

判别分析法用于 ALEPH 实验的 π - μ 鉴别

谢一刚 金山 黄秀萍

(中国科学院高能物理研究所,北京)

摘 要

利用多元统计的判别分析方法,对 ALEPH 实验的强子量能器模型以及实际探测器的强子量能器和 μ 子探测器进行了 π - μ 粒子鉴别的研究。

一、引 言

ALEPH 实验是目前世界上能量最高的正负电子对撞机 LEP 上的四个大型实验之一^[1]。LEP 已于 1989 年 8 月正式运行,并已取得一定的数据和最新的物理成果^[2]。 μ 子的鉴别和强子量能器的参数分析是物理分析的重要工具。近三年来, ALEPH 合作组曾在 CERN 的 π 实验束流上进行了强子量能器及 μ 子探测器模型实验^[3],并试用了几种使 π - μ 鉴别及强子量能器分析优化的方法,其中也选用了多元统计中的判别分析法 (Discriminant Analysis)^[4],这是第一次将这种方法用于高能物理实验分析的尝试。

数理统计学科中的多元统计方法^[5],近年来在社会科学和自然科学等方面得到了愈来愈广泛的运用,并已发展成为一些系统的软件包,例如 SPSS^[6] 就是其中的一种。本文利用这一软件包,对 ALEPH 强子量能器模型(如图 1 所示)中的 π - μ 鉴别问题进行了判别分析,并比较了直接判别法和逐步判别法的结果。另外,从目前已投入运行的 ALEPH 探测器(如图 2 所示)的蒙特卡洛模拟软件包 GALEPH 的有关数据库中得出 π - μ 粒子在强子量能器和 μ 探测器中的击中信息,并将它们换算成多元变量后进行判别分析。

二、判别分析法简单原理^[7,8]

判别分析法是对大量样本按其多种特征进行分类的一种统计方法,也称为分类法。当找不出任何单一特征能明显地将样本分类时,这种方法特别有用。

N 个样本(事例)按其 P 个特征变量 Z_p 划分为 G 个类(群)。为此需按特定的判别条件建立 I 个判别函数 D_i ,以使这 G 个类相互分离得最明显。

$$D_i = d_{i1}Z_1 + d_{i2}Z_2 + \cdots + d_{iP}Z_P,$$

其中 $i = 1, 2, \cdots, I$, 且 $I = G - 1$ 。从几何观点看,任何一个事例相当于 P 维空间的一个代表点,而在 I 个判别函数组成的 I 维新的几何空间中,由这 N 个事例分成的 G 个类

(群)可以分离得最远。将每一个样本(事例)的 P 个特征变量代入判别函数后,就得到一套(共 I 个) D_i 值,称为判别函数的记分,各类(群)的平均记分 \bar{D}_i 称为重心。标准判别函数的系数 d_i 称为权重,它表示第 i 个变量 Z_i 对判别分离程度的贡献。判别函数的确定是一个复杂的理论问题。例如,许多判别法之一的 Fisher 判别法的基本思想是:两群事例(参照图 4)一般具有高斯分布形式,第一群各个特征量的平均值为 $\mu_{11}, \mu_{12}, \dots, \mu_{1P}$, 第二群各特征量的平均值为 $\mu_{21}, \mu_{22}, \dots, \mu_{2P}$, 两群间的平均差值即为 $\sum_{j=1}^P d_j(\mu_{1j} - \mu_{2j})$ 。

而群分离程度最大的条件,可由求群间平均差值平方和各群内事例代表点与各自重心的方差之比最大而得到。如图 4 中第一群重心位置为 $\alpha/2 = +3.8$, 第二群重心为 $\alpha/2 = -3.8$, 这两群间的距离 α 称为马氏 (Mahalanobis) 距离,它比一般方差型的欧几里德几何距离多一个逆协方差矩阵。因为 P 个特征量的量纲各不相同,为了消除量纲差异对判别能力的影响,在多元统计中常采用马氏距离。

由判别函数可进一步导出 G 个区分函数 G_i :

$$G_i = C_{i0} + C_{i1}Z_1 + C_{i2}Z_2 + \dots + C_{iP}Z_P,$$

其中 C_{ij} 称为区分系数, G_i 表示事例属于第 i 群的几率。显然事例被划分到 G_i 最大的那个群内。

可以用直接判别法和逐步判别法进行分类。直接法是一次处理全部变量的方法。逐步法是将变量逐步输入,首先挑选出对判别贡献最大的变量,然后逐步增加变量直至对判别没有明显改进为止。由此可以筛选变量得到比直接法更少的一套变量。根据不同的选择标准有 Wilks, Mahal, Rao 等五种方法。本文对其中的几种方法与直接法的结果进行了比较。

