粒子物理与核物理实验中的数据分析

陈少敏
清华大学

第四讲：统计检验

本讲要点

- 假设，检验统计量，显著水平，功效
- 两种假设下的统计检验
- 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

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：若宇称守衡，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，是否可断定宇称守衡？

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。

参数的准确由相关的误差大小来表示。

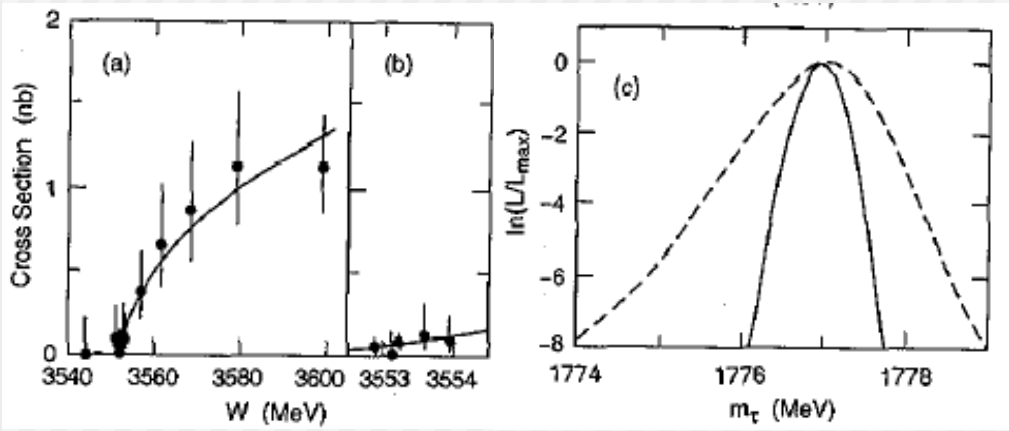
例子：标准模型的检验

标准模型是描述基本粒子之间相互作用的理论，包含十九个参数。其中包含电子，μ，τ等粒子的质量。它预言了费米弱作用藕合常数具有普适性

$$\left(\frac{G_\tau^F}{G_\mu^F}\right)^2 = \left(\frac{m_\mu}{m_\tau}\right)^5 \left(\frac{t_\mu}{t_\tau}\right) B_{\tau\to e\nu\bar{\nu}}$$

- G^F：费米藕合常数
- m：轻子质量
- t：轻子平均寿命
- $B\tau$：τ 纯轻子衰变的百分比

北京正负电子对撞机上的北京谱仪实验通过寻找产生τ-轻子对的最低能量阈，测量了τ–轻子质量参数



$$m_\tau = 1776.96^{+0.18}_{-0.21}\ \text{MeV}$$

$$\Rightarrow \left(\frac{G_\tau^F}{G_\mu^F}\right)^2 = 0.9886 \pm 0.0085 \Rightarrow$$

在1.3个标准偏差范围内与模型预言 1 符合。

假设检验

假如测量结果为 $\vec{x}=(x_1,x_2,...,x_n)$, 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; $x_2=$粒子的平均横动子; $x_3=$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

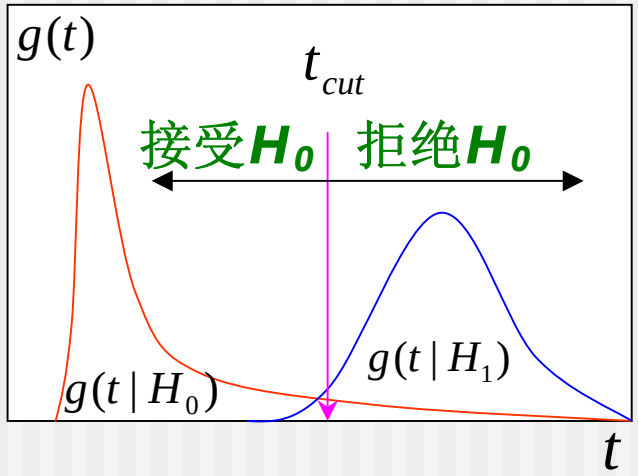
通常情况下很难处理多 维的 $\vec{x}$ 问题，因此， 常常构造低维的统计检 验,在不失去甄别各种假设能 力的条件下，使得 $t(\vec{x})$成为精简后的数据样本 。

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

拒绝域、第一与第二类误差

考虑统计检验量t 服从 $g(t|H_0), g(t|H_1), \ldots$

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误差

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

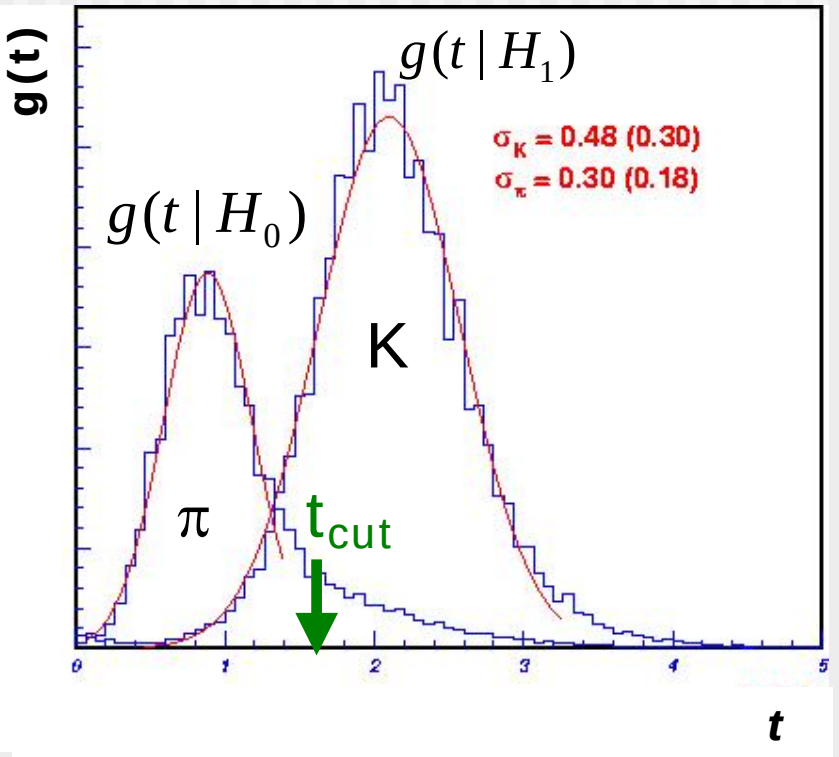
假若接受H_0,但实际情况却是H_1为真的可能性构成第二类误差

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

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，因电离所产生的能损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

