



(八) A.I.D. 方法的一般理论

张 强

第七讲的方法，如果不是用极差进行分割，而是用变差进行分割，就成为一般书上所说的A.I.D.方法（筛选因子的一个新方法）。实际上，A.I.D.方法仅是最优分割法的一种灵活运用。为了能比较清楚地说明这一点，我们就依照一般书上的写法来介绍A.I.D.方法，然后，再说明它与最优分割法的关系。

本讲着重分析介绍一些概念和推理，就不再举例了，读者可将上一讲的实例移植过来，用以说明这一讲的方法。

设有一个因子 x 和一个预报量 y ，它们相应的观测资料为 $x_1, x_2, x_3, \dots, x_n; y_1, y_2, y_3, \dots, y_n$ ；

$\{(x_i, y_i) \mid i=1, 2, \dots, n\}$ 。考虑对 x 这个因子所有可能的分割，任给一个常数 c ，可以把 n 年的资料 (x_i, y_i) 分成两组：

$$I(c) = \{(x_i, y_i) : x_i \geq c\}$$

$$II(c) = \{(x_i, y_i) : x_i < c\}$$

即按因子 x 的值 $x_i \geq c$ ，还是 $x_i < c$ ，分成两组。对每一组计算 y 各自的均值和变差：即计算

或
 $\bar{y}(I(c)) = \bar{y}(I) = \frac{1}{n_1} \sum_{x_i \geq c} y_i$, (第 I 组 y 的均值);

或
 $\bar{y}(II(c)) = \bar{y}(II) = \frac{1}{n_2} \sum_{x_i < c} y_i$, (第 II 组 y 的均值);

n_1, n_2 分别为 $I(c), II(c)$ 资料的个数，例如 n 年的资料 $(x_i, y_i), i=1, 2, \dots, n$ 中有 n_1 个 $x_i \geq c$, n_2 个 $x_i < c$, n_1, n_2 就是这两个数，因此自然有 $n_1 + n_2 = n$ 。

或
 $s^2(I(c)) = S(I) = \sum_{x_i \geq c} (y_i - \bar{y}(I))^2$, (第 I 组 y 的变差)

或
 $s^2(II(c)) = S(II) = \sum_{x_i < c} (y_i - \bar{y}(II))^2$, (第 II 组 y 的变差)

用 $S(C)$ 表示相应于以 C 作为临界值分组后两组的总变差，即 $S(C) = S(I) + S(II)$ (或更仔细地 $= S(I(C)) + S(II(C))$)。

很容易看出，每给定一个 C ，依上述方法就可以算得一个 $S(C)$, $S(C)$ 的大小反映了依 C 来分组是否合

适的程度。 $S(C)$ 小，表示用 C 来分就好，确实把 y 分成了两个不同的部分，内部的差异比较小； $S(C)$ 大，表示用 C 来分就不好，虽然用 C 把 y 分成了两组，但每一组内部并不整齐，说明这样分的意义不大。由于 $S(C)$ 是两个平方和 $S(I), S(II)$ 加起来的，它永远不是负的，它对 C 而言，总可以选某一 C^* 使 $S(C^*)$ 在全部 $S(C)$ 中达到最小值，即有

$$S(C^*) = \min_{C} S(C).$$

如此的 C^* 就是最好的分组临界值，依它分组后 y 内部的变差总和达到最小，因此就用这个 C^* 来进行分组。

仔细分析一下上面的过程，就可以发现它确实是最优分割法的一个应用。如果把资料依 x 取值的大小排列，即重新排列使

$$\begin{matrix} x_{11} \leq x_{12} \leq x_{13} \leq \dots & \dots & < x_{1n} \\ \downarrow & \downarrow & \downarrow & \downarrow \\ y_{11} & y_{12} & y_{13} & \dots & y_{1n} \end{matrix}$$

为了方便，无妨假定

$$\begin{matrix} x_1 \leq x_2 \leq \dots & \dots & \leq x_n \\ y_1 & y_2 & & & y_n, \end{matrix}$$

于是对任何一个给定的 C ，按 $x_i \geq C$ 还是 $x_i < C$ 分成两组，就是比较 x_i 与 C 的大小。如果 $x_i \geq C$ ，而 $x_{i+1} < C$ ，于是分组成 $(y_1, \dots, y_{i-1}) (y_i, y_{i+1}, \dots, y_n)$ 它就是 y_1, \dots, y_n 的一个 2 分割，所有各种不同的 C 所对应的分组就是全部可能的 2 分割（顺序按因子 x 的大小来排列）。由此可见， $S(C)$ 实际上就是 C 作为临界值给出的 2 分割所相应的总变差，所以 C^* 使 $S(C)$ 达到最小值就是 C^* 给出的分割就是最优 2 分割。这样，就弄清楚了 A.I.D. 方法的基本思想与最优分割法的关系：A.I.D. 方法是依因子 x 的取值大小的顺序考察对预报量 y 的最优 2 分割。

