第 26 卷 第 5 期
2004 年 9 月 (523～528)

地 震 学 报
ACTA SEISMOLOGICA SINICA

Vol. 26, No. 5
Sep., 2004

文章编号: 0253-3782(2004)05-0523-06

多元统计组合模型在地震综合预报中的应用*

韩天锡[1)] 蒋 淳[2)] 魏雪丽[1)] 韩 梅[1)] 冯德益[2)]

1) 中国天津 300191 天津理工大学
2) 中国天津 300201 天津市地震局

摘要 针对目前地震综合预报中的一些问题，利用近 30 年来迅速发展的多元统计分析中主成分分析、判别分析组成多元统计组合模型，在众多的地震预报指标(预报因子)中采用信息最大化方法，选择对中期预测信息累积贡献率大于 90%地震预报指标，分别进行相关分析、预测、检验，最终应用马氏距离判别作外推综合预报；并以华北地区(30°～42°N，108°～125°E)为例进行模型的应用检验，初步研究已取得了较好的效果.

关键词 多元统计组合模型 主成分分析 判别分析 地震综合预报

中图分类号: P315.7 **文献标识码**: A

引言

我国地震界在过去 30 多年的预报实践和几次地震预报攻关中，探索了许多新的地震学方法及预报指标. 目前在地震中期和中短期预报中，常使用如反映介质所受应力的状态和介质的均匀程度 b 值(陈锦标，李全林，1989)，表示地震活动是否增强 m_f 值(王炜等，1994)，考虑一个地区的地震活动性、震级和频次各方面的因素，描述各地区地震活动性的定量参数 $A(b)$值(吴佳翼，曹学峰，1983)，地震活动演化指数 Y_H 值(陆远忠等，1994)，地震危险度 D 值(王炜，刘震华，1987)，以及调制小震法 R_m 是利用经常作用于地壳的周期性固体潮对不均匀地壳(指介质的不均匀和应力集中程度的不均匀)的调制作用所诱发的小震活动图象，探测地震活动带或区内地壳介质强弱分布以及寻找高应力集中区(秦保燕，刘江峰，1989)等等. 诸如此类的测震学指标百余种，这是地震学家经数十年的努力探索地震预报途径辛勤劳动结晶. 但是，在多指标的综合预测研究中发现，往往由于指标个数太多，并且彼此之间存在着一定的相关性，不同程度上反映了信息的重叠. 而且当指标较多时，在高维空间中研究样本的分布规律比较麻烦. 这就使人联想到，如果能通过统计方法，对所有测震学指标进行筛选，综合考虑地震活动在中期、短期震兆特征，从携带信息量大的多元地震活动指标来探求未来强震发生前中、短期阶段可能存在的孕震信息，从而进行多元综合判别. 因此，本研究利用统计学中的主成分分析、判别分析方法组成的多元统计组合模型，在众多的测震学指标中，进行信息最大化处理提取预测指标，并使这些预测指

* 中国地震局“十五”科研攻关项目(2001BA601B01-010506)资助.
2003-03-03 收到初稿，2004-02-27 收到修改稿，2004-04-12 决定采用.

标能尽可能多地反映原来的信息量. 通过检验后，确定为多元综合预测变量，建立多元判别模型进行地震综合预测. 这是利用统计学方法各自解决不同问题的优势，进行地震综合预测新方法的有益尝试.

1 多元统计组合模型

多元统计组合模型是由统计学中的主成分分析、判别分析方法组成. 其工作思路为：在众多测震学指标通过主成分分析方法，进行信息最大化处理，选择携带信息量最集中的多元地震活动指标；通过检验后，确定中期地震预报变量集合；利用判别分析方法建模进行学习、检验，外推预测.