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

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

松选择：效率很高，但 K 本底高；
严选择：信号样本纯，但效率低。

π的份额 a_π 可从 t 分布估计

$$f(t;a_\pi) = a_\pi g(t|\pi) + (1-a_\pi) g(t|K)$$

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

→ 两种解释均有道理

通常情况下，需要给出选择样本的纯度

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

$=\pi$ 粒子在区间 $(-\infty, t_{cut}]$ 的概率

注意：$h(\pi|t)$ 有时会被解释为检验统计量。

纽曼- 皮尔森引理与拒绝域

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

问题：如何选择一个最佳的拒绝域或者 **cut**？

纽曼-皮尔森引理：在给定效率条件下，要得到最高纯度的信号样本，或者在给定的显著水平下得到最高的功效，可以选择下列接受域来实现

$$\frac{g(\vec{t}\,|H_0)}{g(\vec{t}\,|H_1)} > c = \text{用以决定效率的常数}$$

对于不含未定参量的最优化一维检验统计量，

$$r = \frac{g(\vec{t}\,|H_0)}{g(\vec{t}\,|H_1)} \quad \Rightarrow \quad \text{简单假设}H_0\text{与}H_1\text{的似然之比}$$

实际应用中，r 最好是单值函数。

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

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

根据纽曼-皮尔森引理，为了选择事例，可选择拒绝域

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

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

→

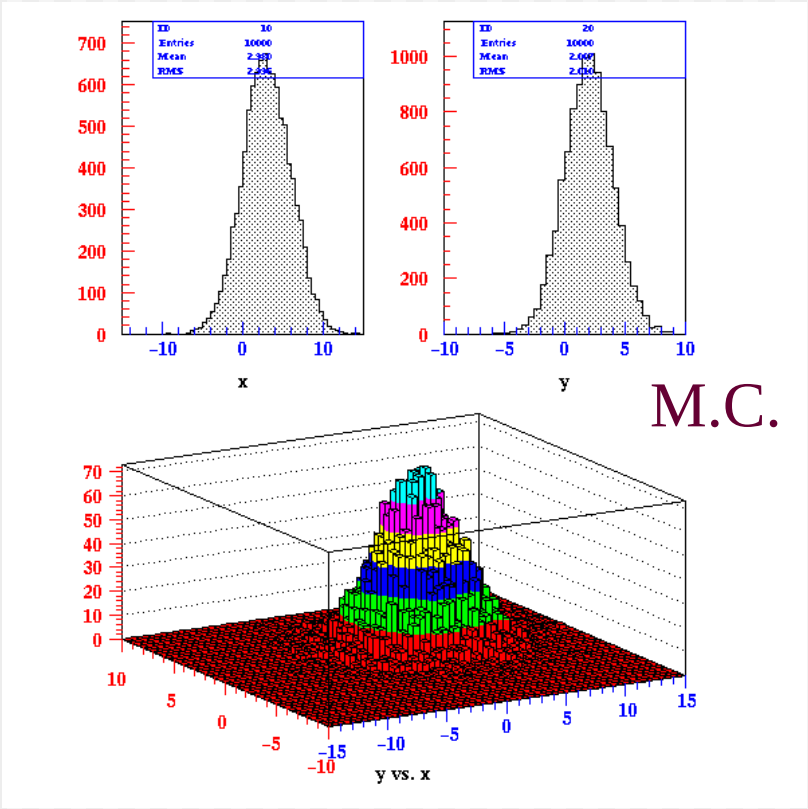
对每个事例，得到测量量$\vec{x}$，并填入$n-$维直方图。如果M为每个分量的区间数，则总单元数为M^n。

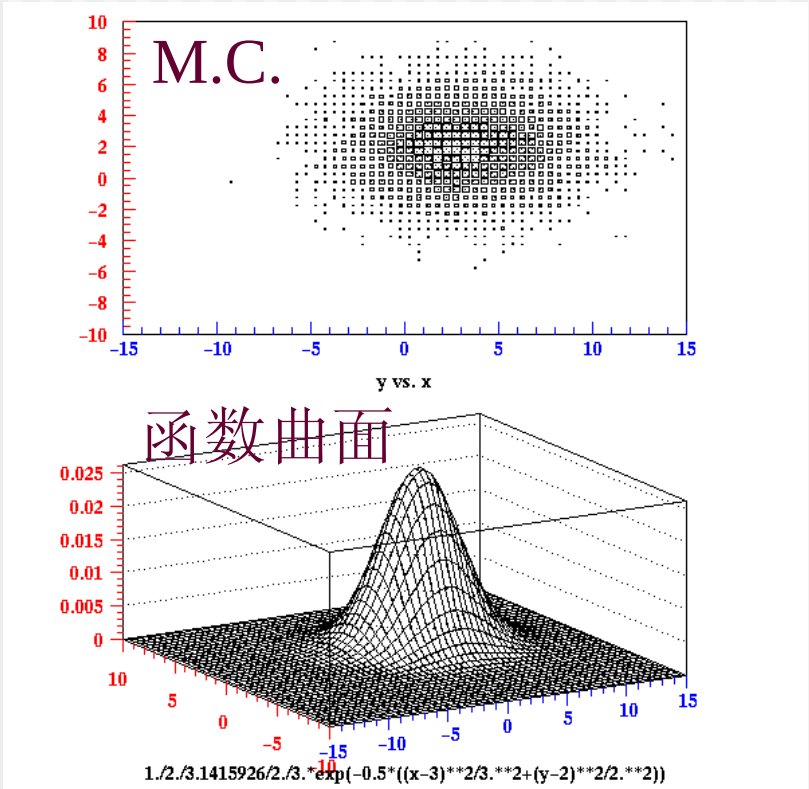
→

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

但是如果 ***n*** 太大时，实际运用会很困难。

例子:蒙特卡罗近似求二维p.d.f.