三、ALEPH 强子量能器模型中的 π - μ 鉴别

ALEPH 强子量能器及 μ 子探测器在 CERN π 和 μ 实验束流上的模型如图 1 所示。

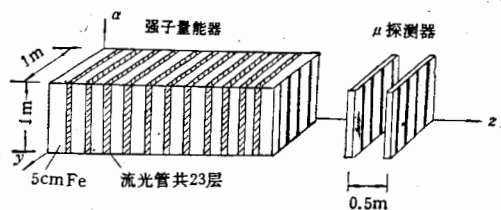


图 1 ALEPH 强子量能器及 μ 探测器实验束模型

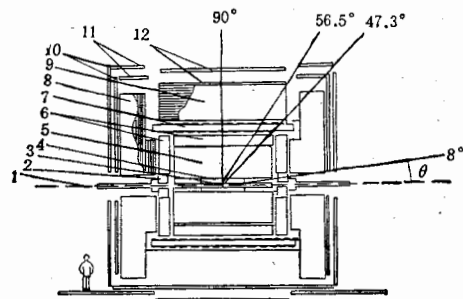


图 2 ALEPH 总体图

1. 束流线; 2. 亮度监测器; 3. 顶点探测器;
4. 内径迹室; 5. 时间投影室; 6. 电磁量能器;
7. 超导磁体; 8. 强子量能器端盖; 9. 强子量能器桶部;
10. μ 探测器端盖; 11. μ -中间角;
12. μ 探测器桶部

模型的 X, Y 方向截面为 1m^2 。沿 z 方向的布局与实际的 ALEPH 探测器(图 2)的强子量能器及 μ 子探测器在极角 $\theta = 90^\circ$ 沿径向截出的断面为 1m^2 的方柱体完全一致。强子量能器模型由 23 层塑料流光管和 23 层 5cm 厚的铁板相间装配而成。在其后方为两层 μ 子探测器, 层间相距 50cm, 每层探测器由双层塑料流光管组成。利用蒙特卡罗模拟程序包 GEANT3, 对上述的探测器几何分别作了 500 个能量 20GeV 的 π 和 μ 事例的蒙特卡罗模拟。因为 ALEPH 实验初期运行阶段未用到 μ 子探测器, 故对只用强子量能器鉴别 μ 子是很感兴趣的。为此我们只选用强子量能器中各层塑料流光管外感应条上一维击中的位置信息。由各层的击中信息换算成 8 种多元变量, 它们是:

1) *DENS*: 每层中击中的平均密度, 定义为

$$DENS = \sum_i^I (N_i/W_i)/I,$$

其中, N_i : 第 i 层中的击中数。

W_i : 第 i 层的簇射宽度, 即有击中的感应条间的最大距离。

I : 击中的层数。

2) *NUMP*: 有击中的平面数。

3) *LASP*: 有击中的最后一层序号。

4) *MAXW*: 最大簇射宽度, 即 W_i 。

5) *CENT*: 重心, 即击中数加权平均的中心, 定义为

$$CENT = \left(\sum_i^I iN_i \right) / \sum N_i.$$

6) *HIT1*: 1—7 层上的总击中数。

7) *HIT2*: 8—14 层上的总击中数。

8) *HIT3*: 15—23 层上的总击中数。

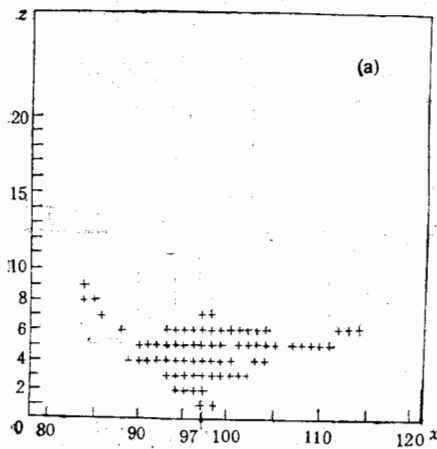


图 3 (a) 强子量能器模型中 π 的击中分布
箭头所指为粒子入射位置

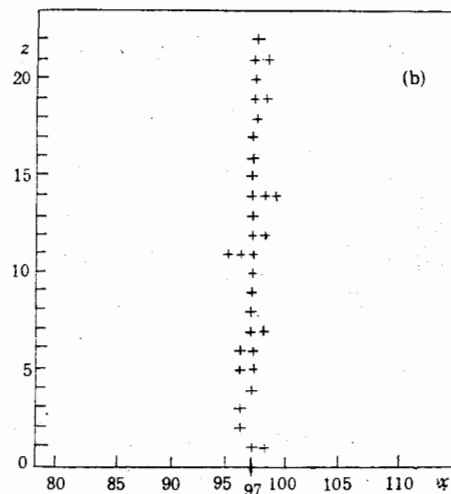


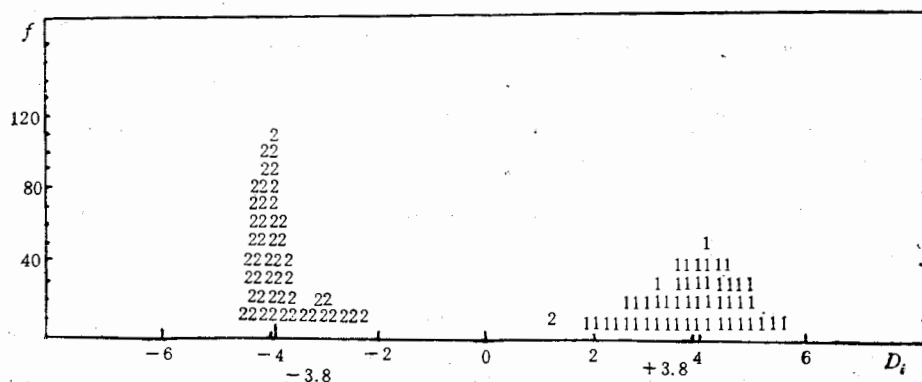
图 3 (b) 强子量能器中 μ 的击中分布
箭头所指为粒子入射位置

表 1 π, μ 在强子量能器中的多元变量平均值及标准和作标准判别函数的系数

| 变量 Z_i | d_{ij} (标准) | d_{ij} (非标准) | Z_i 平均值 | |
|----------|---------------|----------------|-----------|----------|
| | | | π | μ |
| DENS | -0.05980 | -0.99291 | 0.83471 | 0.99961 |
| NUMP | -0.27378 | -0.05510 | 4.32000 | 21.23200 |
| LASP | -0.93160 | -0.29610 | 13.99800 | 23.00000 |
| MAXW | 0.08928 | 0.01156 | 29.11135 | 3.52600 |
| CENT | -0.80910 | -0.36105 | 7.78448 | 11.97742 |
| HIT1 | 0.50078 | 0.02647 | 49.60526 | 10.86800 |
| HIT2 | 1.26874 | 0.06814 | 34.86666 | 10.95800 |
| HIT3 | 1.62738 | 0.12729 | 8.58717 | 13.87600 |
| 常数 | 0 | 6.62296 | | |

入射的 π 和 μ 在强子量能器 23 层感应条上一次事例的击中分布如图 3(a)、(b) 所示。用 20GeV 的 π 和 μ 各 500 个事例得到上述 8 个变量的平均值见表 1。参照图 3 和表 1 中各变量的数据很明显地反映出 π 和 μ 的区别及其各自的特征。例如, μ 的穿透能力强, 其最后作用平面平均值几乎为最后一层, 而 π 的最后作用平面仅为总层数的 1/3 左右。 π 有很大的簇射宽度, 而 μ 一般最多只有 3 个相邻的击中点。 μ 的重心几乎总是在中部, 且三个区域的击中均匀分布, 而 π 的簇射集中在前部。

表 1 中并给出了用直接法得到的标准判别函数的系数, 它反映各量在判别函数中的重要程度, 由于 π 在量能器中后部没有击中分布, 因此 HIT3、HIT2 和 LASP 是最重要的。表 1 中给出的特征变量 Z_i 值因量纲不同(没有标准化), 需与另一套非标准判别函数系数相乘并加一常数后, 才能得到与由标准判别函数得到的同样记分。从标准判别函数系数换算成非标准判别函数系数全部在 SPSS 程序包内进行, 其换算是非常复杂的, 特征变量通常给出的都是非标准的, 它们的标准化过程也是非常复杂的。图 4 给出用直接法得到的两个群的分佈, 两群只对应一个判别函数。横坐标表示判别函数记分 D_1 ; 纵坐标 f 表示两群代表点记分出现的频数。图中并给出两群的重心 $\pm\alpha/2$ 。判别阈在图中原点, 即 $D_1 > 0$ 为 π , $D_1 < 0$ 为 μ 。在同一侧的事例为判中事例, 而在另一侧的事例为误判事例。由区分函数得到的判中几率及误判几率如表 2 所示。由表可见第一群 (π) 误判为第二群