1.1 主成分分析方法

主成分分析是一种降维或把多个指标化为少数几个综合指标的一种统计分析方法. 假设有来自某个总体的 n 个样本，而每个样本测得 p 个指标数据. 这 p 个指标之间往往互有影响，需要从 p 个指标中去寻找少数几个综合性的指标，而这几个综合性的指标既能反映原来 p 个指标的信息，又能达到彼此之间互不相关.

设有测震学指标向量 $\boldsymbol{X}=(x_1, x_2, \cdots, x_p)'$，$\boldsymbol{V}\geqslant 0$ 是 $\boldsymbol{X}$ 的协方差矩阵. 若 $\boldsymbol{V}$ 的非零特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$，相应的特征向量 $\boldsymbol{L}_i=(l_{i1}, l_{i2}, \cdots, l_{ip})'$，则 $\boldsymbol{X}$ 的第 i 个主成分为 $y_i=L_i'\boldsymbol{X}=l_{i1}x_1+l_{i2}x_2+\cdots+l_{ip}x_p$，$i=1, 2, \cdots, p$. 其中，第 i 个主成分的信息贡献率是：$\lambda_i/\sum_{i=1}^{p}\lambda_i$；前 r 个主成分的累积贡献率是：$\sum_{i=1}^{r}\lambda_i/\sum_{i=1}^{p}\lambda_i$(方开泰，1989).

我们取测震学指标的信息累积贡献率达到 $\sum_{i=1}^{r}\lambda_i/\sum_{i=1}^{p}\lambda_i\geqslant 85\%$ 以上的主成分个数 r，并在前 r 个主成分中依次选取权最大的若干个预报指标作为提取的地震预报变量集合.

计算步骤：设 $\boldsymbol{X}_i$ 来自总体 $\boldsymbol{X}$ 的样本 $i=1, 2, \cdots, n$.

1) 计算 $\boldsymbol{V}$ 的估计值：$\hat{\boldsymbol{V}}=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})'$.

2) 计算 $\hat{\boldsymbol{V}}$ 的特征值：λ_i 及特征向量 $\boldsymbol{L}_i(i=1, 2, \cdots, p)$.

3) 计算第 i 个主成分：$y_i=L_i'\boldsymbol{X}(i=1, 2, \cdots, p)$.

4) 计算 $\boldsymbol{X}$ 的第 i 个主成分的贡献率：$\lambda_i/\sum_{i=1}^{p}\lambda_i$.

5) 在前 r 个主成分的因子载荷阵中，依次选取权最大的若干个预报因子作为提取的地震预报变量.

6) 应用相关分析对预报变量进行检验，其最大震级 M 和预报变量的相关系数由下式计算可得：

$$r_{M,x_j}=\frac{\sum_{i=1}^{n}(M_i-\overline{M})(x_{ij}-\overline{x_j})}{\sqrt{\sum_{i=1}^{n}(M_i-\overline{M})^2}\sqrt{\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2}}\qquad j=1, 2, \cdots, p \tag{1}$$

根据相关系数的大小，对主成分分析提取的预报变量进行检验，剔除相关系数小的预报变量，引入相关系数大的预报变量，作为马氏距离判别预报模型预报变量集合.

1.2 判别分析方法(Johnson, Wichern, 1992)

判别分析是多元统计分析中用于判别样本所属类型的一种统计分析方法. 它所要解决

的问题是在一些已知研究对象用某种方法已分成若干类的情况下，确定新的观测数据属于已知类别中的哪一种. 判别分析方法处理问题时，通常要给出一个新样品与各已知类别接近程度的描述指标，即判别函数，同时也指定一种判别规则，以判定新样品的归属. 决定样品归属时，只须考虑判别函数值的大小. 本研究选用马氏距离判别模型.

设马氏距离判别预报变量集合为 x_1，x_2，…，x_k（根据主成分分析方法所提取）. 根据历史数据由于地区年度最大震级均为 5 级以上，故设有两个地震级别的总体：G_1 为年度最大震级 5.0～5.9；G_2 为年度最大震级 6.0 以上. 其样本为 $\boldsymbol{X}_i=(x_{i1}, x_{i2}, \cdots, x_{ik})'$，$i=1, 2, \cdots, n$.