分格子 ➡ 统计每个格子的频数 ➡ 近似的二维函数

如两者不相关 ➡ 两个一维边缘分布 ➡ $f(x,y)=f(x)f(y)$

线性检验统计量

当维数>2时，用蒙特卡罗法找出多维概率密度函数依然较复杂。假设每一维研究均需要分M个区间，对于n-维问题，需要M^n个格子方能将密度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：$t(\vec{x}) = \sum_{i=1}^{n} a_i x_i = \vec{a}^T \vec{x}$

(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 $\vec{a}$ 最大地区分 $g(t|H_0)$与$g(t|H_1)$ 的目的。

必须定义所谓的区分量。

线性检验统计量（续一）

首先对各测量量，我们可以计算对应的期待值与协方差

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

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

类似地，我们还可以导出计算 $t(\vec{x})$ 平均值与方差的公式

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

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

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。

线性检验统计量（续二）

Fisher 定义了一个甄别法

$$J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\sum_0^2 + \sum_1^2}$$

$$= \sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j = \sum_{i,j=1}^{n} a_i a_j B_{ij} = \vec{a}^T B \vec{a}$$

$$= \sum_{i,j=1}^{n} a_i a_j (V_0 + V_1)_{ij} = \vec{a}^T W \vec{a}$$

则

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

令 $\dfrac{\partial J}{\partial a_i} = 0$ ⇒ $\vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$ （证明见习题）

因此定义了**Fisher** 线性甄别函数。

线性检验统计量（续三）

若将 $t(\vec{x})$ 写成

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

用任意标度和偏置 a_0 去固定 τ_0，τ_1

求 $J(\vec{a}) = \dfrac{(\tau_0 - \tau_1)^2}{\sum_0^2 + \sum_1^2}$ 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 $J(\vec{a})$ 的最大值就是以后介绍的最小二乘法原理中的一种。

Fisher 定理与高斯变量

假设 $f(\vec{x}\,|\,H_k)$是多变量高斯分布，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 $V_0=V_1\equiv V$

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

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

$\propto e^t$

→ $t \propto log(r)$ + 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。

Fisher 定理与高斯变量(续)

具有相同协方差矩阵的多变量 $\vec{x}$ 还可给出验后概率的简单表达式，例如

$$P(H_0 \mid \vec{x}) = \frac{f(\vec{x} \mid H_0)P(H_0)}{f(\vec{x} \mid H_0)P(H_0) + f(\vec{x} \mid H_1)P(H_1)} = \frac{1}{1 + \dfrac{P(H_1)}{P(H_0)r}}$$

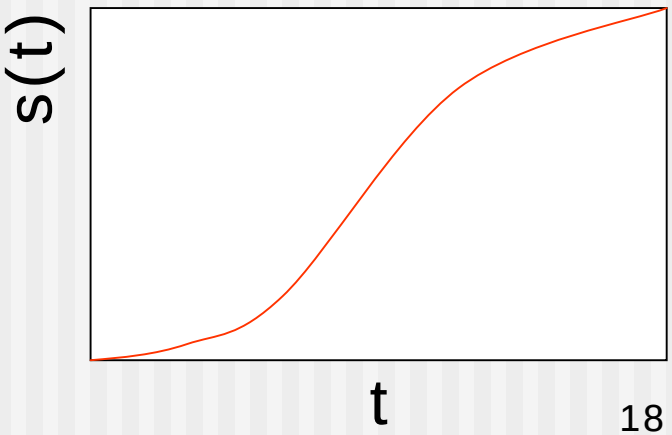
贝叶斯定理

验前概率

选择恰当的偏置a_0，上式可写为

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

也就是所谓的“逻辑σ”函数



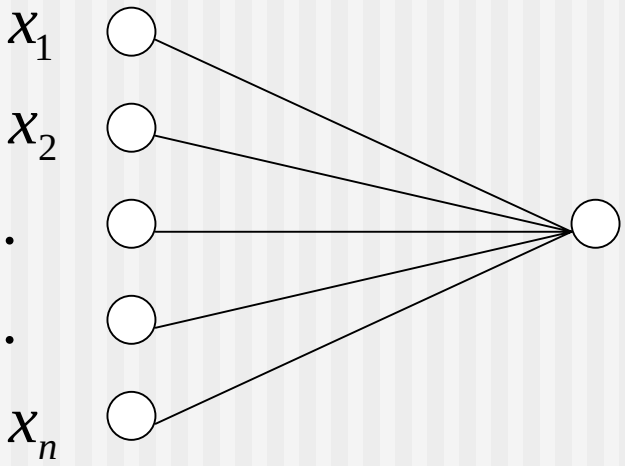
神经网络(一)

神经网络可应用在神经生物学，模式识别，理财预测等等方面，这里它仅作为一种类型的检验统计量。假设 $t(\vec{x})$ 可表示为

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

$$s(u) = (1 + e^{-u})^{-1} = \text{激活函数}$$

是单层的感知器。S是单调的, 因此等效于线性的 $t(\vec{x})$

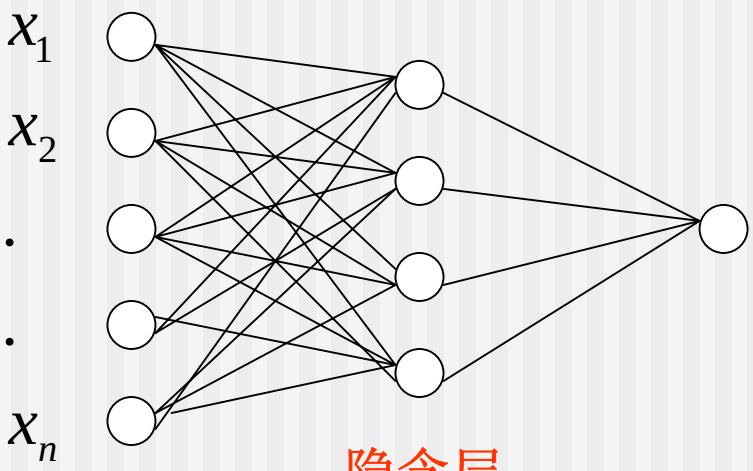


输出节点(可以有多个)

输入层

神经网络(二)

推广到多层感知器



隐含层

输出定义为

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

上一层节点函数可写为

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

a_i, w_{ij}为权重或者联结强度。

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!

