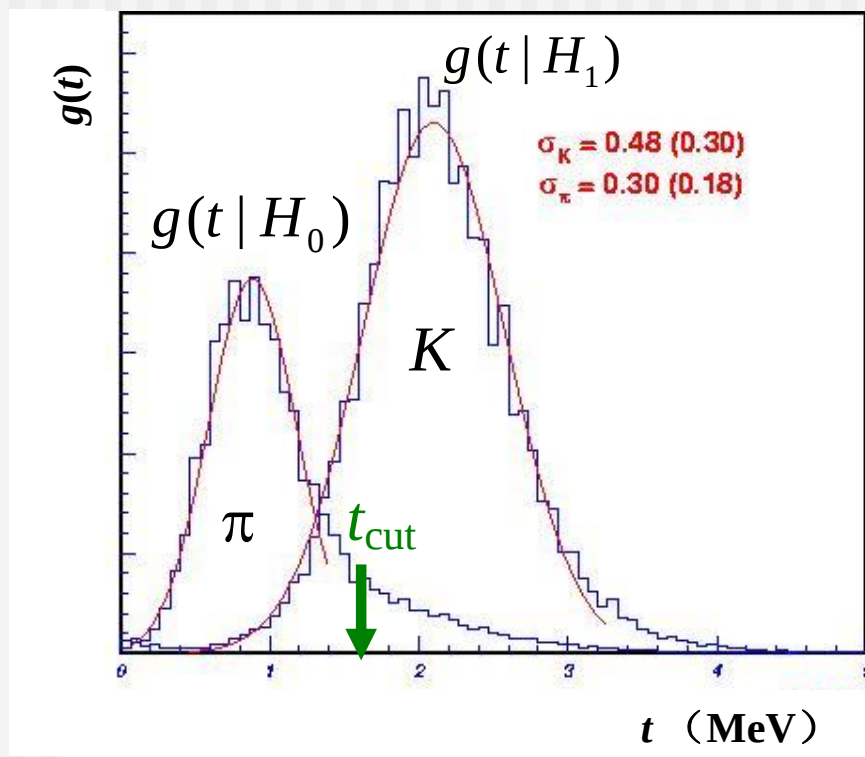
通过要求 $t < t_{cut}$ 来选择 π 粒子,
选择效率为

$$\varepsilon_{\pi} = \int_{-\infty}^{t_{cut}} g(t|\pi) dt = 1 - \alpha$$

$$\varepsilon_K = \int_{-\infty}^{t_{cut}} g(t|K) dt = \beta$$

松选择: 效率很高, 但 K 本底高;
严选择: 信号样本纯, 但效率低。

t 分布估计为: (π 的份额 a_{π}) $f(t; a_{\pi}) = a_{\pi} g(t|\pi) + (1 - a_{\pi}) g(t|K)$



选择到的粒子数及误差

选择到的总事例数：

$$\begin{aligned} N_{acc} &= \epsilon_{\pi} N_{\pi} + \epsilon_K N_K \\ &= \epsilon_{\pi} N_{\pi} + \epsilon_K (N_{tot} - N_{\pi}) \end{aligned}$$

选择到的信号事例数：

$$N_{\pi} = \frac{N_{acc} - \epsilon_K N_{tot}}{\epsilon_{\pi} - \epsilon_K}$$

可以根据信号事例数的数目或统计误差来确定拒绝域

粒子鉴别的概率问题

对于一个具有测量值 t 的粒子，如何估计是 K 还是 π 的概率？

贝叶斯定理



$$h(K | t) = \frac{a_K g(t | K)}{a_K g(t | K) + a_\pi g(t | \pi)}$$
$$h(\pi | t) = \frac{a_\pi g(t | \pi)}{a_K g(t | K) + a_\pi g(t | \pi)}$$

对于贝叶斯论者：上式为粒子是 K 或 π 的可信程度

对于频率论者：给定 t 条件下，粒子是 K 或 π 的比率

通常情况下，需要给出选择样本的纯度



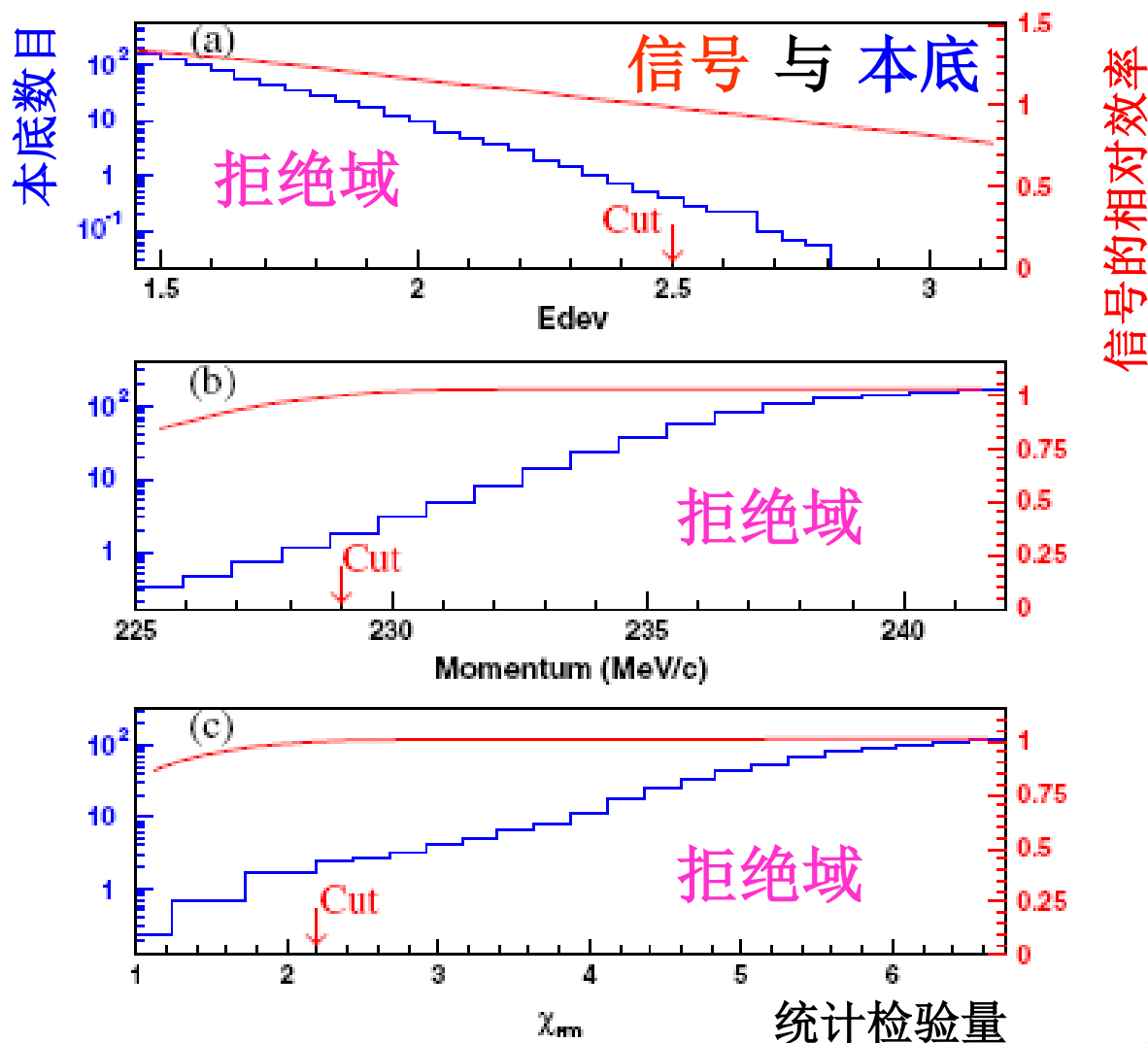
两种解释
均有道理

$$p_\pi = \frac{N_\pi(t < t_{cut})}{N_{all}(t < t_{cut})} = \frac{\int_{-\infty}^{t_{cut}} a_\pi g(t | \pi) dt}{\int_{-\infty}^{t_{cut}} [a_\pi g(t | \pi) + (1 - a_\pi) g(t | K)] dt} = \frac{\int_{-\infty}^{t_{cut}} h(\pi | t) f(t) dt}{\int_{-\infty}^{t_{cut}} f(t) dt}$$

= π 粒子在区间 $(-\infty, t_{cut}]$ 的概率

可以根据信号的纯度来确定拒绝域

实验中拒绝域的选取



cut的确定
较难!

较易!

较易!

Phys. Rev. D77, 052003 (2008)

2026-07-22

纽曼-皮尔森引理与拒绝域

考虑一个多维检验统计量 $t=(t_1,\dots,t_m)$ ，有信号假设 H_0 与本底假设 H_1 。

问题：给定显著水平如何选择一个最佳的拒绝域或者 **cut**?

纽曼-皮尔森引理：对于一个给定的显著水平（或选择效率），在 t 空间给出最高功效（或最高信号纯度）的接收域满足

$$\frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} > c; \quad c - \text{由期望的接收效率而决定的参数}$$

上面定义式左侧为一维检验统计量，简单假设 H_0 与 H_1 的似然比

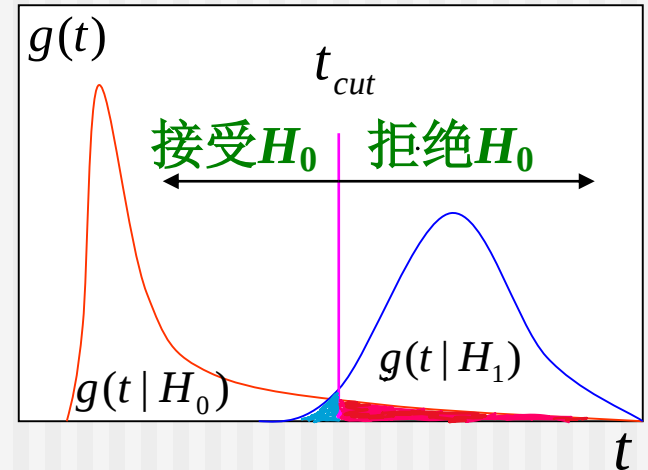
$$r = \frac{g(\vec{t} | H_0)}{g(\vec{t} | H_1)} \quad \longrightarrow \quad r > c$$

纽曼-皮尔森引理 说明

问题：给定 α ，如何选定区域 ω_c ，使 $1 - \beta$ 最大？

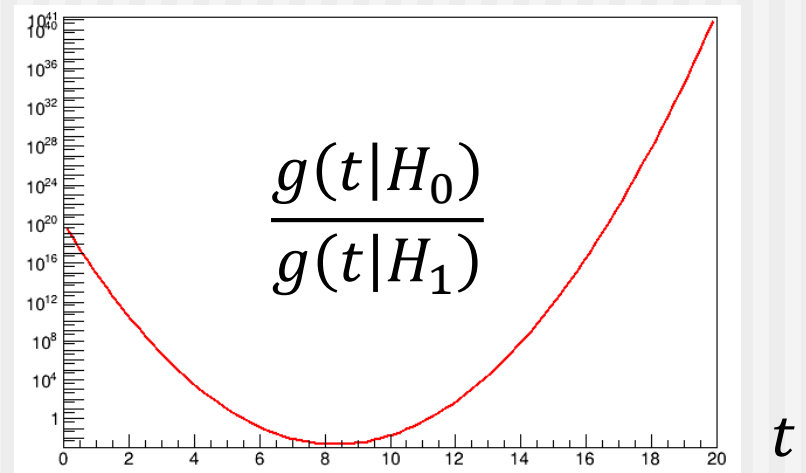
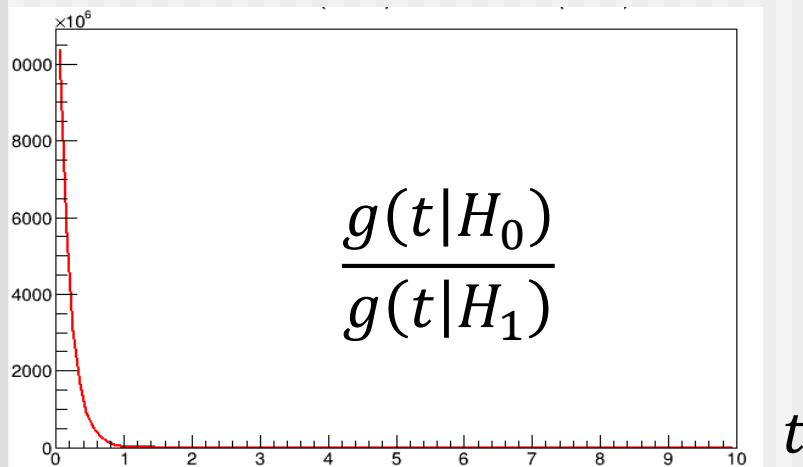
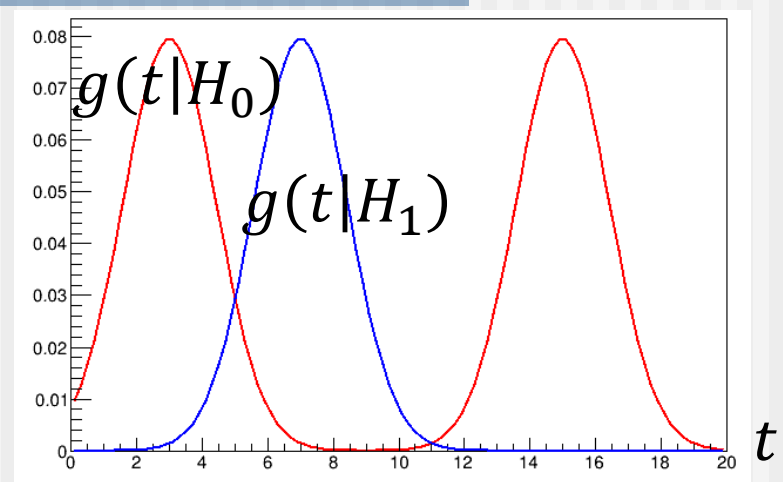
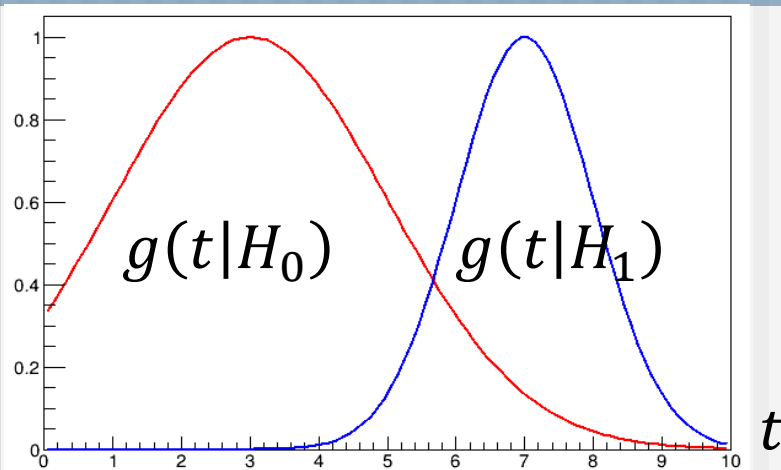
$$\alpha = \int_{\omega_c} g(t|H_0) dt$$

$$\begin{aligned} 1 - \beta &= \int_{\omega_c} g(t|H_1) dt \\ &= \int_{\omega_c} \frac{g(t|H_1)}{g(t|H_0)} g(t|H_0) dt \\ &= E_{\omega_c} \left(\frac{g(t|H_1)}{g(t|H_0)} \mid H_0 \right) \end{aligned}$$