定义：$\boldsymbol{X}_i$ 到两类总体 G_1，G_2 的马氏距离分别为

$$\begin{cases} d^2(\boldsymbol{X}_i, G_1) = (\boldsymbol{X}_i-\boldsymbol{\mu}^{(1)})'\boldsymbol{V}_1^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}^{(1)}) \\ d^2(\boldsymbol{X}_i, G_2) = (\boldsymbol{X}_i-\boldsymbol{\mu}^{(2)})'\boldsymbol{V}_2^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}^{(2)}) \end{cases} \tag{2}$$

其中，$\boldsymbol{\mu}^{(1)}$，$\boldsymbol{\mu}^{(2)}$，$\boldsymbol{V}_1$，$\boldsymbol{V}_2$ 分别是 G_1，G_2 两类总体的均值向量和协方差矩阵. 其样本的估计公式为

$$\begin{cases} \hat{\boldsymbol{\mu}}^{(1)} = \overline{\boldsymbol{X}}^{(1)} = \dfrac{1}{n_1}\sum\limits_{i=1}^{n_1}\boldsymbol{X}_i^{(1)} \\ \hat{\boldsymbol{\mu}}^{(2)} = \overline{\boldsymbol{X}}^{(2)} = \dfrac{1}{n_2}\sum\limits_{i=1}^{n_2}\boldsymbol{X}_i^{(2)} \\ \hat{\boldsymbol{V}}_1 = \dfrac{1}{n_1-1}\sum\limits_{i=1}^{n_1}(\boldsymbol{X}_i^{(1)}-\overline{\boldsymbol{X}}^{(1)})(\boldsymbol{X}_i^{(1)}-\overline{\boldsymbol{X}}^{(1)})' \\ \hat{\boldsymbol{V}}_2 = \dfrac{1}{n_2-1}\sum\limits_{i=1}^{n_2}(\boldsymbol{X}_i^{(2)}-\overline{\boldsymbol{X}}^{(2)})(\boldsymbol{X}_i^{(2)}-\overline{\boldsymbol{X}}^{(2)})' \end{cases} \tag{3}$$

其中，n_1，n_2 分别为 G_1，G_2 的样本容量；$\boldsymbol{X}_i^{(1)}(i=1, 2, \cdots, n_1)$ 和 $\boldsymbol{X}_i^{(2)}(i=1, 2, \cdots, n_2)$ 分别为 G_1，G_2 的样本；$\overline{\boldsymbol{X}}^{(1)}$，$\overline{\boldsymbol{X}}^{(2)}$ 分别为 G_1，G_2 样本均值向量.

马氏距离判别函数

$$\begin{aligned} W(\boldsymbol{X}_i) &= d^2(\boldsymbol{X}_i, G_2) - d^2(\boldsymbol{X}_i, G_1) \\ &= (\boldsymbol{X}_i-\hat{\boldsymbol{\mu}}^{(2)})'\hat{\boldsymbol{V}}_2^{-1}(\boldsymbol{X}_i-\hat{\boldsymbol{\mu}}^{(2)}) - (\boldsymbol{X}_i-\hat{\boldsymbol{\mu}}^{(1)})'\hat{\boldsymbol{V}}_1^{-1}(\boldsymbol{X}_i-\hat{\boldsymbol{\mu}}^{(1)}) \end{aligned} \tag{4}$$

判别规则

$$\begin{cases} \boldsymbol{X}_i \in G_1 & \text{当 } W(\boldsymbol{X}_i) > 0 \\ \boldsymbol{X}_i \in G_2 & \text{当 } W(\boldsymbol{X}_i) < 0 \\ \text{待判} & \text{当 } W(\boldsymbol{X}_i) = 0 \end{cases} \tag{5}$$

其计算步骤：设从 G_i 中抽取的样本为 $\boldsymbol{X}_1^{(i)}$，$\boldsymbol{X}_2^{(i)}$，…，$\boldsymbol{X}_{ni}^{(i)}(i=1, 2)$.