神经网络(三)

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

实际应用中，通常以蒙特卡罗的训练样本平均值来取代期待值。
(调整参数值=神经网络的学习过程)

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

PAW 用户：http://paw.web.cern.ch/paw/mlpfit/pawmlp.html

ROOT用户：http://root.cern.ch/root/html/examples/mlpHiggs.C.html

例子:超新星爆发中微子

产生平均能量为十几个MeV的$\nu_e, \bar{\nu}_e$...

距离 L

因中微子质量造成的时间延迟 Δt

$$\frac{\Delta t}{L} = \frac{1}{\beta} - 1 \approx \left(\frac{5.1ms}{10kpc}\right)\left(\frac{10\text{MeV}}{E_\nu}\right)^2\left(\frac{m_\nu}{1\text{eV}}\right)^2$$

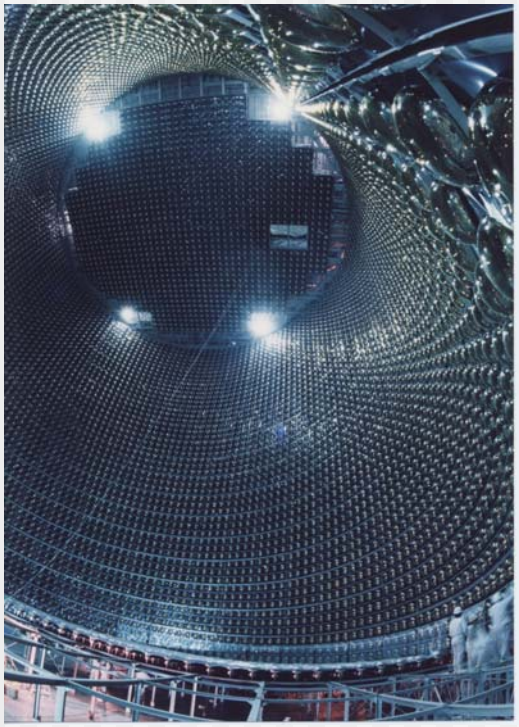
1987A 超新星距地球 52±5 *k*pc

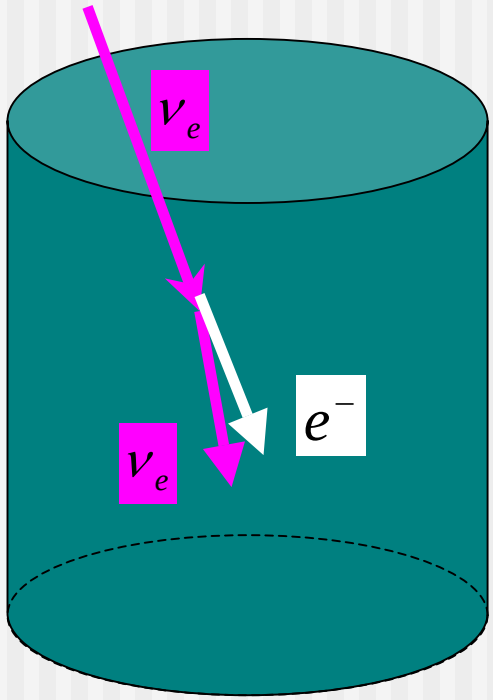
根据中微子到达的时间差，可以给出中微子质量测量。

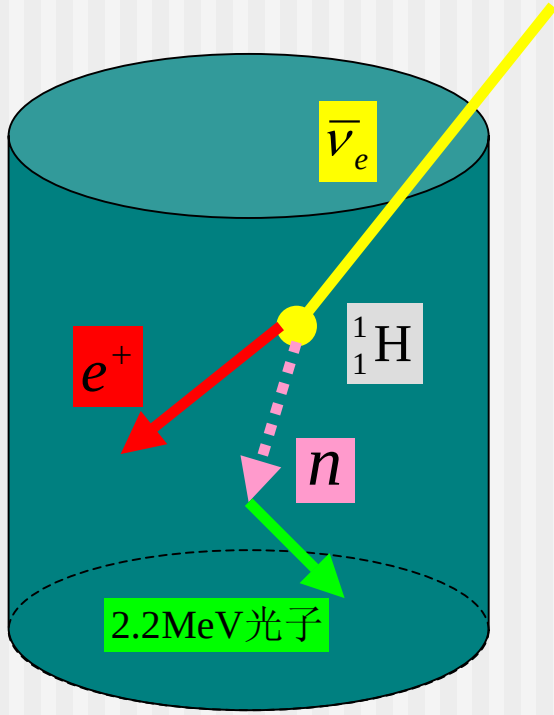
中微子在纯水中的反应

弹性散射：$\nu_e + e^- \rightarrow \nu_e + e^-$

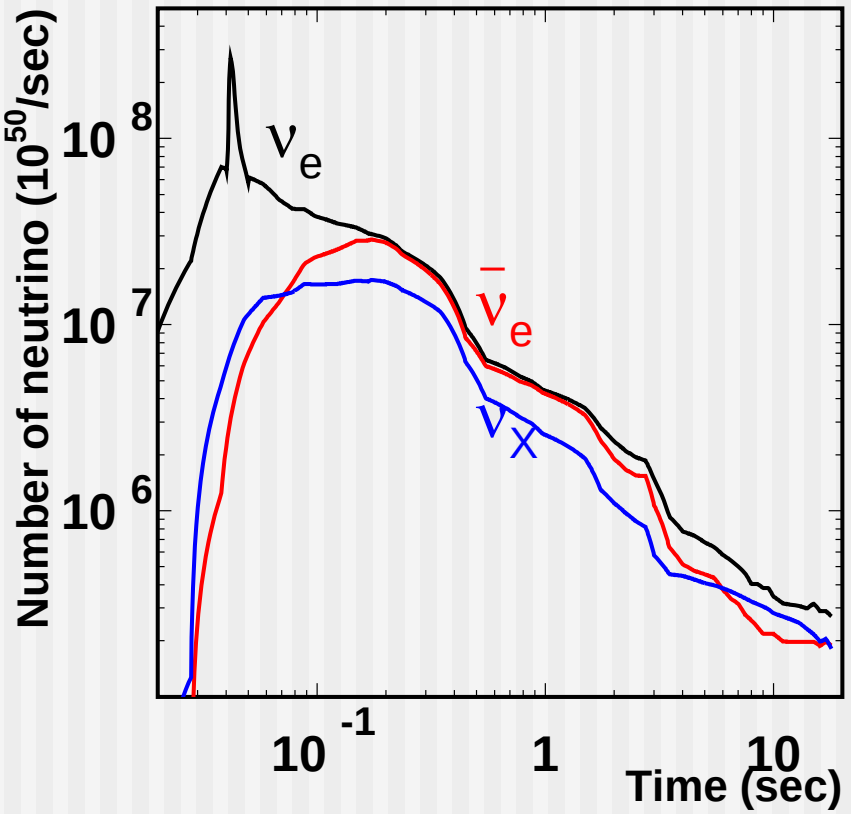
非弹散射：$\bar{\nu}_e + p \rightarrow n + e^+$ ⟹ 反 β 衰变