和变差最优分割法一样，我们可以进一步问：虽然 C^* 给出了 y 的最优 2 分割，但这样分割成两组是否真有意义？也即问这两组是否存在显著性差异？此时可以作一个 F 检验，即第三讲中介绍过的方法。如果 F 检验认为是显著的，即这样分割有意义，否则就无需分割成两组（注意尽管 $S(C^*)$ 达到最小，它也没有意义）。利用 F 检验，就可以进行因子的逐步筛选。

设有 K 个因子 x_1, x_2, \dots, x_k ，预报量是 y ，于是全部的资料就是下述矩阵：

$$\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_k \\ y \end{array} \left(\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ x_{31} & x_{32} & \cdots & x_{3n} \\ \vdots & \vdots & & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \\ y_1 & y_2 & \cdots & y_n \end{array} \right)$$

1. 第一步，对每一个因子 x_i ，依 x_i 的大小顺序，把 y 的资料排成有顺序的 n 个资料，考察它的最优 2 分割，最优 2 分割相应的总变差记为 S_i ，比较 S_1, S_2, \dots, S_K ，找出一个最小的，设为 S_{11} ，即 $S_{11} = \min S_i, 1 \leq i \leq K$

这就是说依 x_{11} 的大小顺序来对 y 作最优 2 分割，可以使分割后两组内部的变差总和达到最小。然后，对 S_{11} 所对应的 2 分割作一次 F 检验，检验分割后两组是否有显著差异。如有， x_{11} 就入选，就转入第二步；如没有显著差异，就表明这样分割是没有意义的，筛选因子和分割的步骤就终止，这是因为 S_{11} 已经达到最小值了，最小值相应的分割都没有什么意义，因此，就无法再选出可以用来预报的因子，筛选就此停止。

2. 第二步， x_{11} 入选后，就依 x_{11} 的顺序考虑最优 2 分割，将全部资料分成两组。每一组都单独重新处理，即回到第一步。

这样，不断地入选因子，不断地将资料分组，直到再也无法入选因子，无法分组为止，于是全部资料就分成了若干组，就可以象上一讲所举的例子那样，画出一张分组的图，按照这张图就可以作预报。

现在，我们从理论上来探讨一下，这样的方法有什么优缺点，如何进一步克服它的一些缺点。在讨论之前，首先说明，凡是在上一讲中说过的问题这里就不再重复了。

1. A. I. D. 方法为什么要用因子的大小顺序来给预报量排一个顺序，然后考虑最优 2 分割？如果因子与预报量之间关系是比较密切的，那么在它们的数量之间一定有些联系。例如因子值越大，预报量的值也越大即通常所说的正相关；或因子值越大，预报量的值就越小即通常所说的负相关。然而，这样明显的规律是不多见的，比较多的是，因子值的大小与预报量之间有一种趋势的联系，但在具体数字上可以出入很大，例如上一讲的例子，“冬雷”和“干秋”的关系。因为世界上事物之间的联系是复杂的，一个预报量往往并不能被一、二个因子的取值所决定，而且预报量与因子之间的关系也不是始终是一样的。这样用因子

的次序来对预报量进行分割就有下列的优点：

如果因子与预报量之间确有比较密切的正相关或负相关，那进行 2 分割之后，总变差确实会变小很多，A. I. D. 方法能反映这种关系。

如果因子与预报量之间正、负相关不明显，而是分段之后才有可能反映出来，我们遇到过这样的例子，将预报量 y 依因子 x 的大小顺序进行最优 2 分割后，发现 $x < C$ 的一组内， x 与 y 是正相关， $x \geq C$ 的一组内， x 与 y 是负相关，说明 C 确实是一个分界点，这两组内 x 与 y 的联系规律是不同的。如何发现这样一些规律，这是过去熟悉的一些预报方法不容易达到的，而 A. I. D. 方法能比较好地反映这一点，特别当资料较长时，可以不只是考虑 2 分割，还可以考虑 3 分割、4 分割、…等手段来发现分组后因子与预报量之间的关系。

A. I. D. 方法的缺点，是将资料分组到最后，有可能每一组内的变差还很大，或者有些组内历史资料非常少，这样预报时，报对的把握就不大。而且 A. I. D. 方法将全部历史资料都分组了，因此预报的结果必然是和历史资料中出现过的某些年相似，它不能自动判别今年所遇到的情况是否是历史资料所不能反映的情况，亦即不能判别今年是不是一个异常年（相对于历史资料来说的异常！）。

2. 进一步考虑就会发现，既然要依 x 来分组由 F 检验来决定，那在考察分割时，不是用总的变差最小，而是用 F 的值来衡量，F 越大就越好，F 值最大的就是最优 2 分割，这不是更好吗？这个想法是合理的，主要问题涉及到两点：一是由于 F 值对分割而言并没有可加性，对 2 分割考虑它的最优分割还不困难，对 3 以上的分割就复杂了；二是 F 值的计算稍麻烦一些，但在计算机上是不困难的。因此，在条件许可时，是可以用 F 值的大小来衡量分割的好坏，而 A. I. D. 方法每一步都是只考虑最优 2 分割，用 F 值是完全可行的。