1）计算 $\hat{\boldsymbol{\mu}}^{(i)}=\overline{\boldsymbol{X}}^{(i)}=\dfrac{1}{n_i}\sum\limits_{j=1}^{n_i}\boldsymbol{X}_j^{(i)}$　　$(i=1, 2)$ (6)

2）计算 $\hat{\boldsymbol{V}}_i=\dfrac{1}{n_i-1}\sum\limits_{j=1}^{n_i}(\boldsymbol{X}_j^{(i)}-\overline{\boldsymbol{X}}^{(i)})(\boldsymbol{X}_j^{(i)}-\overline{\boldsymbol{X}}^{(i)})'$　　$(i=1, 2)$ (7)

3）对待判样本 $\boldsymbol{X}$ 按式(4)计算马氏距离判别函数值 $W(\boldsymbol{X})$.

4）按照判别规则式(5)对待判样本 $\boldsymbol{X}$ 进行判别.

2 多元统计组合模型在地震综合预报中的应用

该方法应用的基本思路是：对众多的测震学指标及数据作主成分分析，提取携带信息

表 1 华北地区多元统计组合模型学习样本判别结果及检验表

序号	震例			样本原属类别	判别类别	检验(判对为√，判错为×)
	年-月-日	地点	M			
1	1973-12-31	河间东	5.3	G_1	G_1	√
2	1974-04-22	江苏溧阳	5.5	G_1	G_1	√
3	1975-02-04	海城	7.3	G_2	G_1	×
4	1976-07-28	唐山	7.8	G_2	G_2	√
5	1977-05-12	宁河	6.5	G_2	G_2	√
6	1978-05-18	营口	5.9	G_1	G_2	×
7	1979-07-09	溧阳	6.0	G_2	G_2	√
8	1980-02-07	宁河	5.2	G_1	G_1	√
9	1981-11-09	隆尧东	5.8	G_1	G_1	√
10	1982-10-19	卢龙	5.3	G_1	G_1	√
11	1983-11-07	菏泽	6.0	G_2	G_1	×
12	1984-05-21	南黄海	6.2	G_2	G_1	×
13	1985-11-30	任县	5.6	G_1	G_1	√
14	1986-05-23	南黄海	5.1	G_1	G_1	√
15	1987-02-17	射阳东南	5.0	G_1	G_1	√
16	1988-07-23	阳原	5.0	G_1	G_1	√
17	1989-10-19	大同	6.1	G_2	G_1	×
18	1990-02-10	常熟	5.5	G_1	G_1	√
19	1991-03-26	大同	5.8	G_1	G_1	√
20	1992-01-23	黄海	5.3	G_1	G_1	√

表 2 华北地区多元统计组合模型检验样本判别结果及检验表

序号	震例			样本原属类别	判别类别	检验(判对为√，判错为×)
	年-月-日	地点	M			
1	1993-08-29	杭锦后旗	5.1	G_1	G_1	√
2	1994-07-26	黄海	5.3	G_1	G_1	√
3	1995-09-20	苍山	5.2	G_1	G_1	√
4	1996-05-03	包头	6.4	G_2	G_1	×
5	1997-07-28	黄海	5.1	G_1	G_1	√
6	1998-01-10	张北尚义	6.2	G_2	G_1	×
7	1999-03-11	张北	5.6	G_1	G_1	√
8	2000-01-12	岫岩	5.1	G_1	G_1	√
9	2001-06-05	乌拉特前旗	5.0	G_1	G_1	√

表 3 华北地区多元统计组合模型 2002～2003 年判别预报结果及检验表

序号	震例			样本原属类别	判别类别	检验(判对为√，判错为×)
	年-月-日	地点	M			
1	2002-04-22	隆尧	5.4	G_1	G_1	√
2	2003			未知	G_1	