图 4 事例按判别函数记分 D_1 的分佈

(μ) 的几率为 1.2%, 而 μ 误判为 π 的几率为 0.2%。从物理角度考虑, 了解 π 误判为 μ 比 μ 误判为 π 重要。

表 2 由区分函数得到的判别几率

| 实际群 | 事例数 | 预期事例 | |
|--------------|-----|--------------|--------------|
| | | 群 1(π) | 群 2(μ) |
| 群 1(π) | 500 | 494, 98.8% | 6, 1.2% |
| 群 2(μ) | 500 | 1, 0.2% | 499, 99.8% |

利用逐步分析法的 Mahal 及 Rao^[6] 方法也处理了以上问题, 得到的结果是: 用这两种方法得到的标准判别函数的系数与直接法的系数完全一样。按照 Mahal 法中逐步使马氏距离最大和 Wilks'A 值最小, 在 Rao 法中逐步使 Rao'V 判据最大的步骤逐一增加判别变量, 最后到对判别没有明显改善为止, 所得到的变量顺序都一样, 且本计算结果并未得到更少的最佳变量数。多元变量的顺序为: MAXW, NUMP, LASP, HIT3, HIT2, CENT, HIT1, DENS。它们的顺序不一定按对判别的贡献大小排列。因为在已选中的变量中, 较早选中的变量可能因其在后来的显著性比较中不能提供更多的附加信息而可能被剔除掉。

四、ALEPH 探测器中的 π - μ 鉴别

利用描写 ALEPH 探测器的蒙特卡罗模拟软件包 GALEPH201, 研究了在实际探测器(图 2)中的 π - μ 鉴别问题。利用一系列自编的程序从有关的 GALEPH 运行结果的动态存贮数据库 (BOS BANK) 中提取信息, 进而换算成与第三节中类似的多元变量, 形成数据文件, 再输入 SPSS 软件包进行判别分析。主要由 HWDI BANK 中提取强子量能器各层击中束团的地址信息, 得出第三节中所述的 1—8 变量。由 HTHT BANK 中提取不同主层中的沉积能量 ENHC。由 MUHT BANK 中得到 μ 探测器有无击中的信息 MUDI (当 MUDI = 1 时为有击中, MUDI = 0 时为无击中)。对能量为 10GeV 的 π 和 μ 各 1000 个事例, 在有 μ 探测器和无 μ 探测器情况下, 用直接法和 Mahal 逐步法分别进行了判别分析。 π - μ 粒子的入射方向取 ALEPH 探测器极角坐标 $\theta = 90^\circ$, 相当于垂直地入射到强子量能器和 μ 探测器的桶部。其标准判别函数 d_{ij} 及误判率结果见表 3 和表 4。关于 π , μ 粒子按不同能量及不同入射角 θ 对判别的响应参见另文^[7]。

由以上各表的结果可以得出以下结论:

- 1) 有 μ 探测器的判中率(预期事例/实际事例)比无 μ 探测器的判中率高。上述第三节中给出的结果与文献[4]中的结果相比也有同样的结论。
- 2) 参数 MUDI 的权重是最大的(有 μ 探测器的情况), 可见 μ 探测器的作用是明显的。
- 3) 在有 MUDI 的情况下其他变量的权重与无 μ 探测器的变量权重很不一样。例如, 对无 μ 探测器的情况, 最后击中层 LASP (或第三节中的 HIT3) 很重要, 而在有 μ

表 3 用直接判别法和逐步判别法在有无 μ 探测器时的标准判别函数

| d_{ij} | 直接法 | | 逐步法 (Mahal) | |
|----------|-------------|-------------|-------------|-------------|
| | 有 μ 探测器 | 无 μ 探测器 | 有 μ 探测器 | 无 μ 探测器 |
| DENS | 0.08136 | 0.01916 | 0.08748 | — |
| NUMP | 0.01268 | 0.43337 | — | 0.43744 |
| LASP | 0.00075 | 0.98906 | — | 0.99137 |
| MAXW | -0.02979 | -0.13677 | -0.03304 | -0.14592 |
| CENT | 0.16445 | -0.14608 | 0.16757 | -0.14437 |
| HIT1 | 0.04864 | -0.36411 | — | -0.37200 |
| HIT2 | 0.09076 | -0.59694 | -0.06985 | -0.60014 |
| HIT3 | 0.05120 | 0.06077 | -0.04077 | 0.05700 |
| ENHC | 0.07636 | -0.12737 | -0.11838 | -0.12276 |
| MUDI | 0.95451 | — | 0.95437 | — |