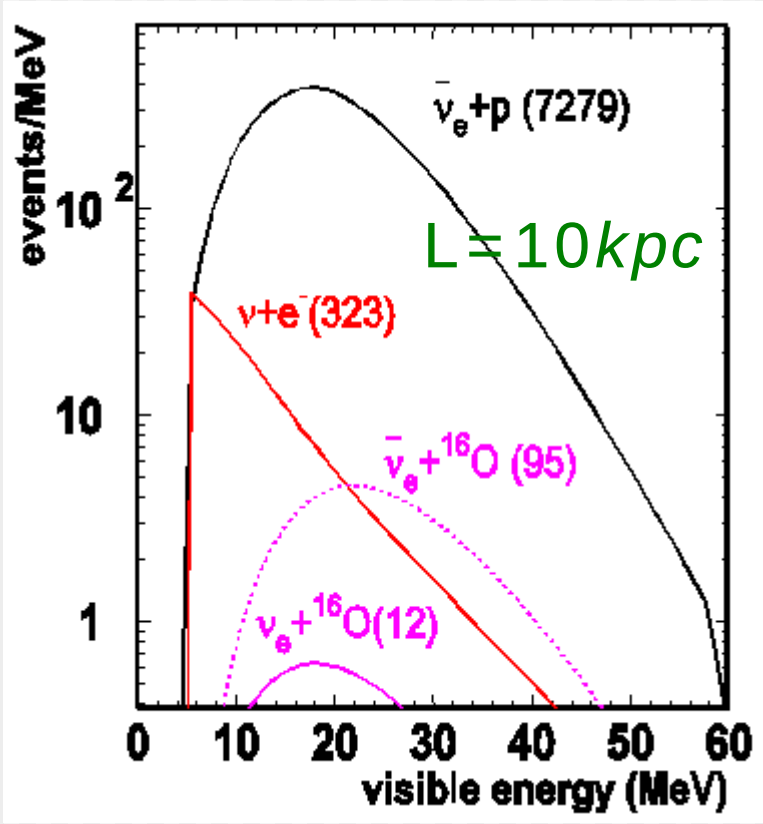






在超级神冈探测器的接受情况

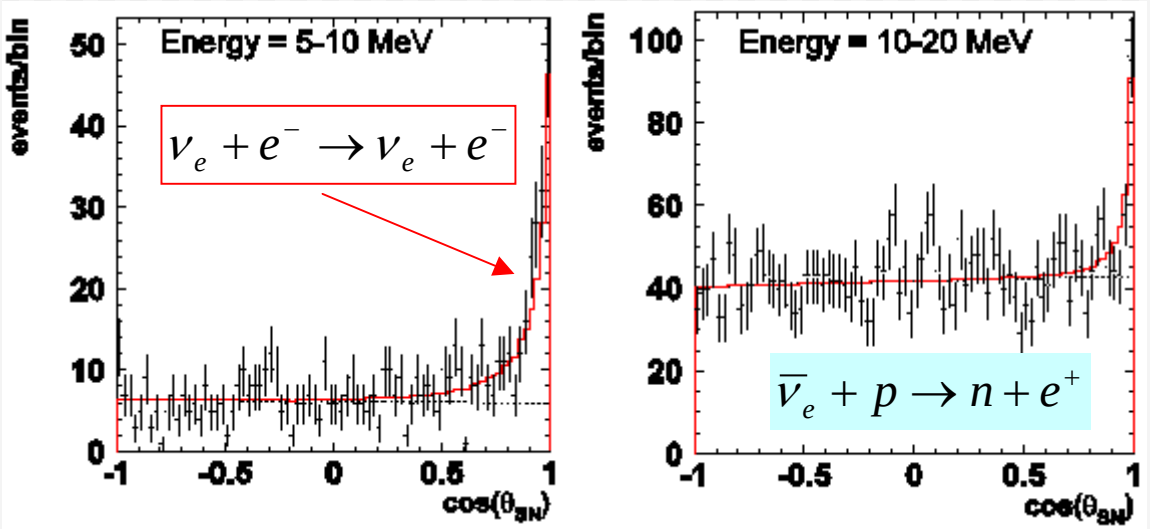


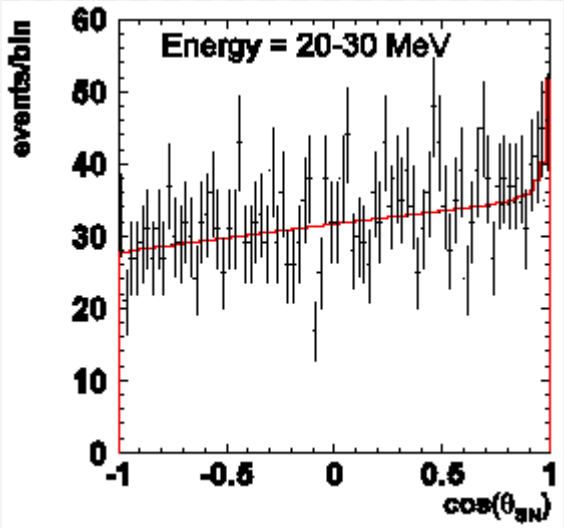


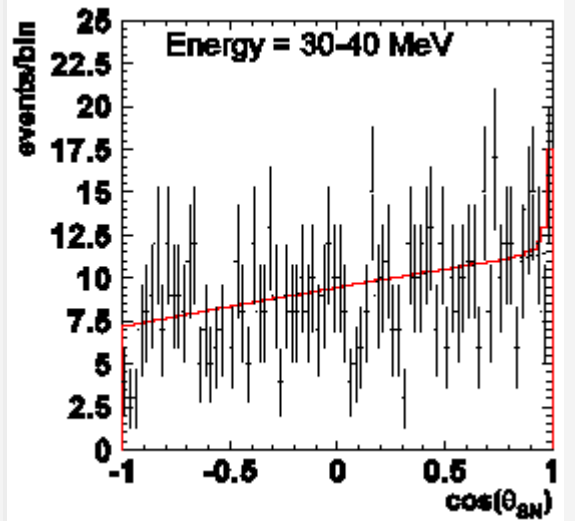
$$\sigma(\nu_e + e^- \to \nu_e + e^-) \cong 0.9\times10^{-43}\left(\frac{E_\nu}{10\text{MeV}}\right)cm^2;\quad \sigma(\bar{\nu}_e + p \to n + e^+) \cong 9.75\times10^{-42}\left(\frac{E_\nu}{10\text{MeV}}\right)cm^2$$

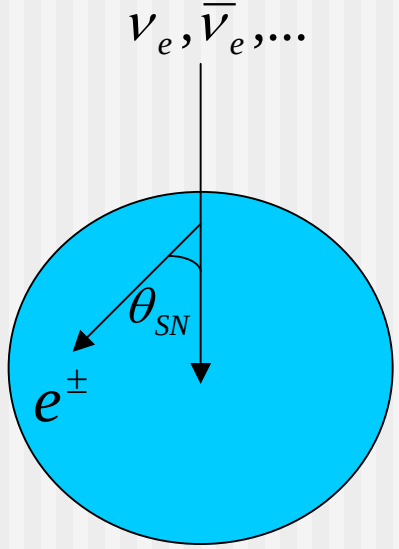
非弹性散射比弹性散射大两个量级。

反 β 衰变过程的影响







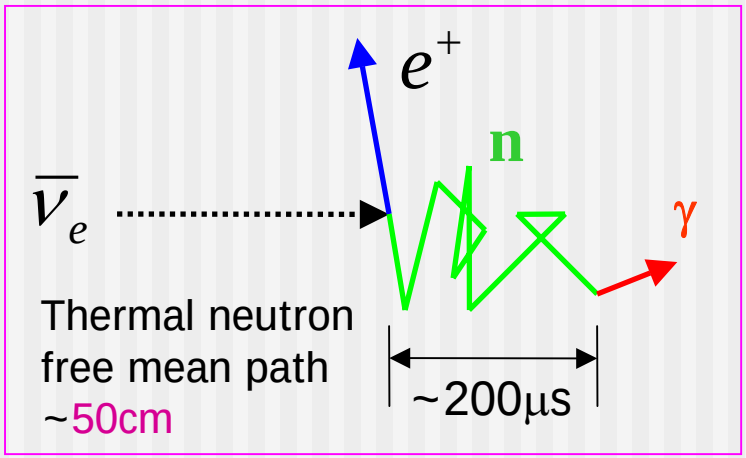


当能量大于**10MeV**时，存在大量非弹散射过程，严重影响**SN**的方向及最初到达时刻的确定。

如何探测反 β 衰变过程

在超级神冈中微子探测器上探测反 β 衰变是一项困难的工作，因为

$$\bar{\nu}_e + p \rightarrow n + e^+ \quad \Rightarrow \quad n \rightarrow p \rightarrow d + \gamma$$



- 瞬时产生的 e^+

$$E_{e^+} \approx E_{\bar{\nu}_e} - 1.3\text{MeV}$$

- 延迟产生的 γ，极低的能量2.2MeV，平均约6光电倍增管击中，难以通过探测器的触发阈值。

在鉴别反 β 衰变过程中，主要的本底来自光电倍增管自身噪音产生的“假击中”。 ⇒ 即 $e^{\pm}$ 与“假击中“的偶然符合。

与反 β 衰变过程有关的测量量

定义

信号：$\bar{\nu}_e + p \rightarrow e^+ + n;\ n + p \rightarrow d + \gamma(2.2\text{MeV})$

本底：$e^{\pm}$ +同时产生的几个光电倍增管"假击中"