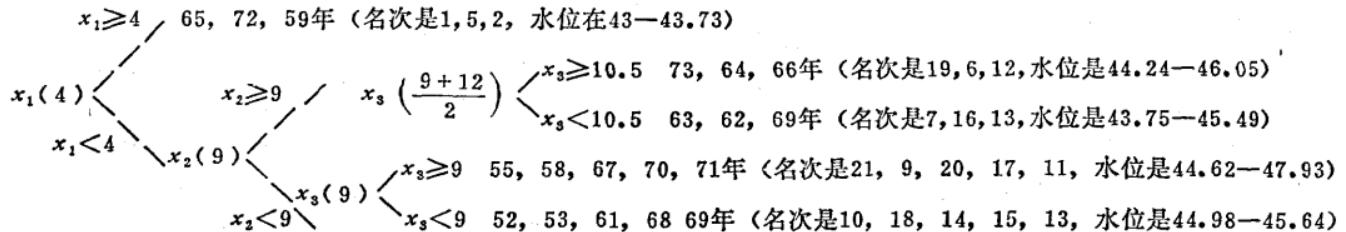
3. A. I. D. 方法并不需要因子的具体数值，而只需要在资料中对因子的值给出一个大小的顺序（由小到大或由大到小都可以），实际上只要它的顺序号，如果对预报量 y 不考虑它的具体的值而只考虑它的顺序号，把 A. I. D. 方法用到这样一类的数据上去，就得到一种预报的方法，下面用一个数字的例子来说明。

用弋阳前一年 11 月至当年 1 月的降水总量 x_1 ，前一年 10 月至当年 1 月最大一次降水量 x_2 ，当年 1 月的平均气温 x_3 ，来预报当年的最高水位 y ，历史资料见附表：略去计算的过程（采用极差分割，2 分割），读者可以作为练习去计算一下，最后得下面一张附图：

附表 1952—1973 年的资料

| 年份 | 1952 | 1953 | 1954 | 1955 | 1956 | 1957 | 1958 | 1959 | 1960 | 1961 | 1962 | 1963 | 1964 | 1965 | 1966 | 1967 | 1968 | 1969 | 1970 | 1971 | 1972 | 1973 |
|-----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| x_1 的名次 | 15 | 9 | — | 11 | 5 | 7 | 10 | 3 | 13 | 9 | 8 | 4 | 16 | 1 | 18 | 9 | 6 | 16 | 12 | 14 | 2 | 17 |
| x_2 的名次 | 12 | 18 | — | 9 | 13 | 2 | 17 | 5 | 8 | 16 | 6 | 4 | 1 | 3 | 5 | 15 | 11 | 19 | 10 | 11 | 14 | 7 |
| x_3 的名次 | 9 | 9 | — | 1 | 3 | 15 | 8 | 6 | 9 | 10 | 4 | 2 | 13 | 16 | 15 | 7 | 10 | 11 | 3 | 5 | 14 | 12 |
| y 的名次 | 10 | 18 | — | 21 | 8 | 4 | 9 | 2 | 3 | 14 | 16 | 7 | 6 | 1 | 12 | 20 | 15 | 13 | 17 | 11 | 5 | 19 |

名次是指资料由小到大排列时相应的顺序号，以后就用这个术语“名次”来说了。



附图

从图就可以看到，依 x_1 ， x_2 这两个因子 $x_1 < 4$ ， $x_2 < 9$ 就可以将大部分水位较高的年份分离出来，而 $x_1 \geq 4$ 或 $x_2 \geq 9$ ， $x_1 < 4$ 的那些年份即使用 x_3 也分不好，说明在 $x_1 \geq 4$ 或 $x_2 \geq 9$ ， $x_1 < 4$ 时，只用这三个因子是确定不了的，还需要考察其他有关的因子，才能改善预报的结果。

讨论变量“名次”之间的关系，有所谓“秩相关法 (Rank Correlation Methods)”，它与回归有关，将在今后介绍，请读者注意和现在这样的做法来进行比较。

从上面的讨论不难看出，如果我们对预报量只考虑它的级别（比名次还要粗一些来划分预报量的值），上述的方法还可以简化，而且对预报的效果更容易考察，这些就不一一细说了，请读者自己去进行分析。

4. A.I.D. 方法告诉我们，在考虑预报问题时，

用距平来反映因子的情况并不总是合理。在气象上常常常用距平来反映一个气象要素的历史状况，距平的正负号就把历史资料分成了两大类，实际上是一个以均值为临界值的一个2分割，这个2分割从预报 y 来看是不是最合适的方法呢？不一定，A.I.D. 法就是在所有可能的全部临界值 C 中去寻找对预报 y 来说最好的一个临界值 C^* ， C^* 往往不一定就是因子 x 的均值。这就告诉我们，当我们研究事物之间的关系时，绝不能孤立地从各自的情况去分析考察，而必须在相互的关系中去分析考察，A.I.D. 方法的优越之处，也就是在于这一点。而这样一个基本思想，却可以引导我们来改进过去熟悉的一些图上操作的方法。在下一讲我们将运用这个思想来介绍一种不用手算，直接在点聚图上进行划分的预报方法，称为A.I.D. 的图上作业法。