所以，我们应该尽量选取 $\frac{g(t|H_1)}{g(t|H_0)}$ 最大的那一组做拒绝区域
或者说，尽量选取 $\frac{g(t|H_0)}{g(t|H_1)}$ 最大的那一组做选择域

纽曼-皮尔森引理一维应用



$g(t|H_0)$, $g(t|H_1)$ 分别为两个简单的高斯分布

$g(t|H_0)$ 形态变复杂的情况

利用直方图近似得到似然比

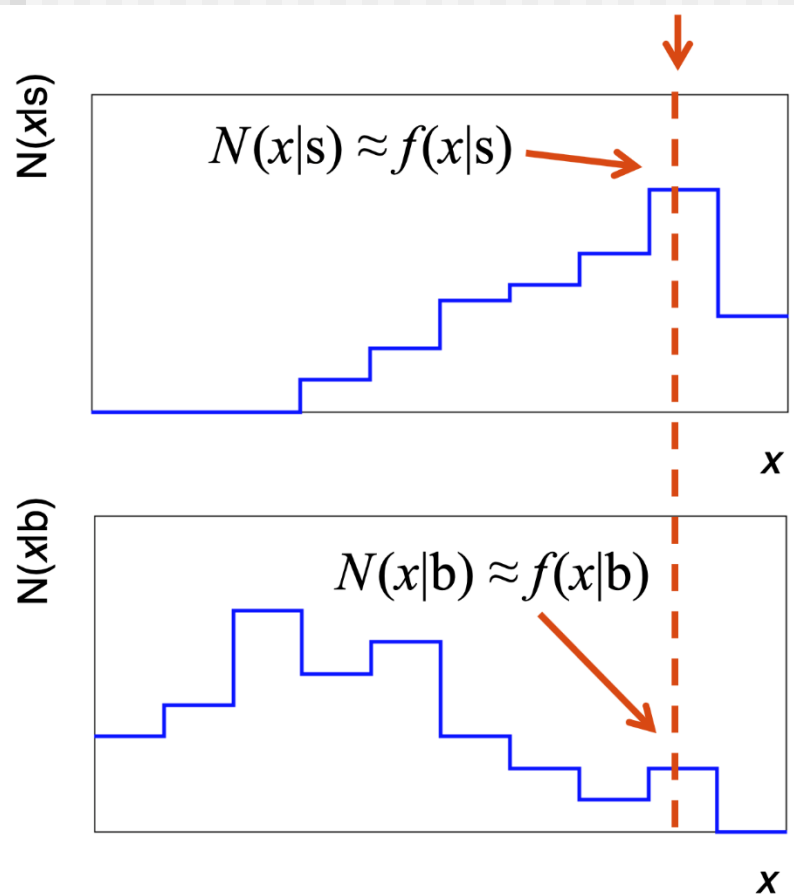
希望得到 $t(x) = \frac{f(x|s)}{f(x|b)}$

可以产生**MC**数据并得到信号和本底的直方图。

利用归一化的直方图的值近似得到似然比

$$t(x) = \frac{N(x|s)}{N(x|b)}$$

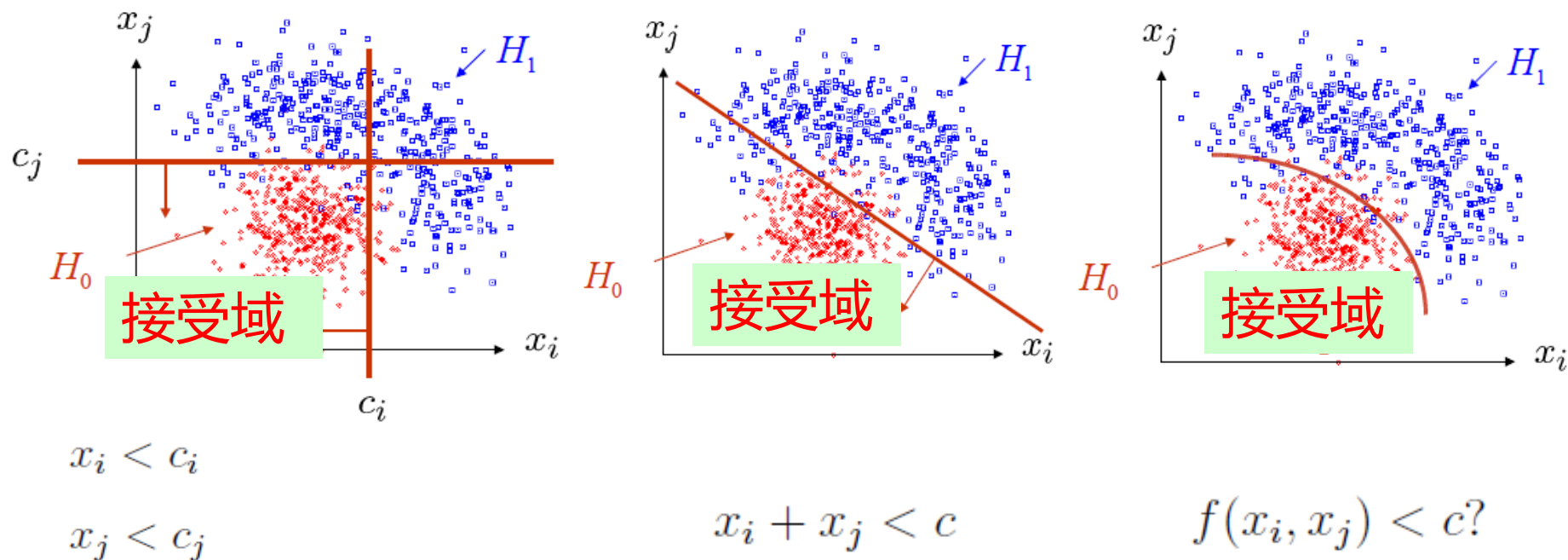
对于单变量问题，这种方法很有效。



二维事例筛选

假设实验数据包含两种不同事例，对应于假设 H_0 和 H_1 ，我们希望选择出 H_1 类的事例（位于临界域）。

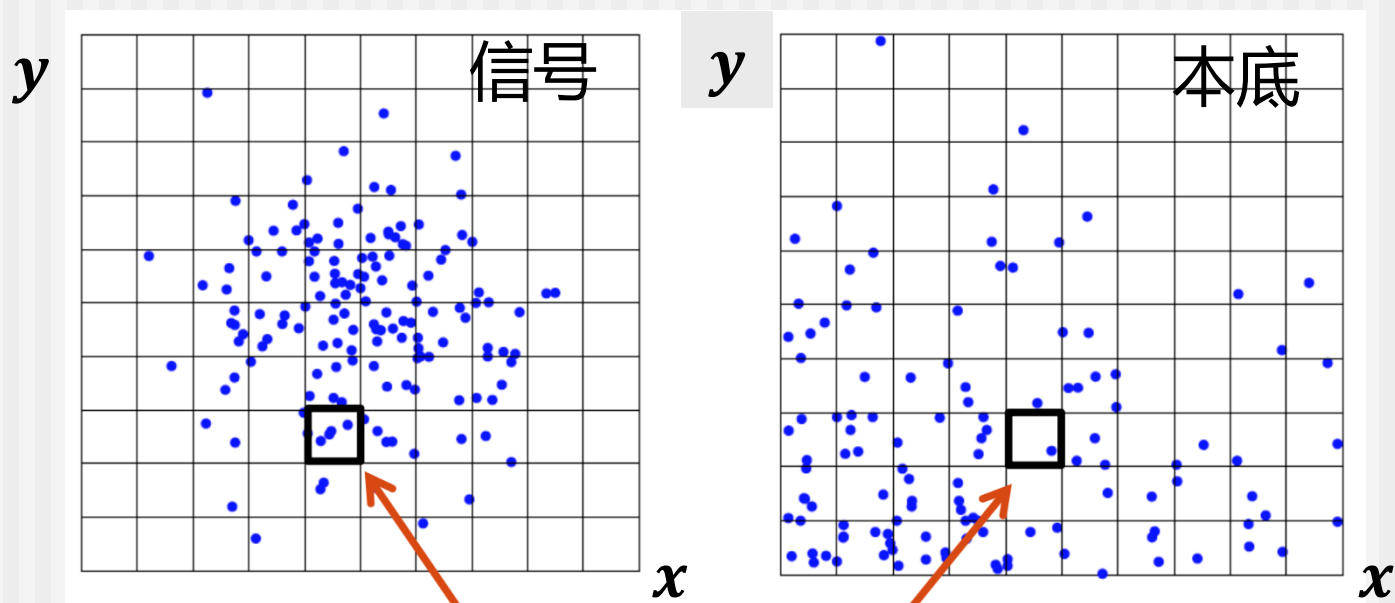
每个事例对应于 $\vec{x}$ 空间的一个点。应当如何选择判别边界？



如何给出“最优化”的选择？

利用二维直方图近似得到似然比

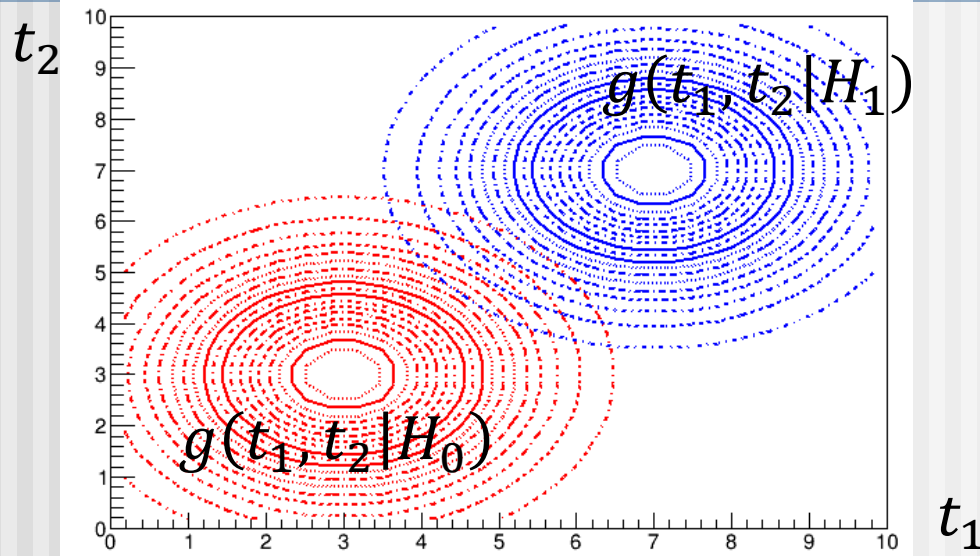
假设问题中有两个变量(x, y)，可以使用二维直方图



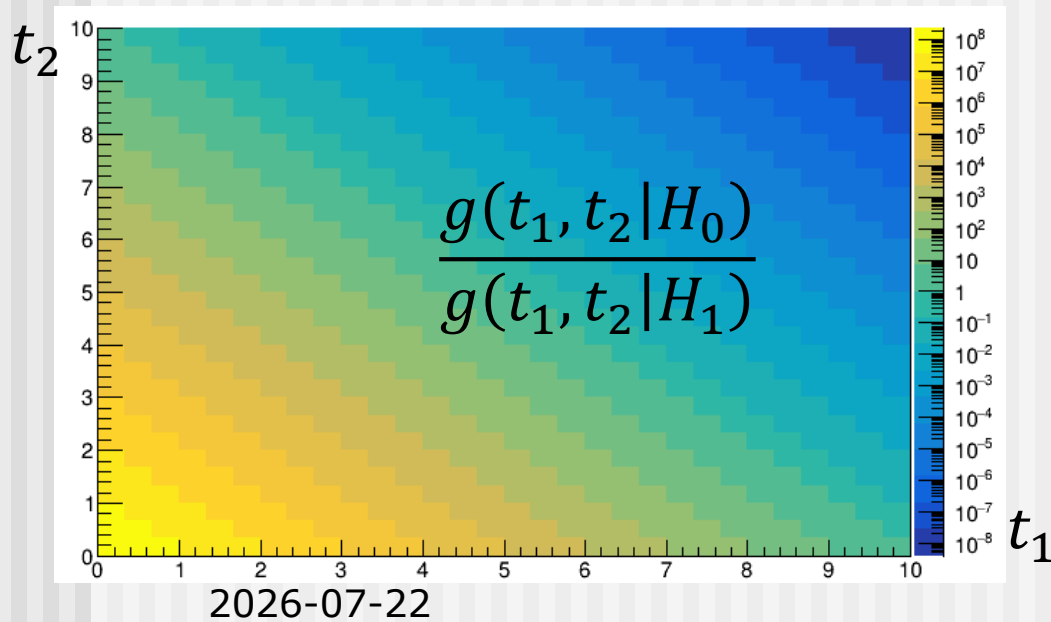
利用对应网格中的 $N(x, y|s)$ 和 $N(x, y|b)$ 近似二维概率密度。
如果希望每个变量分 M 个区间，对于 n 维问题，需要 M^n 个网格；
很难产生足够的训练数据。