测量量

$\vec{x} = (x_1, x_2, ..., x_n)$

$x_1 = anisotropy =$ 光电倍增管到作用点与正电子飞行方向的平均张角

$x_2 = dirks =$ 光电倍增管到作用点与正电子飞行方向的平均极角

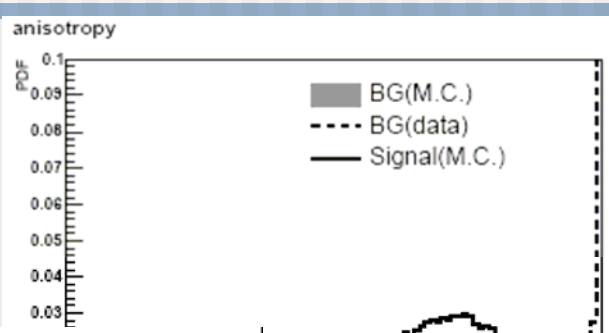
$x_3 = tdiv =$ 光电倍增管时间测量残差的均方根RMS

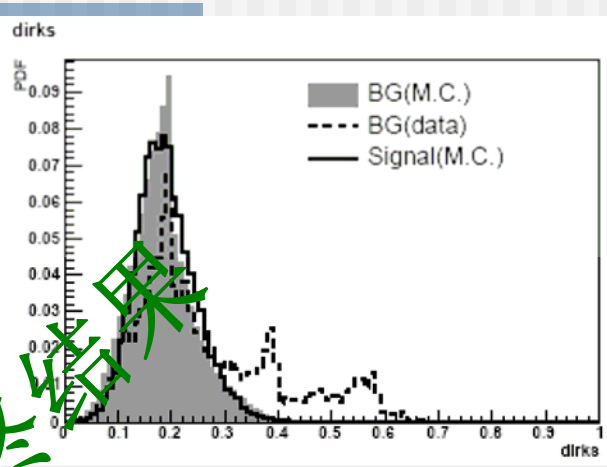
$x_4 = ddiv =$ 光电倍增管到它们几何重心的距离平均值

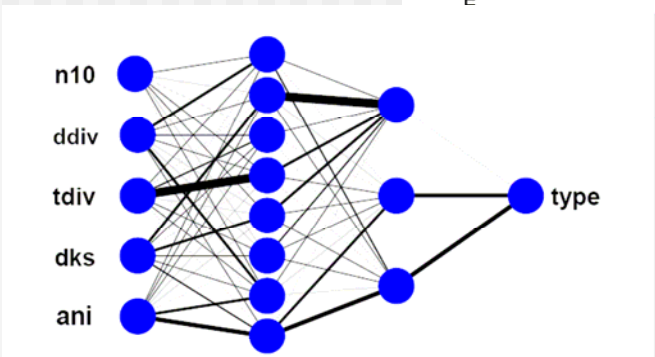
...