量比较大的指标集合作为判别预报的变量集合；然后利用该地区每年变量集合的数值判别预报该地区次年最大震级属于那一类. 以华北地区(30°～42°N，108°～125°E)为例，参照前人有关地震活动性指标的研究成果(罗兰格，2002；李平等，2000；周翠英等，1999)，从时间扫描的 30 个地震参数中初选相对独立的、映震能力较强、普遍应用的地震活动指标 20 个：A_C 值、η 值、A 值、$A(b)$值、$P(b)$值、b 值、C 值、D 值、ΔF 值、m_f 值、Y_H 值、缺震、频度、活动强度熵、活动频度熵、活动时间熵、调制比、响应比、等效震级、地震活动度 S. 应用基于 GIS 的地震分析预报系统计算软件(陆远忠等，2002)，统计了 1972～2001 年的数据，取时间窗长为 3 年、时间步长为 1 年进行计算. 所得结果作为多元统计组合模型的初始测震学指标和数据. 震级资料取自国家地震局分析预报中心编辑的《中国地震目录》. 为避免余震序列对计算结果的影响，计算时对所有 $M_S \geqslant 5.0$ 地震序列的余震进行了剔除.

2.1　主成分分析选择地震预报变量

对初选 20 个测震学指标进行主成分分析计算，得到前 4 个主成分的累积贡献率为 91.386 5%. 因此，我们取前 4 个主成分已充分提取了这 20 个指标所提供的信息资源. 由 4 个主成分的因子载荷阵看到，90%以上的信息主要是由下述 6 个指标所提供，即 $A(b)$值、地震活动标度 ΔF、缺震、等效震级、地震活动度 S 和 Y_H 值. 另外应用相关分析对上述预报变量进行了检验. 其最大震级 M 和预报变量的相关系数都是比较大的. 因此，上述 6 个指标选为预报变量集合，记 x_1＝$A(b)$值，x_2＝地震活动标度 ΔF，x_3＝缺震，x_4＝等效震级，x_5＝地震活动度 S，x_6＝Y_H 值.

2.2　马氏距离判别预报结果及检验

以华北地区为例，通过主成分分析筛选预报变量集合为 $\boldsymbol{X}=(x_1, x_2, \cdots, x_6)'$. 判别类别集合：$G_1$ 为年度最大震级 5.0～5.9；G_2 为年度最大震级 6.0 以上，建立多元马氏距离判别综合预报模型. 马氏距离判别函数值 $W(\boldsymbol{X})>0$，判 $\boldsymbol{X}$ 所属类别为 G_1；$W(\boldsymbol{X})<0$，判 $\boldsymbol{X}$ 所属类别为 G_2. 利用 1973～1992 年为建模学习样本，1993～2001 年为检验样本，2002～2003 年为预报样本. 统计每年 6 项预报变量与次年在该区发生最大地震震级的关系，即根据该区本年度中各项地震活动指标的数值，预测次年该区发生地震震级属于哪一类别. 所得结果见表 1～表 3.

取华北地区每年最大震级为总体样本，作为多元统计组合模型的方法应用检验. 其学习样本共 20 个，错判 5 个，错判率为 25%；检验样本共 9 个，错判 2 个，错判率为 22%. 判别预报 2002 年结果与实际相符，判别预报 2003 年结果为 G_1 类，即震级为 5～6，可以认为 5.5 级左右. 由此可见所建模型初步取得较好的预报效果.

3　讨论和结论

本研究所设计的多元统计组合模型针对众多地震预报指标，在尽可能多地获取孕震信息资源的前提下，简化数据结构达到降维的目，通过马氏距离判别综合预测. 初步研究取得了较好的效果，在中短期预报中有一定的实用价值，但实际应用中还存在某些问题.