➔ 对于 n 较大的问题，直方图方法往往不直观

两维计算的图示



$g(t_1, t_2|H_0)$, $g(t_1, t_2|H_1)$
分别为两个二维高斯分布



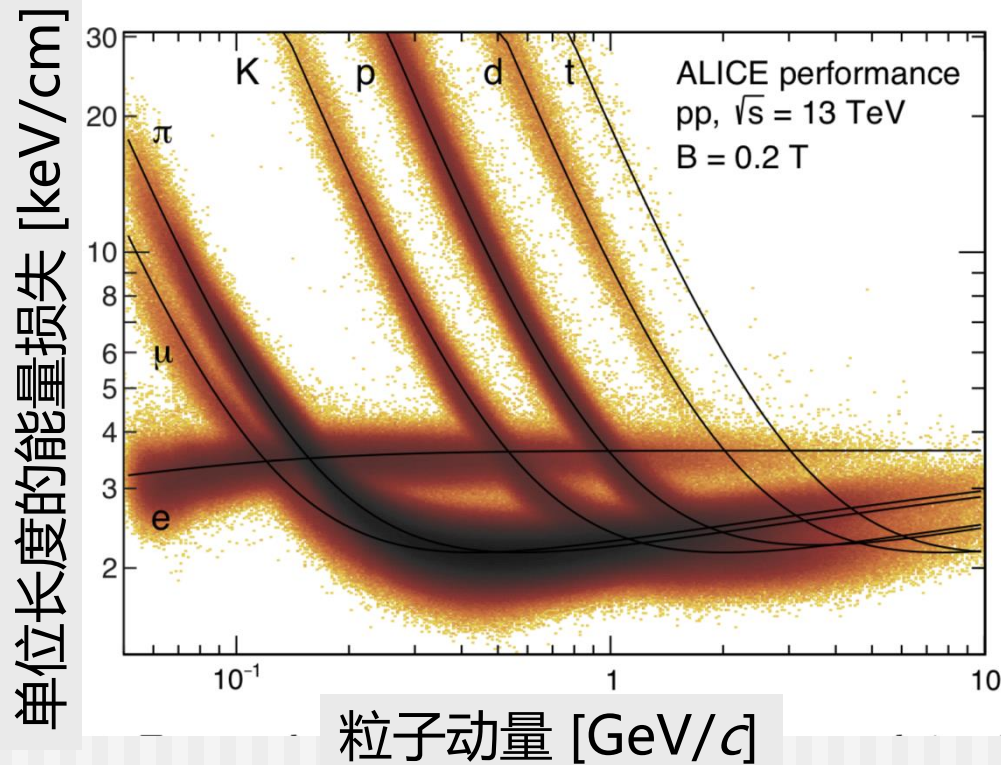
$$\frac{g(t_1, t_2|H_0)}{g(t_1, t_2|H_1)}$$

在确定效率的情况下，可
以选择似然比较高的区域

粒子鉴别的例子——多维

粒子在物质中的任何相互作用原则上都可用来作粒子鉴别，区别在于效果好坏。

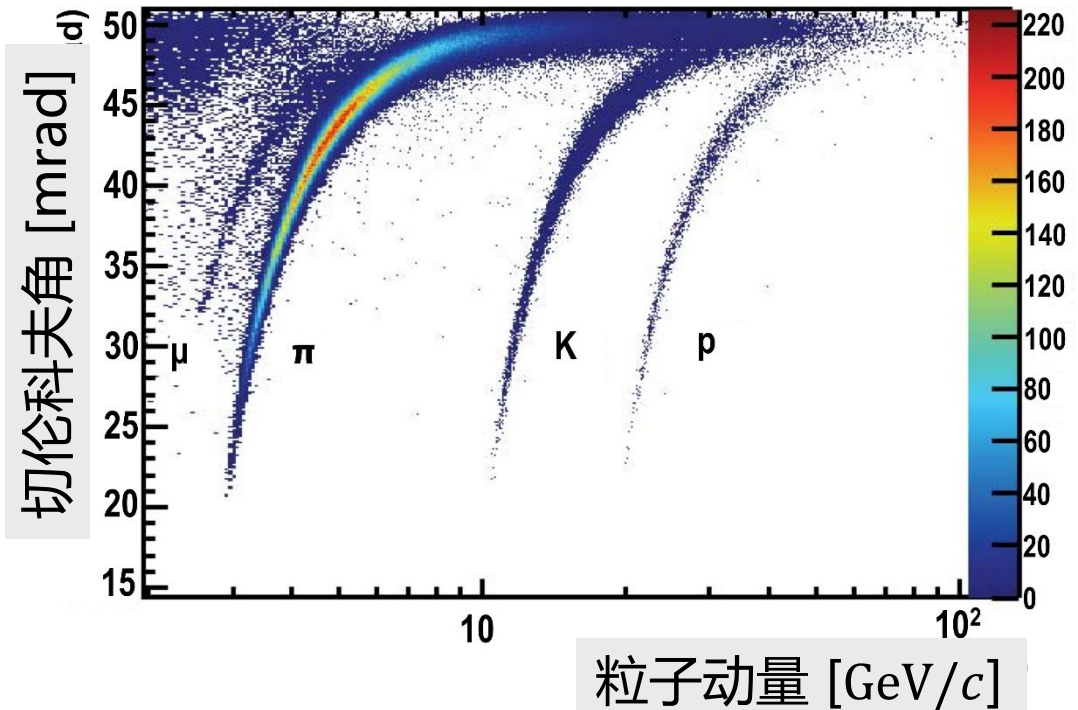
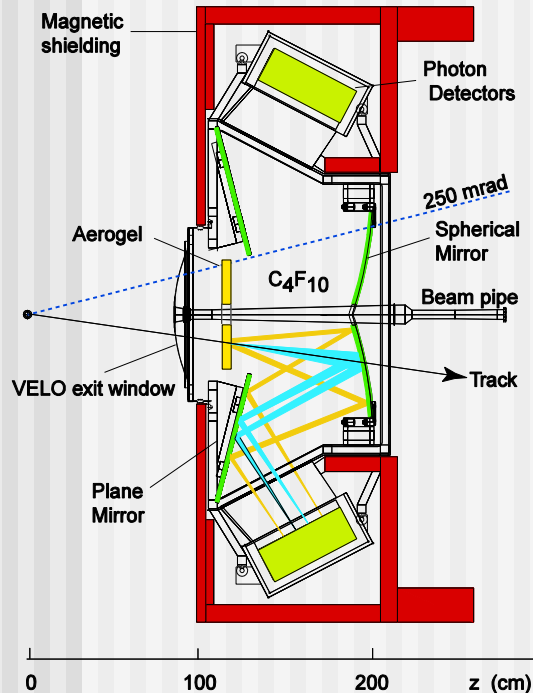
带电粒子在物质中电离造成能量损失，可用来鉴别带电粒子：



粒子鉴别的例子——多维（续）

粒子在物质中的任何相互作用原则上都可用来作粒子鉴别，区别在于效果好坏。

带电粒子在介质中发射切伦科夫光：



如何构造一个检验统计量

在只考虑两种可能性的情况下，对于每个事例，测量

$$\vec{x} = (x_1, \dots, x_n)$$

根据纽曼-皮尔森引理，为了选择事例，可选择拒绝域

$$t(\vec{x}) = \frac{f(\vec{x} | H_0)}{f(\vec{x} | H_1)}$$

问题：如何知道这两个不同假设下的概率密度函数？

实际应用中，可以利用蒙特卡罗方法模拟物理过程与探测器响应，通过产生大量的样本，可以近似地得到上述概率密度函数的表达方式。

分别产生信号
与本底事例，并
经过探测器模拟



对每个事例，得到
测量量 $\vec{x}$ ，并填入 n 维
直方图。如果 M 为每
个分量的区间数，则总
单元数为 M^n 。



$$\frac{f(\vec{x} | H_0)}{f(\vec{x} | H_1)}$$

但是如果 n
太大时，实
际运用会很
困难。

构建线性检验统计量

当维数 >2 时，假设每一维度研究均需要分 M 个区间，那么，对于 n -维问题，需要 M^n 个格子方能将概率密度函数近似确定下来。为了简化处理此类问题，可以采用拟设的方法给出包含少量参数的检验统计量形式，通过确定参数(例如采用蒙特卡罗方法)，最大限度地区分 H_0 与 H_1 。

拟设：
$$t(\vec{x}) = \sum_{i=1}^n a_i x_i = \vec{a}^T \vec{x} \quad (\text{即把测量量做线性叠加})$$

给定一个 $\vec{a}$ ，可以得到相应的概率密度函数 $g(t | H_0), g(t | H_1)$

通过选择 $\vec{a}$ 来到达最大地区分 $g(t|H_0)$ 与 $g(t|H_1)$ 的目的。

对不同假设下的均值与方差要求

对已有的测量量，我们可以计算对应的期待值与协方差

$$(\mu_k)_i = \int x_i f(\vec{x} | H_k) d\vec{x} \quad k = 0, 1 \text{ (假设)}$$

$$(V_k)_{ij} = \int (x - \mu_k)_i (x - \mu_k)_j f(\vec{x} | H_k) d\vec{x} \quad i, j = 1, \dots, n \text{ (}\vec{x}\text{分量)}$$

类似地，我们还可以导出计算 $t(\vec{x})$ 平均值与方差的公式

$$\tau_k = \int t(\vec{x}) f(\vec{x} | H_k) d\vec{x} = \vec{a}^T \vec{\mu}_k$$

$$\sum_k^2 = \int (t(\vec{x}) - \tau_k)^2 f(\vec{x} | H_k) d\vec{x} = \vec{a}^T V_k \vec{a}$$

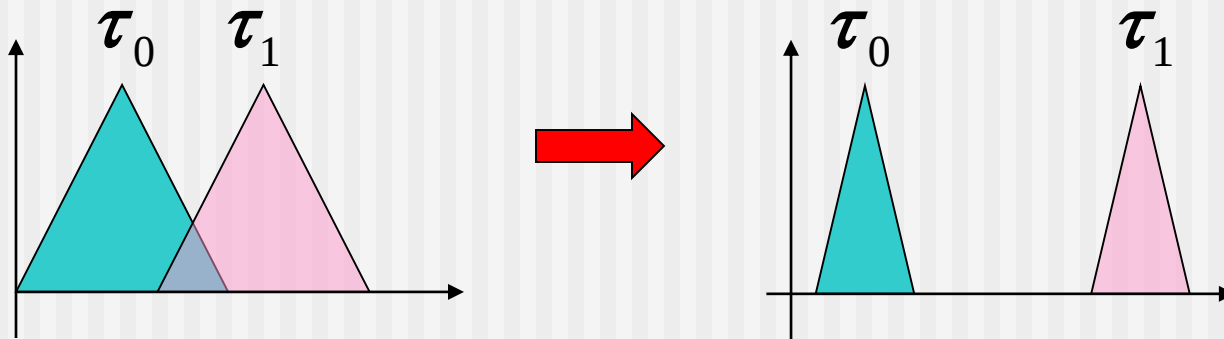
通过优化参数最大区分两种假设

利用算出的 $t(\vec{x})$ 平均值与方差的公式

$$\tau_k = \int t(\vec{x}) f(\vec{x} | H_k) d\vec{x} = \vec{a}^T \vec{\mu}_k$$

$$\sum_k^2 = \int (t(\vec{x}) - \tau_k)^2 f(\vec{x} | H_k) d\vec{x} = \vec{a}^T V_k \vec{a}$$

要求大的 $|\tau_0 - \tau_1|$ 与小的 $\sum_0^2, \sum_1^2$
使得 pdfs 分布集中在均值附近。



Fisher判别函数的定义

Fisher定义了一个判别法

$$J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\sum_0^2 + \sum_1^2} \begin{aligned} &\rightarrow = \sum_{i,j=1}^n a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j = \sum_{i,j=1}^n a_i a_j B_{ij} = \vec{a}^T B \vec{a} \\ &\rightarrow = \sum_{i,j=1}^n a_i a_j (V_0 + V_1)_{ij} = \vec{a}^T W \vec{a} \end{aligned}$$

则

$$J(\vec{a}) = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}}$$

$$\text{令 } \frac{\partial J}{\partial a_i} = 0 \quad \rightarrow \quad \vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$$

因此定义了可求极值的Fisher线性判别函数 $t(x)$ 。

求Fisher甄别函数的最大值

若将 $t(\vec{x})$ 写成

$$t(\vec{x}) = a_0 + c \sum_{i=1}^n a_i x_i$$

用任意整体标度 c 和偏置 a_0 去固定 τ_0 , τ_1 到0和1

求 $J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\sum_0^2 + \sum_1^2}$ 的最大值, 意味着要将下式最小化

$$\sum_0^2 + \sum_1^2 = E_0[(t - \tau_0)^2] + E_1[(t - \tau_1)^2]$$

与假设对应的期待值

求 Fisher 函数 $J(\vec{a})$

最大值就可以利用我们稍后要学习的求极值方法。

高斯分布下Fisher甄别量特点

假如 $f(\vec{x}|H_k)$ 是多量高斯分布，具有平均值

$\vec{\mu}_0$ 为假设 H_0 的均值

$\vec{\mu}_1$ 为假设 H_1 的均值

而且，两者的协方差矩阵为 $V_0 = V_1 = V$

➤ 含偏置的 **Fisher** 甄别量为 $t(\vec{x}) = a_0 + (\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1} \vec{x}$