表 4 用直接法和逐步法在有无 μ 探测器时的 π, μ 判中率和误判率

| 预期事例 | | 直接法 | | | | 逐步法 (Mahal) | | | |
|------------|------|-------------|-------|-------------|-------|-------------|-------|-------------|-------|
| | | 有 μ 探测器 | | 无 μ 探测器 | | 有 μ 探测器 | | 无 μ 探测器 | |
| 群号 | 事例数 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 1(π) | 1000 | 977 | 23 | 958 | 42 | 977 | 23 | 958 | 42 |
| | | 97.7% | 2.3% | 95.8% | 4.2% | 97.7% | 2.3% | 95.8% | 4.2% |
| 2(μ) | 1000 | 8 | 992 | 39 | 961 | 8 | 992 | 39 | 961 |
| | | 0.8% | 99.2% | 3.9% | 96.1% | 0.8% | 99.2% | 3.9% | 96.1% |

探测器的情况下则降到很次要的地位。

4) 直接判别法和逐步判别法的标准判别函数系数 d_{ij} 具有非常接近的值, 而且两种方法的误判率也是一致的。

5) 逐步法的最佳变量数可能比直接法的少。例如表 3 中的有 μ 和无 μ 探测器的情况分别比直接法的变量少 2 个和 1 个。

6) 在 Mahal 逐步判别法中, 变量逐次被结合和剔除的顺序是按马氏距离增大的顺

序排列的,而且某些先结合的变量以后反而被剔除.例如,在有 μ 探测器的情况下,各变量被结合或剔除(用括号表示)的顺序依次为 *MUDI*, *HIT1*, *DENS*, *CENT*, *HIT2*, *ENHC*, (*HIT1*), *HIT3*, *MAXW*. 最后剩表3中的7个变量.

五、结 束 语

除了用判别分析对 π - μ 鉴别进行研究以外,作者亦曾用判别分析中的主分量(PCA)法在文[8]的基础上研究了用 ALEPH 的强子量能器模型侧向流光管感应条上的击中分布模拟强子簇射的产生^[9]. 在本工作进行期间,高能实验物理领域又有一些工作采用了判别分析法.例如,ALEPH 合作的强子量能器组编写了专门的户子程序包 MUCALO 以便在物理分析的事例重建程序 JIJLIA 中进行 π - μ 鉴别^[10]. 也有人利用判别分析法改善 ALEPH 强子量能器的能量分辨率和在电磁量能器中进行 π - γ 鉴别^[11].

参 考 文 献

- [1] 吴为民,张家铨,高能物理,3(1984),37.
- [2] D. Decamp et al., *Phys. Lett.*, B231, 4(1989), 519; D. Decamp et al., [*Phys. Lett.*, B1, 2(1990), 209; 谢一刚,“ALEPH 国际合作及其最新结果”,物理(1990.5),待发表;谢一刚,高原宁,第五届全国粒子理论会议文集(1989.10).
- [3] M. G. Cantanesi et al., *Nucl. Inst. Meth.*, A247(1986), 438.
- [4] R. Baldini et al., *Nucl. Inst. Meth.*, A251(1986), 449.
- [5] 张尧庭、方开泰著,多元统计分析引论,科学出版社(1982);
M. 肯德尔著,多元分析,科学出版社,(1983);
C. R. 劳著线性统计推断及其应用,科学出版社(1987);
Narayan C. Giri, *Multivariate statistical Interference*;
Paul E.Green, *Mathematical Tools for Applied Multivariate Analysis*;
R. Gnanadesikan, *Methods for statistical Data Analysis of Multivariate Observations*.
- [6] SPSS (McGraw-Hill, New York, 1975)
- [7] 金山,上海复旦大学硕士论文(1990.2).
- [8] M. De Palma et al., *Nucl. Inst. Meth.*, A235(1985), 70.
- [9] 谢一刚,第三届高能实验物理会议论文集,(1988年12月西安).
- [10] G. Bagliesi et al., ALEPH Note 89-123 (1989).
- [11] G. Kellner et al., ALEPH Note 88-100 (1988).

π - μ IDENTIFICATION AND HADRON CALORIMETER ANALYSIS OF ALEPH EXPERIMENT WITH MULTIVARIATER STATISTICS METHOD

XIE YIGANG JIN SHAN HUANG XIUPING

(Institute of High Energy Physics, Academia Sinica, Beijing)

ABSTRACT

In this paper π - μ identification of the hadron calorimeter prototype and the real hadron calorimeter with muon detector of ALEPH experiment on LEP is investigated by using discriminant analysis method of multivariate statistics.