1) 作为模型的初步应用研究，将震级分为两类，即 5.0～5.9 和≥6.0. 这样划分比较粗，往往忽略 4.0～4.9 级地震的预报.

2) 影响错判率的因素常常与分类的样本个数有关. 若将 7 级以上地震单独分为一类，

会因为样本个数太少而使错判率增加. 这个问题还需进一步研究解决.

3) 在实际计算中，学习样本建模所取时间段常常依赖于大的地震周期的划分，这需要地震预测方面的经验和模拟计算的结果，否则会在一定程度上影响错判率.

参 考 文 献

陈锦标，李全林. 1989. 震级序列的前兆分析——关于 b 值研究的程式报告[A]. 见：国家地震局科技监测司编. 地震预报方法实用化研究文集，地震学专辑[C]. 北京：学术书刊出版社，163～172

方开泰. 1989. 实用多元统计分析[M]. 上海：上海华东师范大学出版社，179～208，291～336

李平，朱元清，肖兰喜，等. 2000. 华北地区强震前空间动态场中短期综合预报研究[J]. 地震，**20**(2)：37～47

陆远忠，吕悦军，郑月君. 1994. 地震演化指数 Y_H 与细胞自动机模型检验[J]. 地震，(增刊)：12～17

陆远忠，李胜乐，邓志辉，等. 2002. 基于 GIS 的地震分析预报系统[M]. 成都：成都地图出版社，17～41

罗兰格. 2002. 我国地震综合预报方法研究的回顾与展望[J]. 华北地震科学，**20**(4)：1～18

秦保燕，刘江峰. 1989. 甘肃省及邻区中强震前调制小震异常指标研究[A]. 见：国家地震局科技监测司编. 地震预报方法实用化研究文集，地震学专辑[C]. 北京：学术书刊出版社，473～501

王炜，戴维乐，黄冰树. 1994. 地震震级的统计分布及其地震因子 M_f 值在华北中强以上地震前的异常变化[J]. 中国地震，**10**(增刊)：95～129

王炜，刘震华. 1987. 地震时间间隔的统计分布及其地震危险度 D 值在华北大震前的异常变化[J]. 地震学报，**9**(2)：113～127

吴佳翼，曹学锋. 1983. 地震活动性的定量化问题[J]. 地震，(6)：13～16

周翠英，朱元清，王红卫，等. 1999. 华北地区地震学指标的定量对比筛选及其综合预报方法研究[J]. 地震学报，**21**(2)：208～213

Johnson R A, Wichern D W. 1992. *Applied Multivariate Statistical Andysis*, 5th ed[M]. Englewood Cliffs: Prentice Hall, 426～461

JOINT MULTIVARIATE STATISTICAL MODEL AND ITS APPLICATIONS TO SYNTHETIC EARTHQUAKE PREDICTION

Han Tianxi[1)] Jiang Chun[2)] Wei Xueli[1)] Han Mei[1)] Feng Deyi[2)]

1) *Tianjin University of Technology*, *Tianjin* 300191, *China*

2) *Earthquake Administration of Tianjin Municipality*, *Tianjin* 300201, *China*

Abstract: Considering the problems that should be solved in the synthetic earthquake prediction at present, a new model is proposed in the paper. It is called joint multivariate statistical model combined by principal component analysis with discriminatory analysis. Principal component analysis and discriminatory analysis are very important theories in multivariate statistical analysis that has developed quickly in the late thirty years. By means of maximization information method, we choose several earthquake prediction factors whose cumulative proportions of total sample variances are beyond 90% from numerous earthquake prediction factors. The paper applies regression analysis and Mahalanobis discrimination to extrapolating synthetic prediction. Furthermore, we use this model to characterize and predict earthquakes in North China (30°～42°N, 108°～125°E) and better prediction results are obtained.

Key words: joint multivariate statistical model; principal component analysis; discriminatory analysis; synthetic earthquake predication