➤ 利用前面所述的似然比对给定效率条件下的最大纯度

$$\begin{aligned} r &= \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} = \exp \left[-\frac{1}{2} (\vec{x} - \vec{\mu}_0)^T V^{-1} (\vec{x} - \vec{\mu}_0) + \frac{1}{2} (\vec{x} - \vec{\mu}_1)^T V^{-1} (\vec{x} - \vec{\mu}_1) \right] \\ &= \exp \left[(\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1} \vec{x} - \frac{1}{2} \vec{\mu}_0^T V^{-1} \vec{\mu}_0 + \frac{1}{2} \vec{\mu}_1^T V^{-1} \vec{\mu}_1 \right] \end{aligned}$$

$\propto e^t$



$t \propto \log(r) + \text{常数 (单调变化)}$



**Fisher 甄别量
与似然比等效。**

如果不是多变量高斯分布，上式不成立。

Sigmoid, Logistic sigmoid

给定 $\mathbf{x}$, 利用贝叶斯定理判断假设的情况:

$$P(H_0|\vec{x}) = \frac{f(\vec{x}|H_0)\pi_0}{f(\vec{x}|H_0)\pi_0 + f(\vec{x}|H_1)\pi_1} = \frac{1}{1 + \pi_1/\pi_0 r}$$

其中的 r 正好是似然比

利用Fisher甄别量 t 改写将似然比 r :

$$P(H_0|\vec{x}) = \frac{1}{1 + e^{-t}} \equiv s(t)$$

Sigmoid
function-
乙型函数
s型函数

$$a_0 = -\frac{1}{2}\mu_0^T V^{-1}\mu_0 + \frac{1}{2}\mu_1^T V^{-1}\mu_1 + \log \frac{\pi_0}{\pi_1}$$

$s(t)$ 取值在0-1之间,
逻辑Sigmoid function

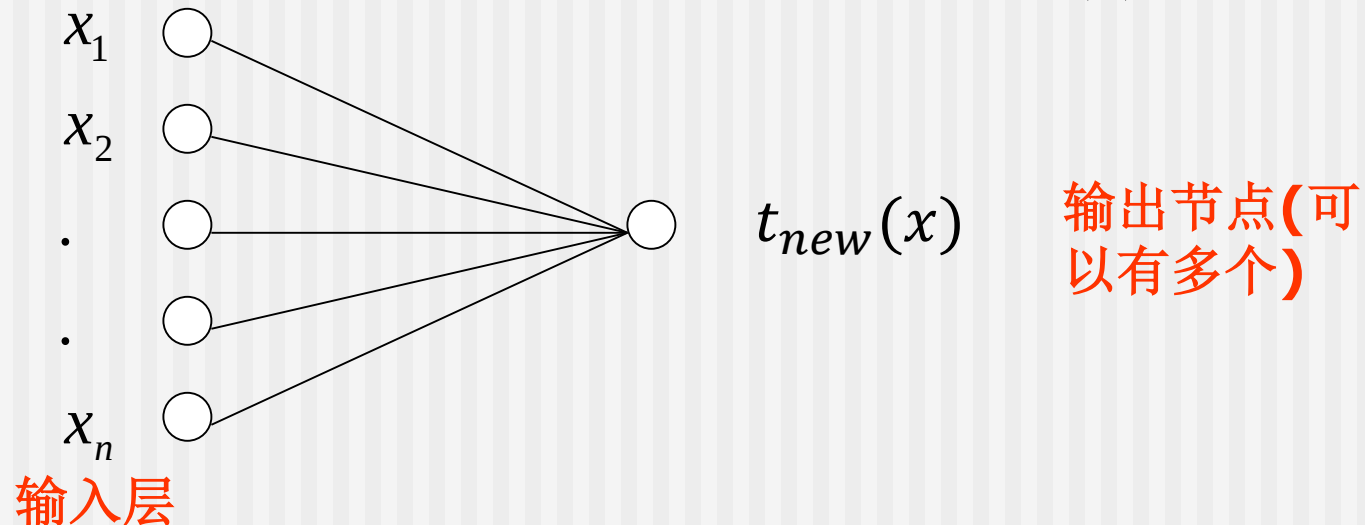
非线性检验统计，神经网络(一)

如果不同假设下得到实验观测量的概率密度函数 $f(\vec{x}|H_0)$ 与 $f(\vec{x}|H_1)$ 不是高斯或无共同的协方差矩阵，**Fisher** 甄别方法不再适用。此时可以采用更为一般的所谓**神经网络方法**

假设统计检验量 $t_{new}(\vec{x}) = s(a_0 + \sum_{i=1}^n a_i x_i)$

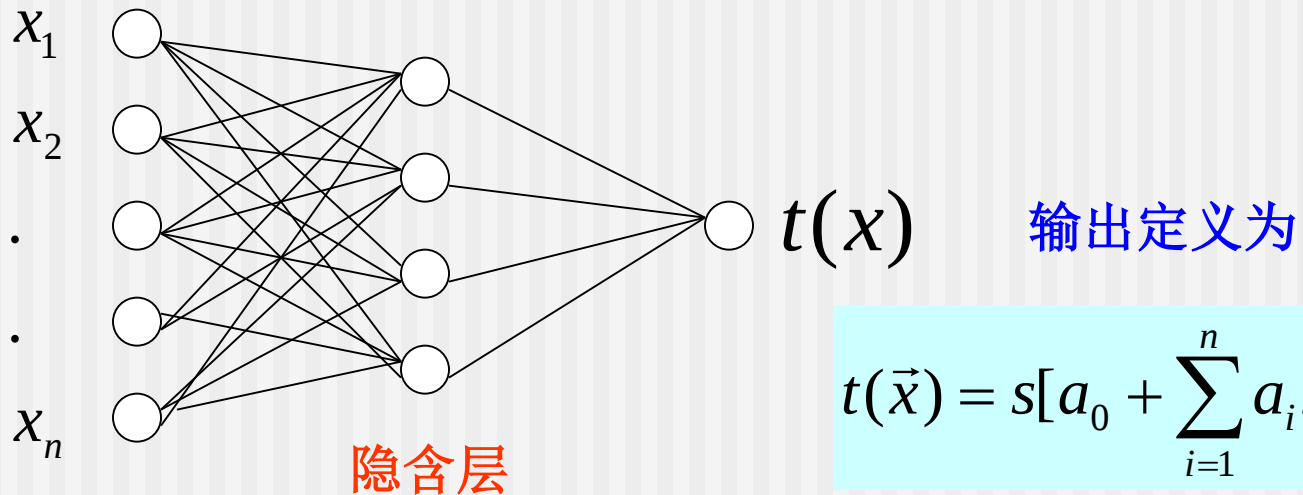
$s(u) = (1 + e^{-u})^{-1}$
激活函数
逻辑S型函数

是单层的感知器。 S 是单调的，因此等效于线性的 $t(\vec{x})$



多层感知器的神经网络

推广到多层感知器



$$t(\vec{x}) = s[a_0 + \sum_{i=1}^n a_i h_i(\vec{x})]$$

上一层节点函数可写为

$$h_i(\vec{x}) = s(w_{i0} + \sum_{j=1}^n w_{ij} x_j)$$

越多节点



神经网络越接近优化的 $t(x)$

但需要定更多的参数!

a_i, w_{ij} 为权重或者联结强度。

神经网络中的误差函数最小化

参数取值通常根据误差函数的最小化结果来决定

$$\varepsilon = E_0[(t - t^{(0)})^2] + E_1[(t - t^{(1)})^2]$$

这里 $t^{(0)}$, $t^{(1)}$ 为目标值, 例如选 0 和 1 的逻辑 σ 函数值

- 如何求解许多的参数的呢? 需要大量的训练样本, 既已经标记好信号或者本底的两种样本, 有点随机寻找的意思, 寻找各个系数。

在核物理与粒子物理研究中, 是通过定义信号与本底两个样本, 从样本中给出每个事例的相关测量量(例如, 动量, 飞行时间...), 然后直接调用欧洲粒子物理实验室(CERN)提供的物理分析软件包ROOT(基于C++)PAW(基于Fortran), 得到训练后的参数与输出量, 并将它们用于待分析的事例来决定其是本底还是信号。具体应用参见下列网站

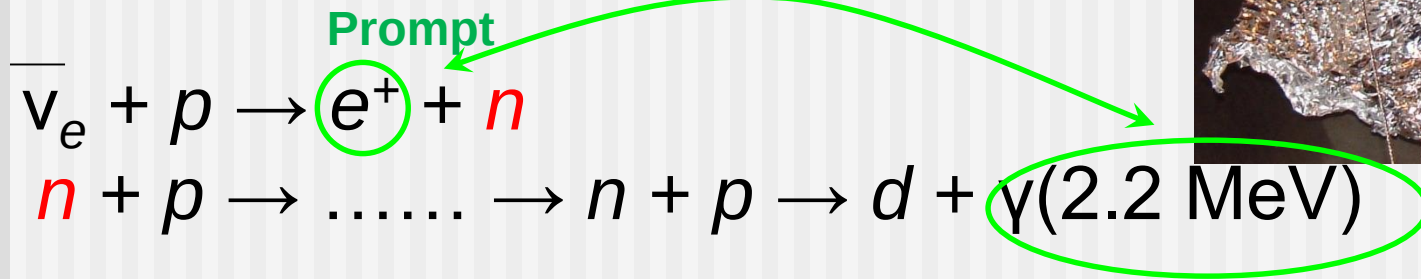
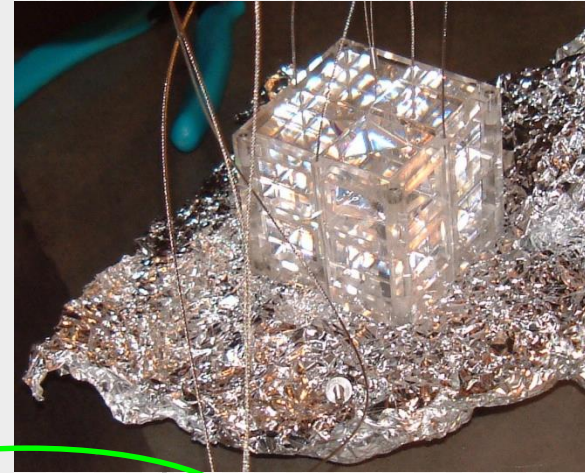
ROOT用户: https://root.cern/doc/master/mlpHiggs_8C.html

(调整参数值=神经网络的学习过程)