神经网络甄别反 β 衰变事例

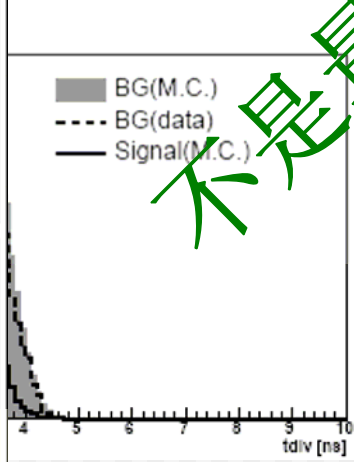
神经网络
输入变量

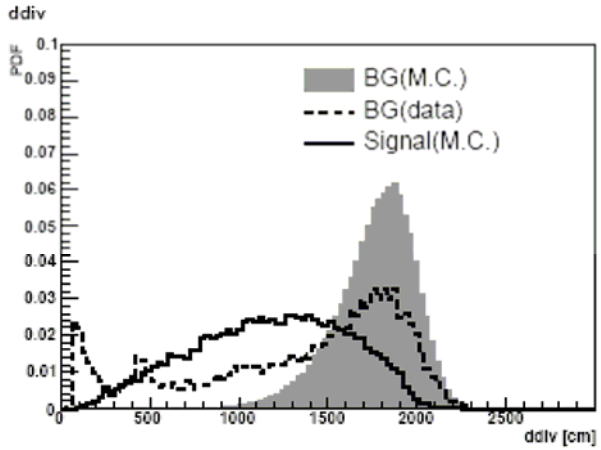


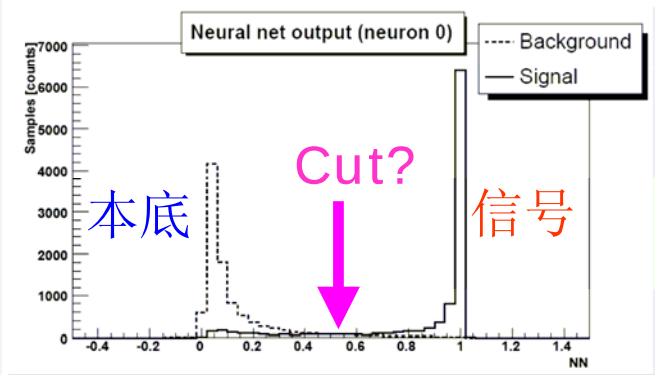




不是最终结果







神经网络
输出量

选择拒绝域使得信噪比与效率最大。

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的相关情形，应只保留一个。

如果输入量对甄别无太大影响，应弃之。

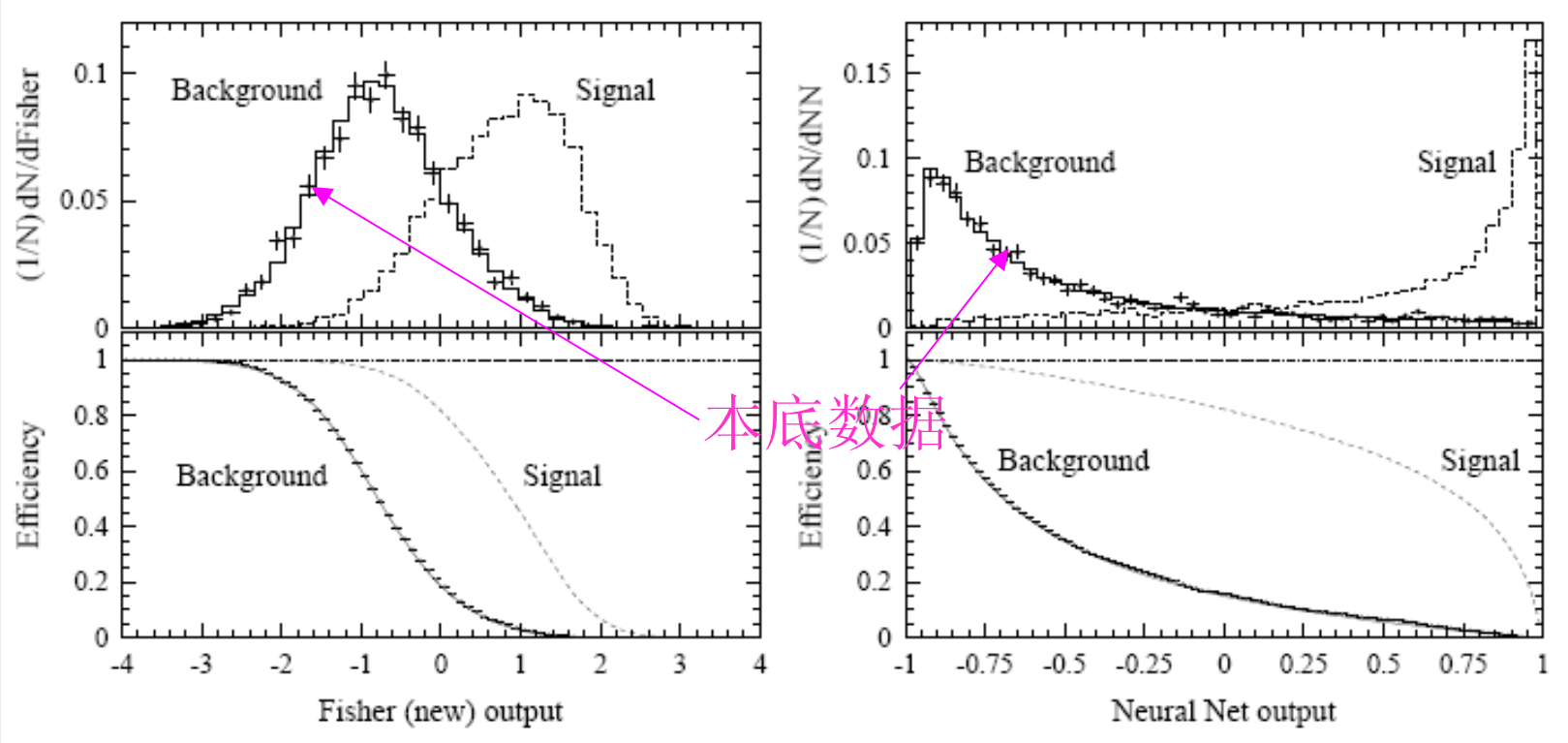
神经网络利用了较高阶矩的联合概率密度函数 $f(\vec{x}\,|\,H)$, 它们也许在训练的样本中找不到较好的模型来描述

→ 最好简化 $t(\vec{x})$, 只要它还能恰当地描述样本。

避免输入量和要研究的信号特征量相关联。

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

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)。

习题

习题3.1:带电粒子横穿一定体积的气体会产生电离,其平均值取决于粒子的类型。假设一基于电离测量的检验统计量 t 经过重建后服从高斯分布。其中心值对电子为 0 ,对$\pi^{\pm}$ 为2,而且标准偏差对于这两种假设均为1。构造一检验统计量使得选择电子时，可通过要求 $t<1$来实现。

a) 该检验的显著水平是多少?(即接受电子的概率是多少)

b) 该检验相对于粒子是$\pi^{\pm}$ 的假设的功效是多少?有多大可能一个$\pi^{\pm}$被当成电子来接收？

c) 假设一粒子样本中已知包含99%的π和1%的电子.当选择t<1时电子的纯度是多少?

(d) 如果要求电子的纯度至少为95%,那么检测的拒绝域(CUT条件)应设何处?在该CUT条件下电子的接收效率是多少?该检验的显著水平是多少?

习题(续)

习题3.2:考虑一检验统计量 t 是基于输入变量 $\vec{x}=(x_1,...,x_n)$ 线性组合,对应的相关系数为 $\vec{a}=(a_1,...,a_n)$

$$t(\vec{x})=\sum_{i=1}^{n}a_i x_i=\vec{a}^T\vec{x}$$

在两种假设H_0与H_1的情况下, 对应的 $\vec{x}$平均值分别为 $\vec{\mu}_0$ 与 $\vec{\mu}_1$,协方差矩阵为V_0与V_1,检验统计量 t 的平均值为τ_0与τ_1,方差为 $\sum_0^2$ $\sum_1^2$

a)证明使粒子分辨达到最大的相关系数

$$J(\vec{a})=\frac{(\tau_0-\tau_1)^2}{\sum_0^2+\sum_1^2}$$

由下式给出

$$\vec{a}\propto W^{-1}(\vec{\mu}_0-\vec{\mu}_1)$$

这里$W=V_0+V_1$。它定义了Fisher线性甄别函数.

习题(续)

b) 假设$V_0=V_1=V$,而且对于输入变量 $f(\boldsymbol{x}|H_0)$与$f(\boldsymbol{x}|H_1)$的p.d.f.是中心在μ_0与μ_1的多维高斯分布.令两种假设的验前概率分别为π_0与π_1.利用贝叶斯理论,找出随 t 变化的验后概率 $P(H_0|\boldsymbol{x})$与$P(H_1|\boldsymbol{x})$.

c)证明为了将检验统计推广到一般而引入偏置

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

其经验概率$P(H_0|\boldsymbol{x})$可以表示为

$$P(H_0\mid x)=\frac{1}{1+e^{-t}}$$

这里,偏置由下式给出

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