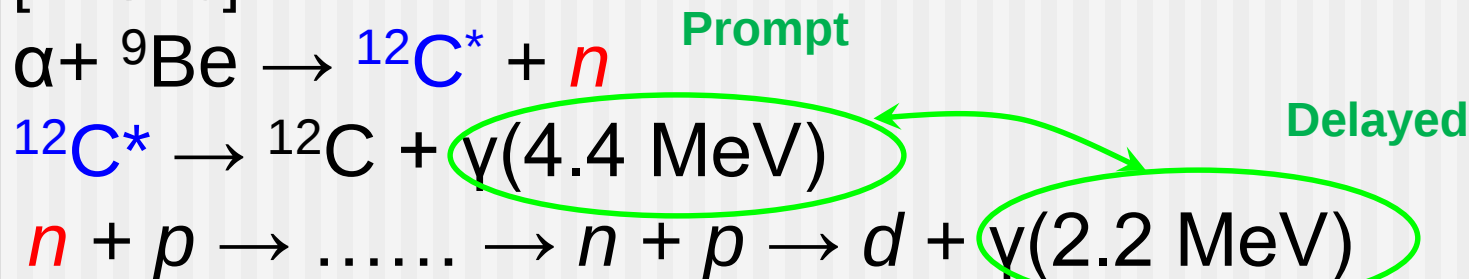
例子：用神经网络甄别中子信号

为了在一个5万吨水的超级神冈实验探测器上探测中子信号(超新星遗迹中微子)，进行了如下实验

~ 200 μ s
Delayed



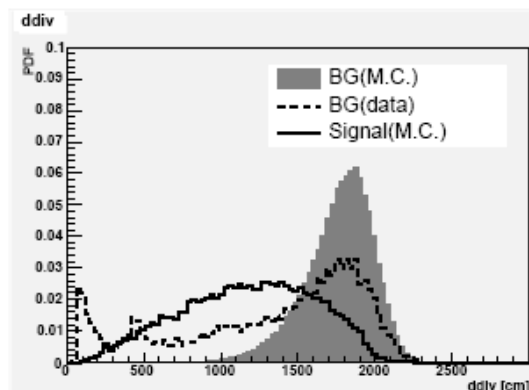
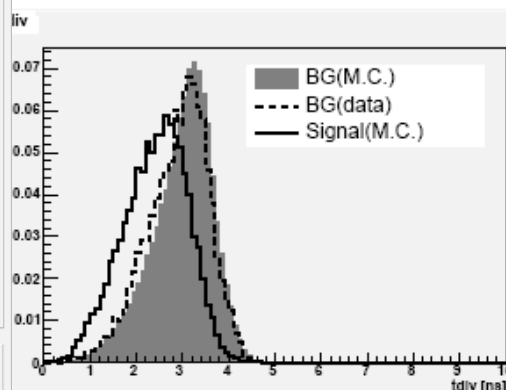
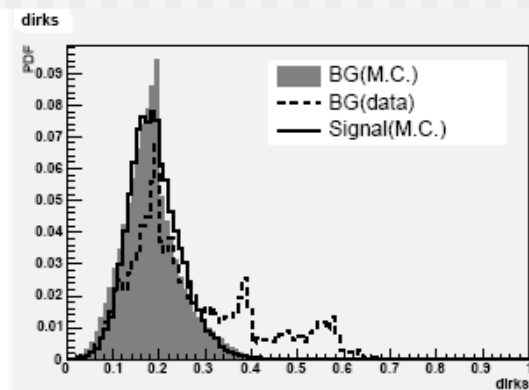
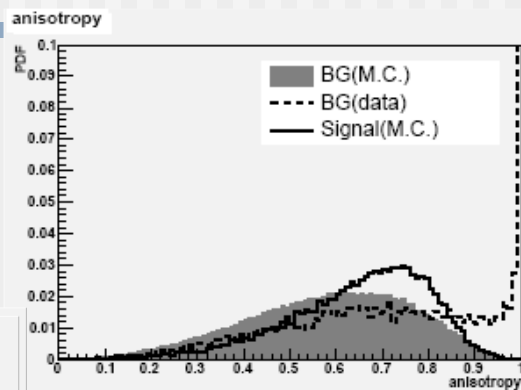
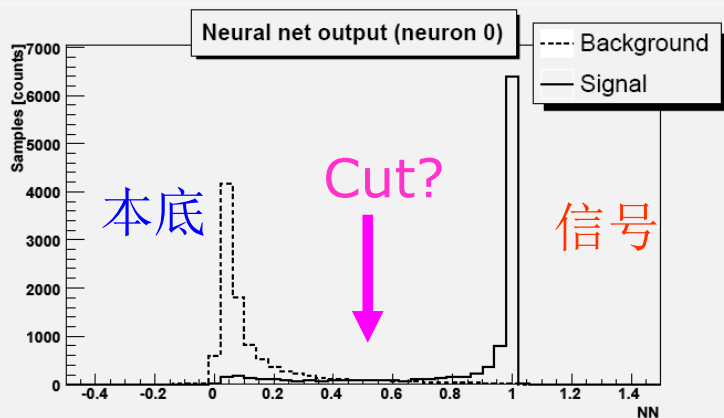
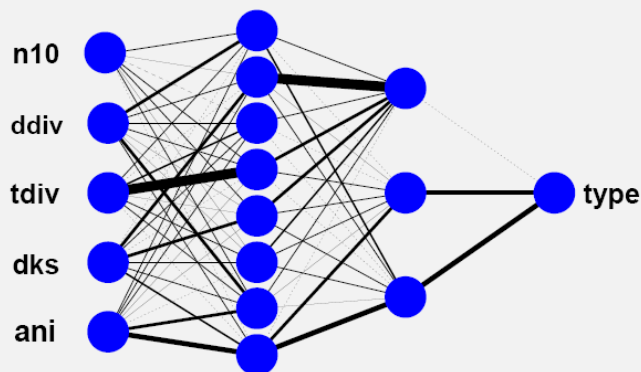
[Am/Be]



希望在纯水中观测到中子被水俘获的现象。已知数据中有大量包括诸如光电倍增管噪音的本底。

神经网络输入量与甄别量

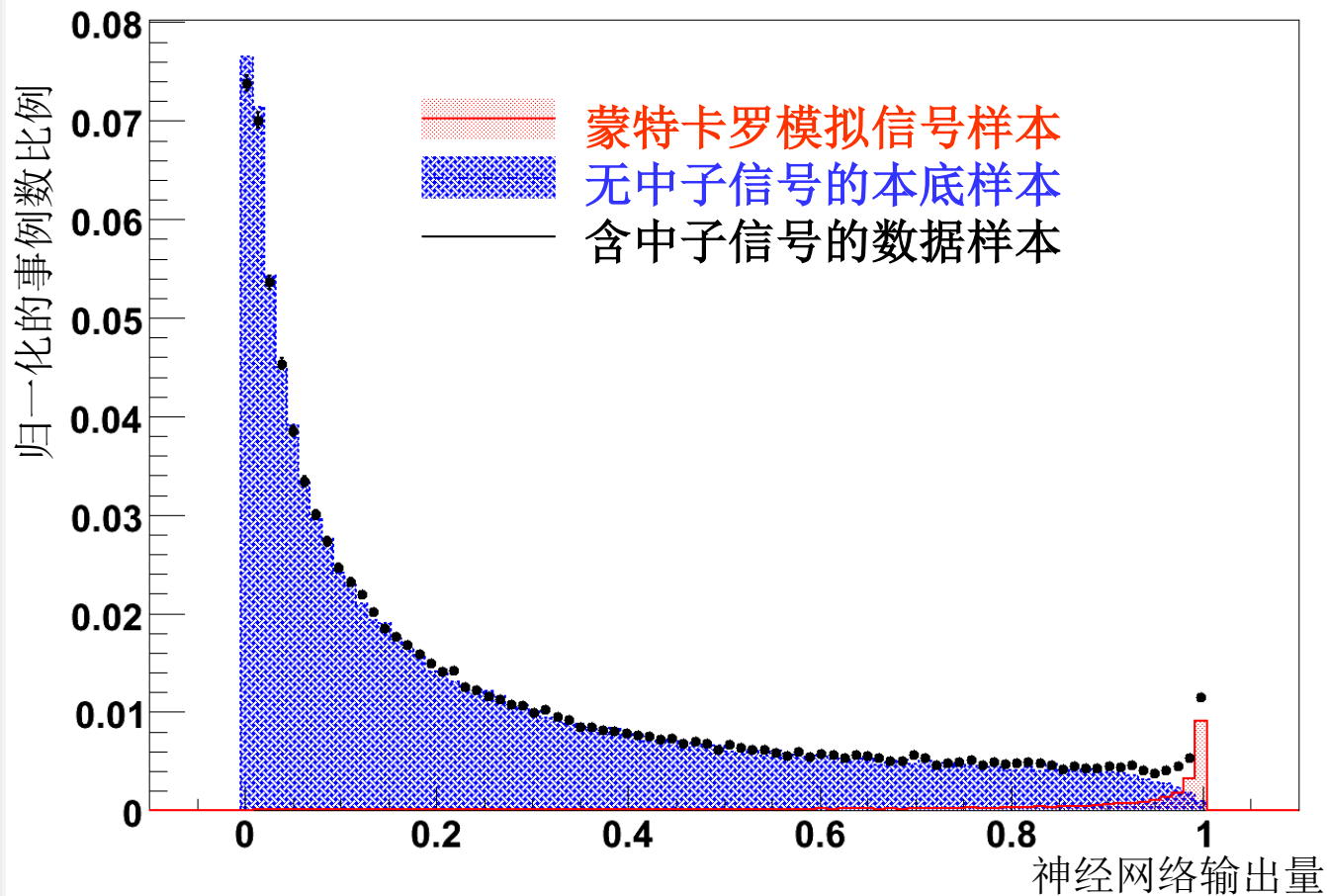
神经网络
输入变量



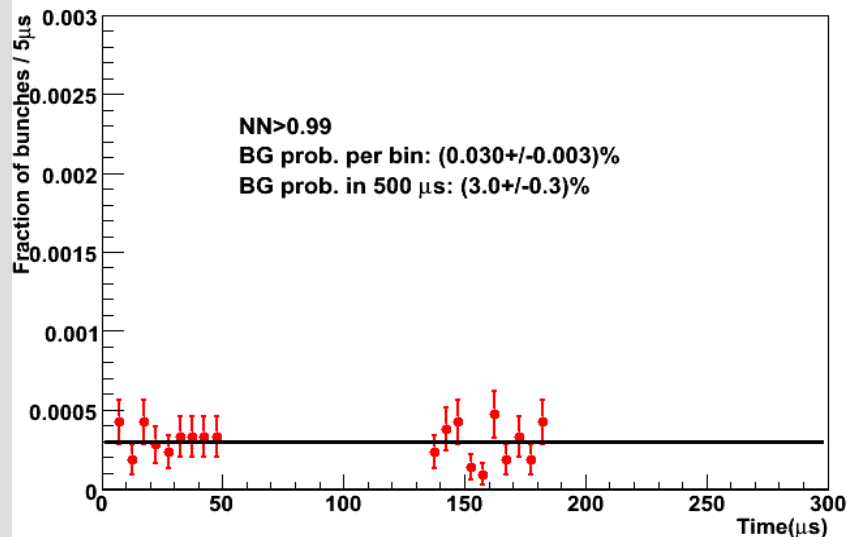
神经网络
输出量

选择拒绝域使得信
噪比与效率最大。

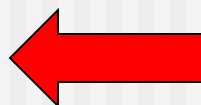
从神经网络输出量中甄别中子



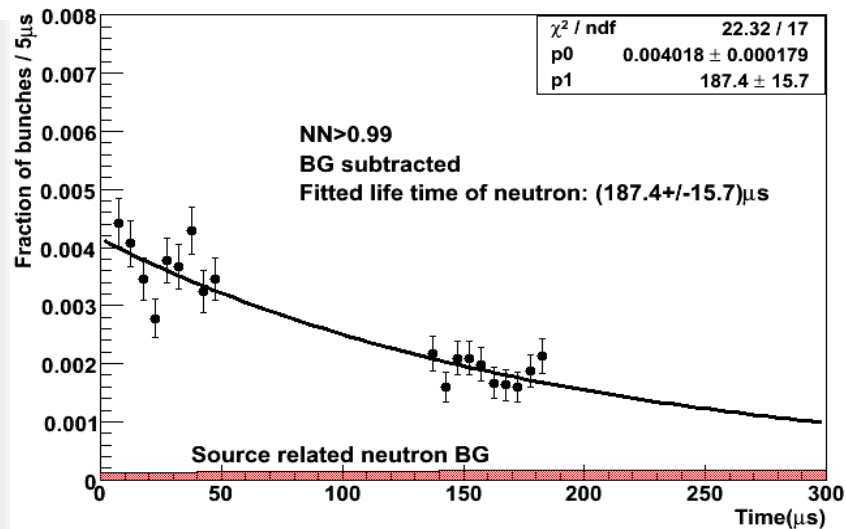
神经网络甄别后的中子信号



在本底无中子信号
数据中看不到体现
中子被水中的氢原子
俘获的寿命特征



在含中子源信号的
数据中看到了体现
中子被水中的氢原子
俘获的寿命特征



关于神经网络的输入变量问题

问题：是否输入量越多越好？

较少的输入量  较少的可调参数

 在有限的样本中，参数可以得到很好的确定

如果输入量之间中有很强相关情形，应只保留一个。

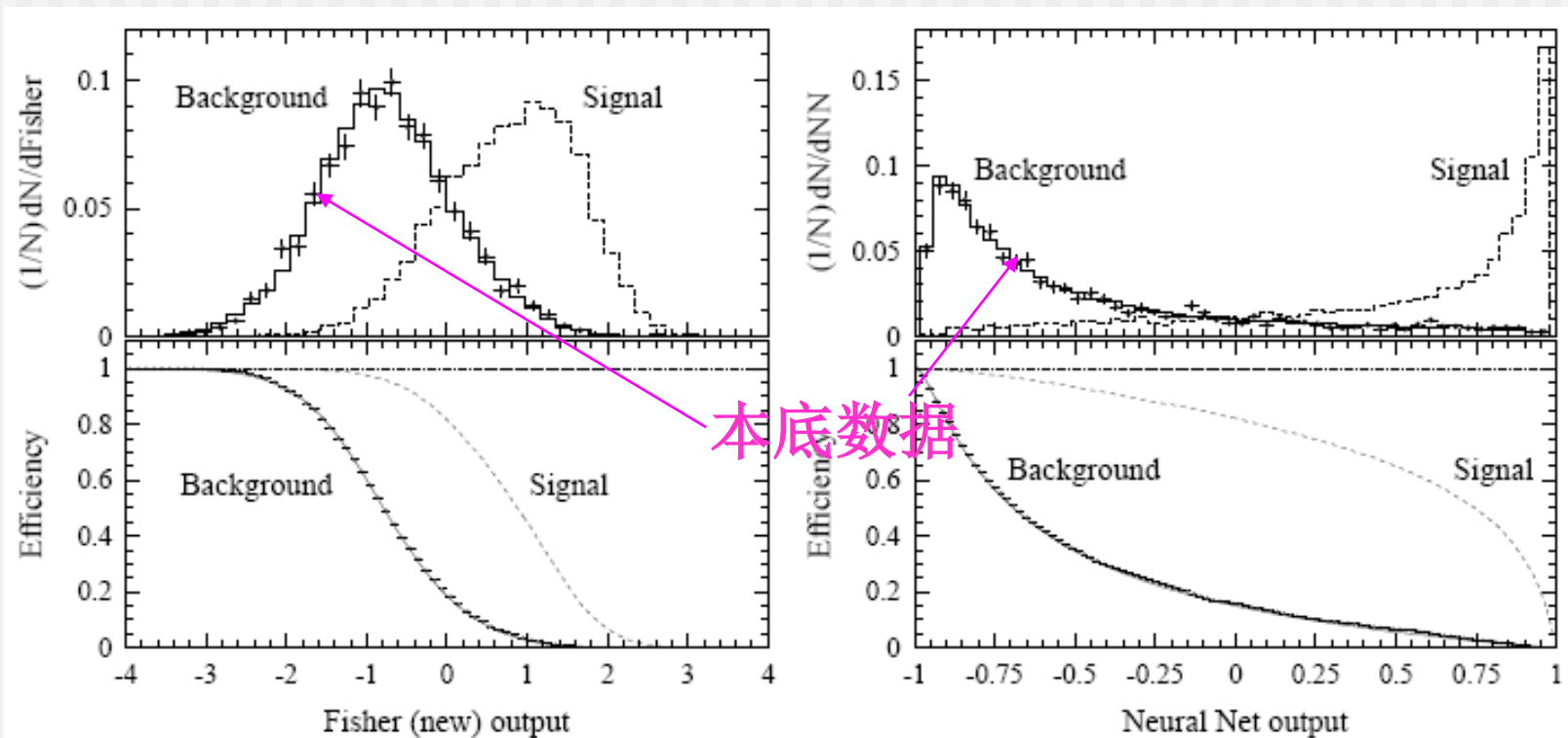
如果输入量对甄别无太大影响，应弃之。

神经网络利用了较高阶矩的联合概率密度函数 $f(\vec{x}|H)$ ，对于太多的输入变量，它们也许在训练的样本中找不到较好的模型来描述

 最好简化 $t(\vec{x})$ ，只要它还能恰当地描述样本。

Fisher方法与神经网络

- **Fisher** 方法只适用于用线性方法构造统计量。
- 神经网络在应用上更具有普遍性和更大的甄别能力。

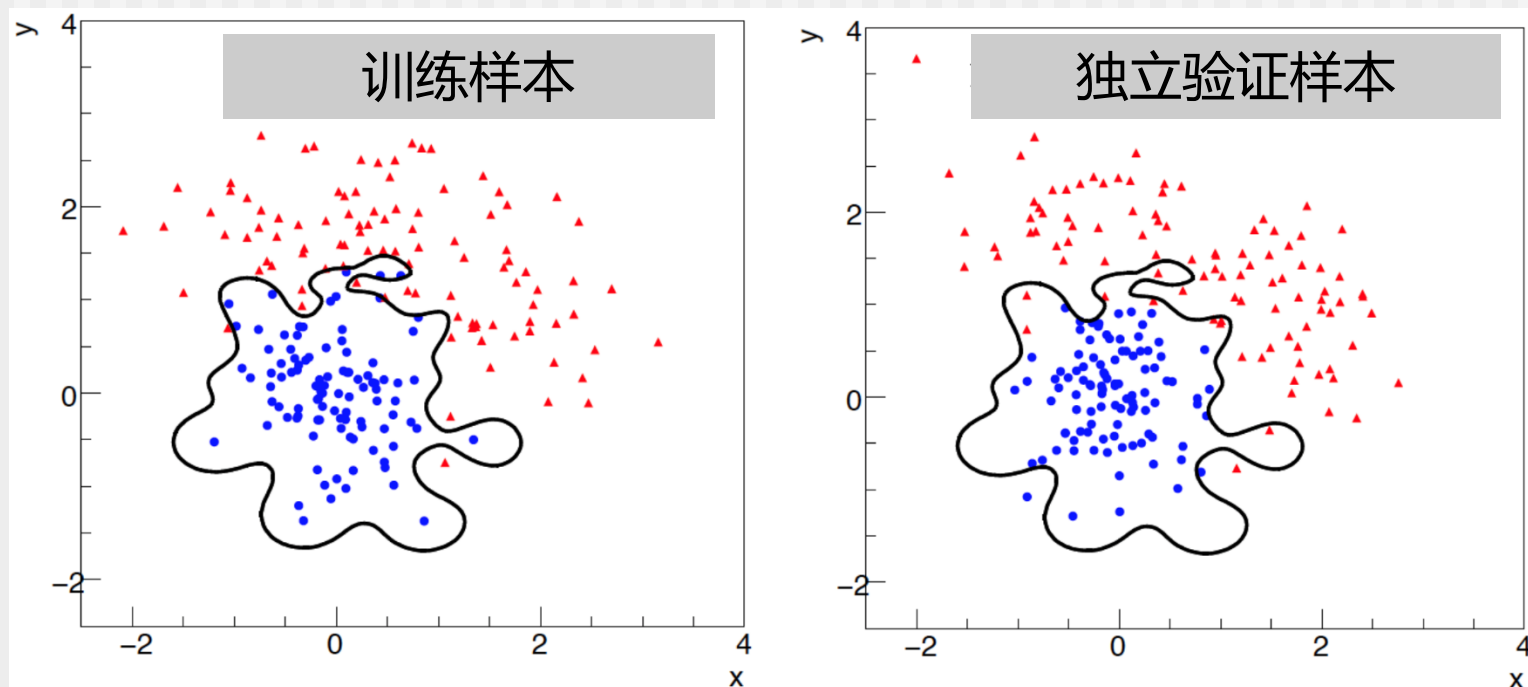


有研究表明，同等本底大小的情况下，神经网络有时能使效率增加15%。参见 [arXiv:hep-ex/0107075](https://arxiv.org/abs/hep-ex/0107075)

过度训练与盲分析

在分类器中包含更多参数可以使得决策边界非常灵活，例如，在神经网络中采用更多节点和层数。

“灵活”的分类器可能会过于迎合训练样本，相同的决策边界在独立的验证数据样本中性能不会太好 → “过度训练”



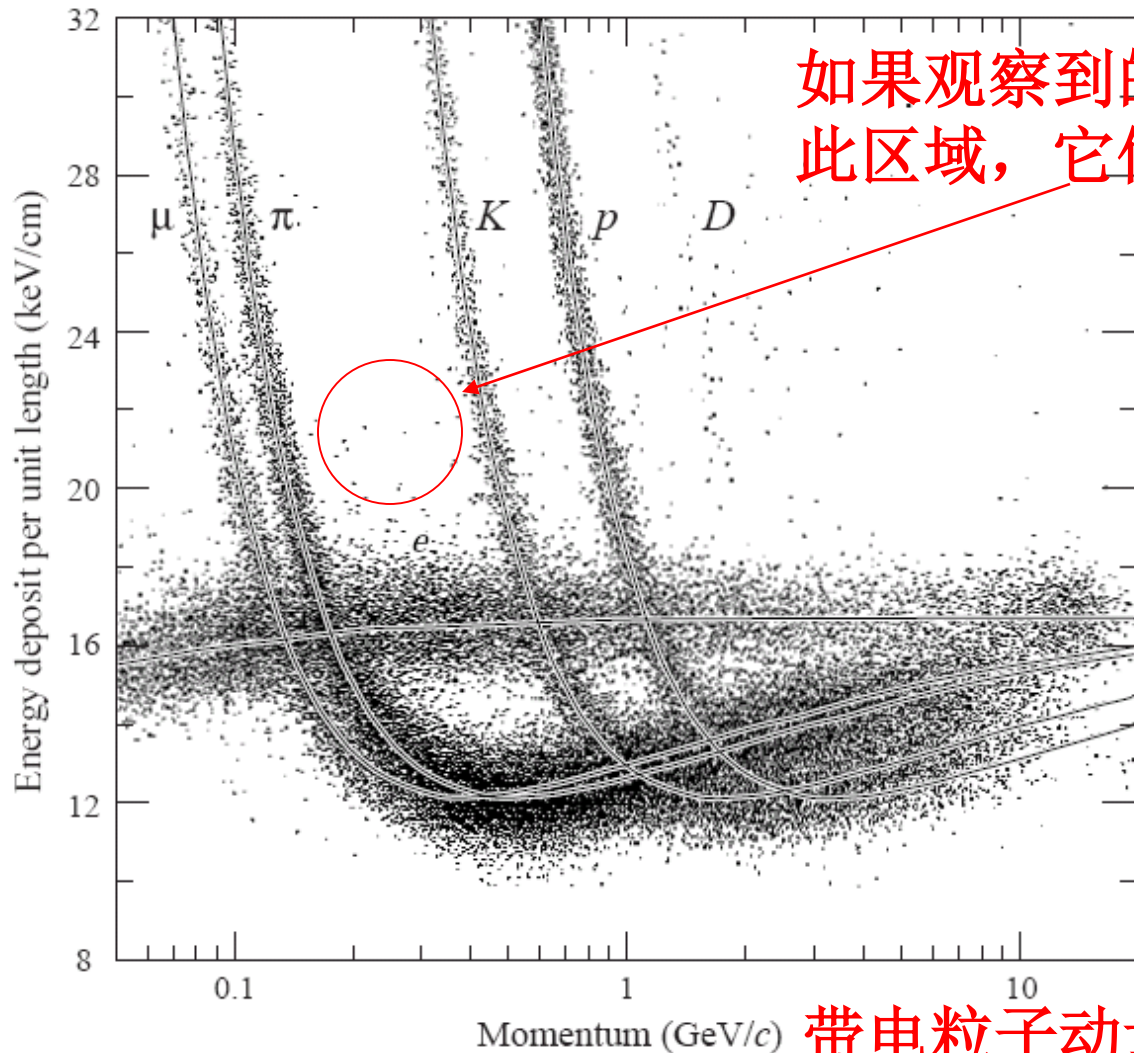
引入盲分析策略

过度训练 / 过拟合



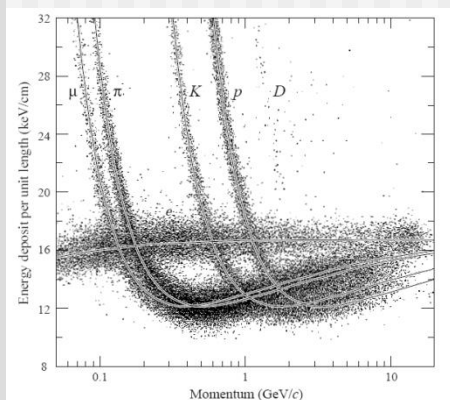
粒子鉴别中常遇到的问题

粒子在每单位长度介质中的能量沉积



检验统计量与拟合优度

通常需要构造统计检验量 $t(\vec{x})$ ，它的大小可以反映出在 $\vec{x}$ 与 H 之间符合的程度。例如



$$\Rightarrow t = \left(\frac{(dE / dx)_K^{Theo} - (dE / dx)_K^{Obs}}{\sigma} \right)^2$$

小的 t



数据与 H 更符合

大的 t



数据与 H 更不符合

(由于概率密度函数 $f(\vec{x}|H)$ 已知，因此在 H 假件下量 t 的概率密度函数 $g(t|H)$ 是完全可以确定。)

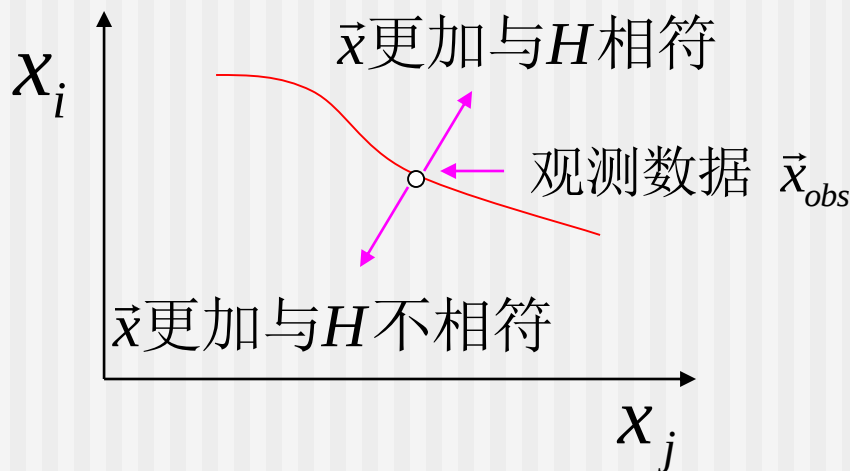
检验拟合优度

如果假设 H 对一组观测量 $\vec{x} = (x_1, x_2, \dots, x_n)$ 预言了 pdf $f(\vec{x} | H)$ 。

如果在 $\vec{x}$ -空间中观察到一个点: $\vec{x}_{obs}$, 那么数据在检验假设 H 的有效性方面, 可以得出什么样的结论呢?



需要决定 $\vec{x}$ - 空间中哪一部分比观测点 $\vec{x}_{obs}$ 更能代表与假设 H 的不相符。



为了达到此目的...

P-值定义

将拟合优度用P-值表示（也称为检验的观测显著水平或置信水平）

P = 观察到实验数据 $\vec{x}$ 或 $t(\vec{x})$ 像 $\vec{x}_{obs}$ 或 $t(\vec{x}_{obs})$ 一样，与假设 H 具有相同或较小符合程度的概率。

注意：这不是 H 为真的概率。

例子:硬币问题中的 P -值计算

任意投掷一枚硬币，结果为正面与反面的概率都是**0.5**。

如果有人声称对此进行了检验。投了**20**次，得到了**17**次正面的结果。那么能否断定得到正面的概率应该是

$$p_h = 0.85 \pm 0.08$$

也就是说与预期值 **0.5** 有 **4** 个标准偏差呢？

问题：理论上允许这样的极端情况出现吗？或者说与这样一种极端情况相等或更高的概率有多大？

例子:硬币问题中的 P -值计算

投 N 次硬币, 观察到 n_h 次头朝上的概率服从二项式分布:

$$f(n_h; p_h, N) = \frac{N!}{n_h!(N - n_h)!} p_h^{n_h} (1 - p_h)^{N - n_h}$$

假设 H : 硬币是公平的(朝上的 p_h = 朝下的 p_t = 0.5)

取拟合优度检验统计量 $t = |n_h - N/2|$

投 $N=20$ 次硬币, 观察到17次头朝上, 则偏离 $= |17 - 20/2| = 7$

在 t -空间中, 具有相同或较少符合的区域为

$$t = |n_h - \frac{N}{2}| \geq 7$$

$$P\text{-值} = P(n_h = 0, 1, 2, 3, 17, 18, 19, 20) = \sum_{i \leq 8} f_i \approx 0.0026$$

拟合优度检验中的问题

问题: 当 P -值等于0.0026, 是否意味着 H 假设是错的?

P -值并不回答此问题。它只是给出与观察到的结果一样, 与 H 假设不符或者高于 H 假设 ($p_h = p_t = 0.5$) 的概率。

P -值=“偶然”得到如此奇怪结果的概率

一种实用的检验方法是在“同样”的假设下, (模拟) 产生同样数目的事例足够多次。检查如此奇怪的结果发生的概率是否与 P -值相当。

实验P值获取的“同样条件”

刚刚的实验中，20次抛硬币，我们说出现3次及以下头朝上和17次及以上头朝上都是和我们的观测17次头朝上同样的与50%概率假设一样的不符或更差的情形，我们计算得到

$$P\text{-值}=0.26\%$$

重新设计实验，抛硬币，一直抛到有三个头，于时停止，我们可以计算出现一直抛到20次或以上才出现以上情形的总概率，作为与我们实验一样不符或更差的概率作为P值

$$P\text{-值}=0.072\%$$

例：观测到一个信号的显著程度

假设观测 n 个事例，包含了

n_b = 已知过程(或本底)的事例数 n_s = 新过程(或信号)的事例数

如果 n_b , n_s 服泊松分布，均值 ν_b , ν_s ，它之和 $n = n_b + n_s$ 也是服泊松分布，均值 $\nu = \nu_b + \nu_s$ ：

$$P(n; \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}$$

如果 $\nu_b = 0.5$ ，而且观测到 $n_{obs} = 5$ 

假设 H : $\nu_s = 0$ ，即只有本底过程出现。



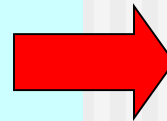
也就是所谓的“无效假设”

可否就此声称该
迹象为新的发现？

观测到一个信号的显著程度(续)

对应的 P -值

$$\begin{aligned} P\text{-值} &= P(n \geq n_{obs}) \\ &= \sum_{n=n_{obs}}^{\infty} P(n; v_s = 0, v_b) \\ &= 1 - \sum_{n=0}^{n_{obs}-1} \frac{v_b^n}{n!} e^{-v_b} \\ &\approx 1.7 \times 10^{-4} \end{aligned}$$



给出了得到这种极端结果的概率：虽然很小但不为零！通常用于决定是否发现了新的效应。

潜在的问题之一

一个误导读者但又常常被使用的结果表示...

对 v_s 估计时得到: $n_{obs} = 5$
估计 n 的标准偏差为: $\sqrt{n} = 2.2$ } 信号



➡ v_s 的估计值: $n_{obs} - v_b = 4.5 \pm 2.2$ 即与零有两倍的标准偏差

但上面的结果表示隐含了均值为**4.5**, $\sigma = 2.2$ 的高斯变量给出零或更少的概率:

➡
$$\int_{-\infty}^0 \frac{1}{\sqrt{2\pi} \times 2.2} \exp\left(\frac{-(x - 4.5)^2}{2 \times 2.2^2}\right) dx = 0.021$$

如果 $v_s \gg 1$, 没有问题, 即 n 服从高斯分布。

实际想要的是: 均值 $v_b = 0.5$ 的泊松变量给出观测量大于 **5** 概率是多少?

➡ 概率为 1.7×10^{-4}

潜在的问题之二

实际问题中会涉及系统误差，例如 $\nu_b = 0.8$ ，则概率变为

$$\begin{aligned} P\text{-值} &= P(n \geq 5; \nu_b = 0.8, \nu_s = 0) \\ &= \sum_{n=n_{obs}}^{\infty} P(n; \nu_b = 0.8, \nu_s = 0) \\ &= 1 - \sum_{n=0}^{n_{obs}-1} \frac{\nu_b^n}{n!} e^{-\nu_b} \\ &= 1.4 \times 10^{-3} \end{aligned}$$

虽然本底只增大了0.3，但P却比 $\nu_b = 0.5$ 时大了一个量级。

→ 建议给出与 ν_b 合理变化相对应的P-值范围

应对 v_b 的系统误差方法

例如 v_b 服从 **Gaus(0.5,0.3)** 的分布，则概率变为

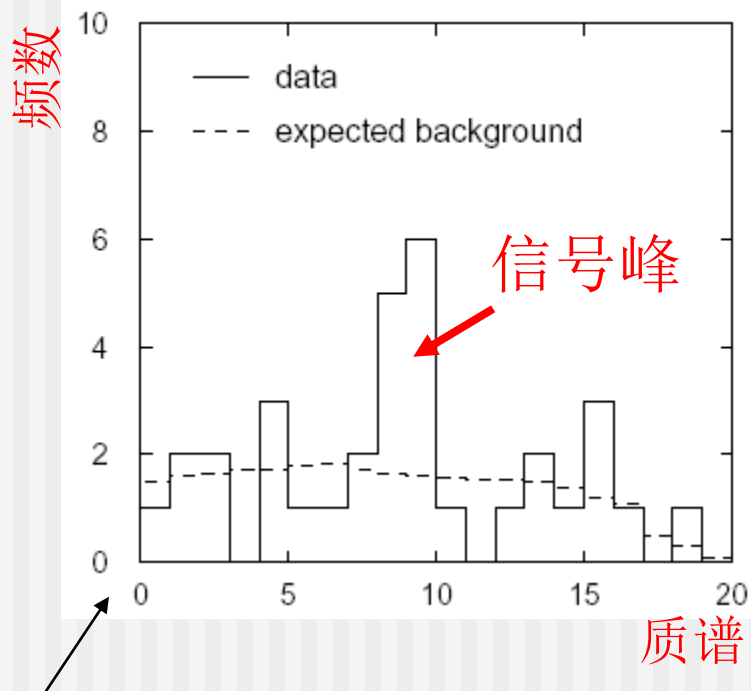
$P - value$

$$= \int Gaus(v_b | 0.5, 0.3) P(n \geq 5; v_b, v_s = 0) dv_b$$

由此我们可以应对 v_b 的变化范围
该计算开始需要很多的数值计算

信号峰的显著性

假设我们不但测量了总的事例数，还测量了每个事例对应的质量。



观察到的实验数据与期待本底大小的直方图，每个区间是泊松分布的一个变量。

在显示信号峰的两个区间，有11个事例，本底估计为 $\nu_b = 3.2$

$$P(n \geq 11; \nu_b = 3.2; \nu_s = 0) = 5.0 \times 10^{-4}$$

Q1: 在哪寻找信号峰?

➡ 计算任何两相连区间的 P -值

Q2: 信号宽度与分辨率相符吗?

➡ 将区间增大至分辨率的几倍

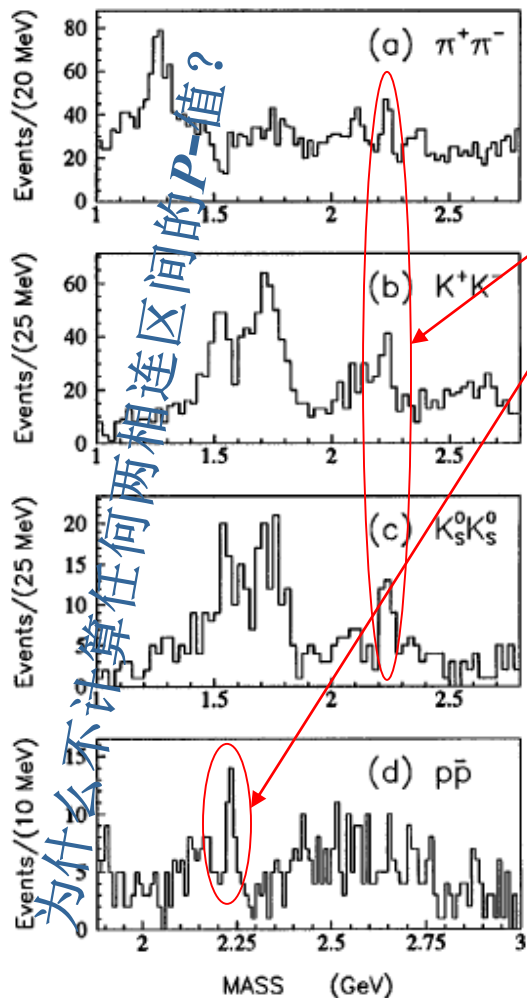
Q3: 信号峰是人为制造出来的吗?

➡ 调整选择条件，分析新数据

...

Qn: 能发表信号峰结果吗?

例子: “ $\xi(2230)$ ” 的观测



PHYSICAL REVIEW LETTERS

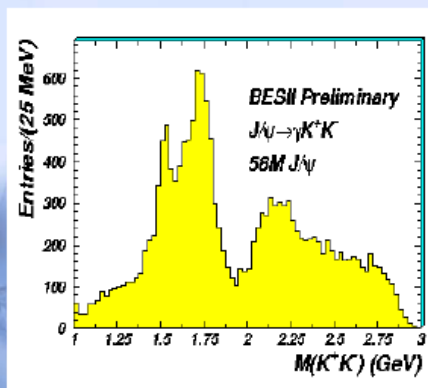
6 MAY 1996

Studies of $\xi(2230)$ in J/ψ Radiative Decays

Y. Chen,¹ S. J. Chen,¹ S. M. Chen,¹ Y. Chen,¹ Y. B. Chen,¹ Y. Q. Chen,¹ B. S. Cheng,¹

the Beijing electron-positron collider has observed the $\xi(2230)$ signal in $J/\psi \rightarrow \gamma K^+ K^-$ final states with 4.6σ , 4.1σ , 4.0σ , and 3.8σ statistical significances. The observations of two nonstrange decay modes of $\xi \rightarrow \pi^+ \pi^-$ and $p \bar{p}$ are in favor of the glueball interpretation of the $\xi(2230)$. [S0031-9007(96)00037-3]

Status of $\xi(2230)$ at BES II



- So far, no clear signal of $\xi(2230)$ has been observed.
- All possible problems are still being checked.

重复实验得不到先前的结果!

皮尔逊的 χ^2 检验与 P -值

在观测的数据 $\vec{n} = (n_1, \dots, n_N)$ 与预言的期待值 $\vec{v} = (v_1, \dots, v_N)$ 之间进行比较的检验统计量

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - v_i)^2}{v_i}$$

如果 n_i 是相互独立而且服从均值为 v_i 泊松分布，所有 v_i 并不太小 (≥ 5)，那么 χ^2 将服从 N 个自由度的最小二乘概率密度函数分布。所观察的 χ^2 可给出 P -值

$$P\text{-值} = \int_{\chi^2}^{\infty} f(z; N) dz$$

这里， $f(z; N)$ 为自由度为 N 的最小二乘概率密度函数。

皮尔逊的 χ^2 检验中的问题

自由度为 N 的最小二乘概率密度函数的期待值为 $E(z)=N$

→ 通常以 χ^2/N 来体现符合的程度

最好分别给出 χ^2 , N , 例如

$$\chi^2 = 15, N = 10 \rightarrow P\text{-值} = 0.13$$

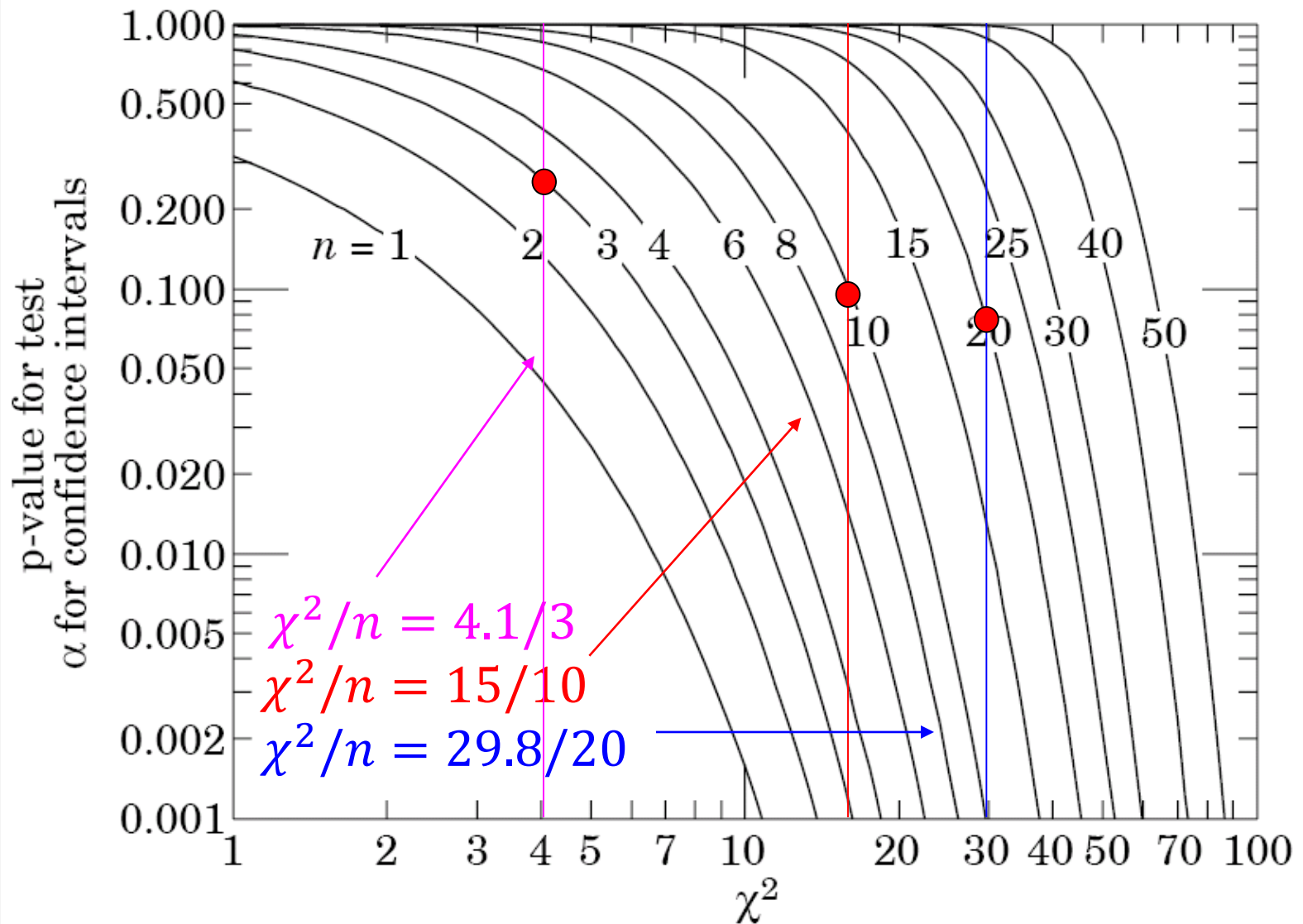
$$\chi^2 = 150, N = 100 \rightarrow P\text{-值} = 9.0 \times 10^{-4}$$

如果 $n_{tot} = \sum_{i=1}^N n_i$ 固定, n_i 服从二项式分布, $p_i = v_i / n_{tot}$, 则

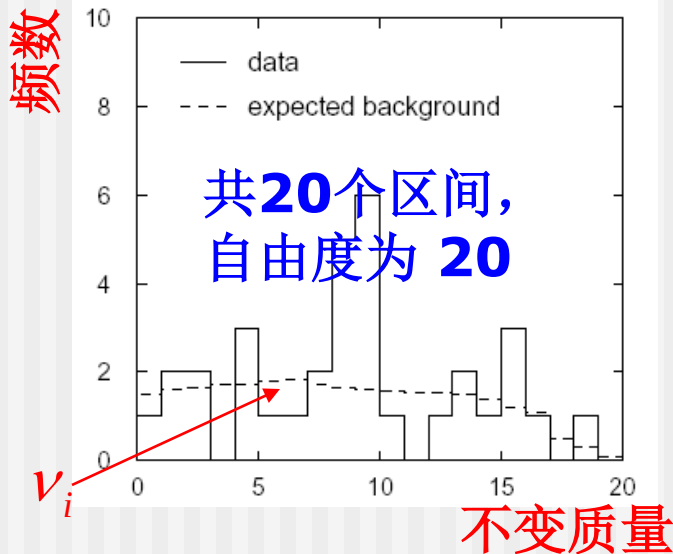
$$\chi^2 = \sum_{i=1}^N \frac{(n_i - p_i n_{tot})^2}{p_i n_{tot}}$$

服从 $N-1$ 自由度的
 χ^2 分布 ($p_i n_{tot} \gg 1$)

P-值与皮尔逊的 χ^2



例：小统计量的 χ^2 检验

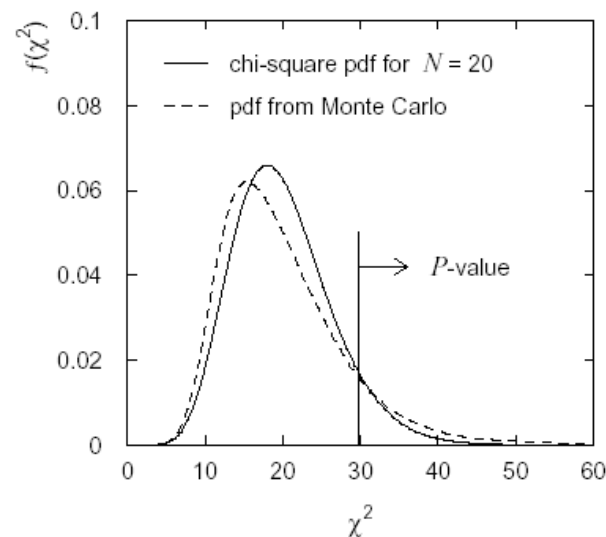


$$\chi^2 = \sum_{i=1}^N \frac{(n_i - v_i)^2}{v_i} = 29.8$$

- 1) 由于许多区间只有很少或根本没有计数，它将不服从 χ^2 的概率密度函数分布。
- 2) 皮尔森 χ^2 仍可以作为一个检验统计量。

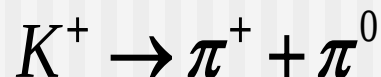
为计算 P -值，先用蒙特卡罗方法
产生 n_i 均值为 v_i 的泊松分布， $i=1, \dots, N$
计算 χ^2 ，填入直方图
重复足够多次得到 $f(\chi^2)$

MC pdf: P -值=0.11
 χ^2 pdf: P -值=0.073



例:大统计量下的 χ^2 检验

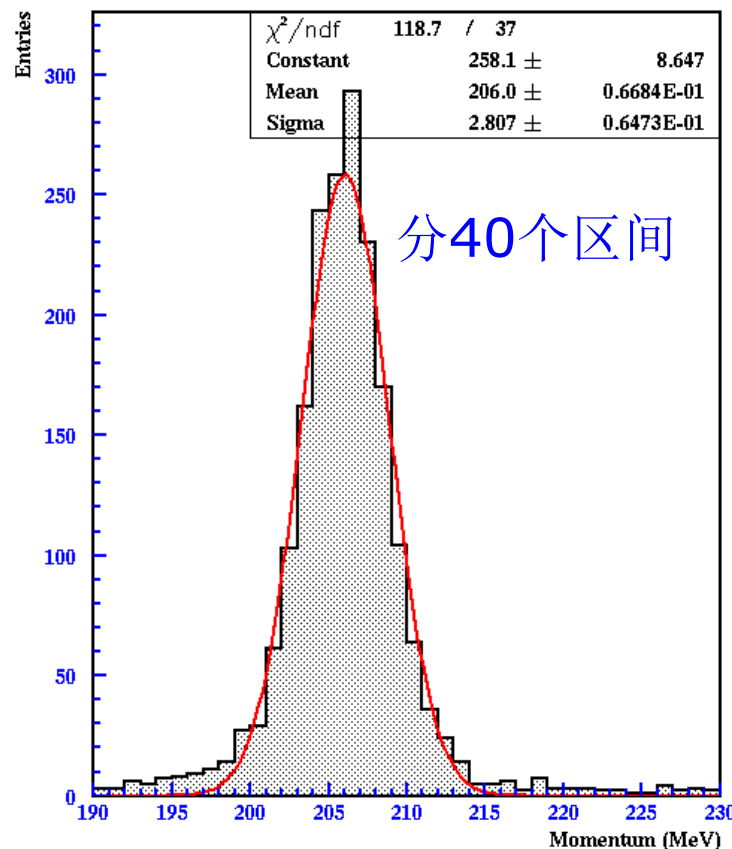
探测器性能研究中经常利用已知的信号进行刻度（或定标），例如



由于母体粒子静止衰变，当测量带电子体动量时，理论上应是单值。为了刻度探测器，需要给出高斯分布的均值与标准偏差。但是

$$\begin{aligned}\chi^2 / ndf &= 118.7 / (40 - 3) \\ &= 118.7 / 37\end{aligned}$$

P -值远小于0.001。由于其它物理或非物理贡献，使得分布不服从高斯分布。采用得到的参数进行刻度可能会有问题。



分析原因，找到解决办法

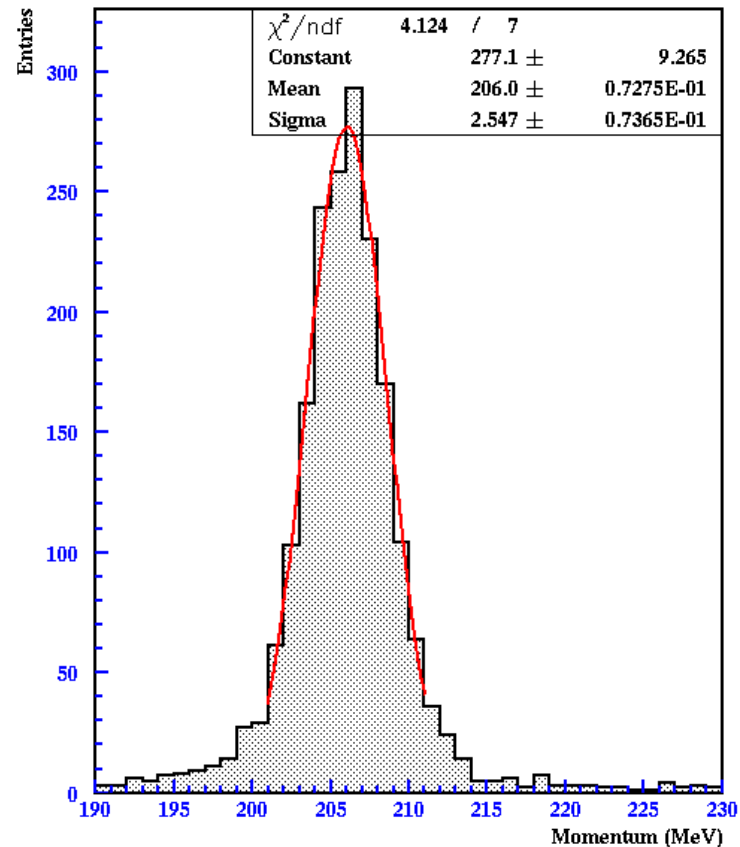
低端尾部分布来自

- 1) $K^+ \rightarrow \pi^+ \pi^0 \gamma$
- 2) π^+ 与探测器物质相互作用

最佳方法是纯化样本。近似处理可采用提取核心贡献项的方法。

$$\begin{aligned}\chi^2 / ndf &= 4.1 / (10 - 3) \\ &= 4.1 / 7\end{aligned}$$

使P-值在很好的容许范围。在不考虑次级效应的情况下，给出核心贡献项的均值与标准偏差。



这里除了检验测量量是否符合高斯分布以外，还涉及到均值与标准偏差的参数估计问题。

P -值与假设检验

对 H 假设拟合优度的检验可以通过计算 P -值的大小来完成。但是应注意：

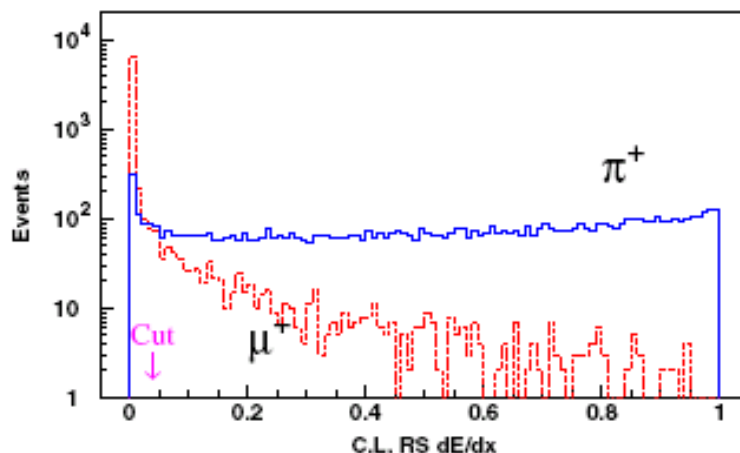
➤ P -值是一个随机变量。

➡ 如果 H 为真，则对于连续的 $\vec{\chi}$ ， P 在 $[0,1]$ 范围内均匀分布。

➡ 如果 H 非真，则 P 的概率密度函数通常很接近零。

例如，在闪烁体叠层中根据每一层测量的电离能损，并与利用测量轨迹长度与不同粒子假设估计出的电离能损进行比较，可以计算 P -值进行 π/μ 粒子鉴别。

Phys. Rev. D77, 052003 (2008)



小结

□ 统计检验:

检验在何种程度上, 数据与假设相符。

□ 检验统计量:

将矢量 $\vec{x}$ 简化为一个或几个分量的矢量 $t(\vec{x})$

□ 检验的要点:

关键区, 显著水平, 功效, 纯度, 效率。

□ 纽曼-皮尔森引理:

在给定效率条件下, 给出纯度最大区。

□ 构造检验统计量:

最好是似然比, 但通常需太多待定参数。

□ 统计分析中两种方法:

Fisher 甄别函数(线性的); 神经网络(非线性的)。

小结（续）

□ 检验拟合优度， P -值定义与应用

P -值为得到数据像已观测的结果一样与假设不符或更不符合的概率。

□ 信号观测的显著程度

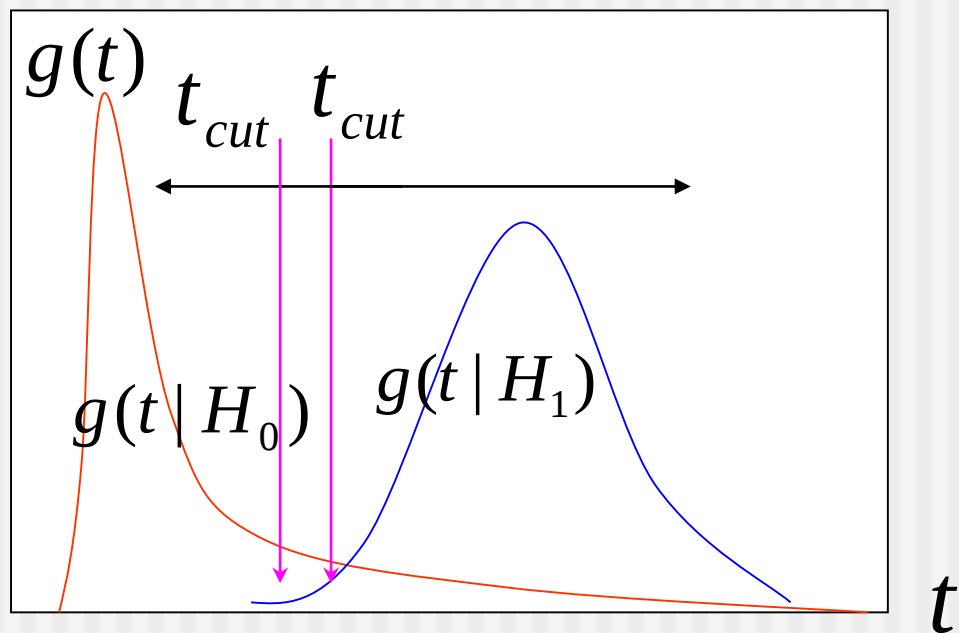
很复杂，许多具有 10^{-4} 效应的结果最终证明是统计涨落的受害者。

□ 皮尔逊 χ^2 检验以及与 P -值的关系

广泛用于检验统计量。对于小样本数据，它将不服从 χ^2 的概率密度函数分布。但仍可用蒙特卡罗得到概率密度函数分布。

讨论题1 定cut位置的原则？

- 有两个待研究的模型 H_0 和 H_1 ，并且经过我们的学习和研究，我们已经找到了一个区分变量 t ，并且也有 t 在 H_0 和 H_1 下的分布，那具体 t 该**cut**在什么位置呢？选取的原则是什么？纯度？统计量？信噪比？



讨论题2 本章的主题

- ❑ 本章的主题是统计检验，我们从来没有讲**P**值为多少既可以声明一个新的发现，这可能是我们真正的哲学或伦理思考，你们的意见如何？
- ❑ 现代实验声明的发现是怎么做到的？例如 Higgs？引力波？