旱涝研究的新方法——投影追踪回归*

史 久 恩 陈 忠 琏
(北京气象学院, 100081) (中国科技大学研究生院)

常 红
(Department of Mathematics, Colorado School of Mines, USA)

提 要

投影追踪回归是一种处理高维问题、特别是高维非正态问题的新兴统计方法. 其基本思想是将高维数据投影到低维 (1—3 维)子空间,使在此投影空间中得到的构型最能反映原来高维数据的结构和特征,从而达到研究分析高维数据的目的.

本文将投影追踪回归应用于旱涝研究,希望能为旱涝数据分析提供新的手段,并解决一些线性模型所不能解决的问题.

关键词 旱涝;投影追踪回归 (PPR).

一、引 言

投影追踪回归 (Projection Pursuit Regression 简称 PPR)[1,2] 是一种处理高维问题、特别是高维非正态问题的新兴统计方法,其基本思想是通过选择数据的低维投影来研究数据的高维结构. 它包含的内容有: 第一是通过观测投影图形来揭示原数据的结构;第二是根据实际问题的要求,选择一种投影指标,通过计算机的程序自动挑选出使投影指标达到极值的投影. 由于将高维数据投影到低维子空间, 就克服了经常遇到的维数太大而样本点相对较少的困难; PPR 的另一特点是, 可以排除与数据结构、特征无关的或关系很小的变量的干扰.

目前, 回归分析是在气象问题中应用最广泛的统计方法之一[3,4], 它在统计气象应用中取得了一定的成效, 但也存在不少问题[5]. 其原因之一是当前所用的回归模型大多是线性回归,它要求数据满足正态性和线性性. 而实际的气象数据往往不满足这些条件. 此外,由于气象数据序列的样本容量较小,当变量个数较多时, 易出现所得的回归方程不稳定. 又由于气象数据中常存在少量的离群点 (outlier), 使得相关系数和回归系数都具有不稳健性. 再者,由于大多数气象问题,事先并不知道数据的结构, 因此常用的参数模型也不一定适合. 但是 PPR 克服了这些不足,它是一种稳健的非参数方法,可以有效地寻找线性投影中的非线性结构,揭示出气象数据所遵循的模型结构. 因此本文将 PPR 应

1993 年 4 月 7 日收到, 5 月 9 日收到修改稿.
* 得到国家自然科学基金和云南省气象局的经费资助.

用到气象中来，希望它能为降水、旱涝等研究提供一种新的、有效的分析和预测手段。

二、投影追踪回归

设 $(\vec{X},Y)$ 是一对随机变量，其中 $\vec{X}$ 是 p 维的，Y 是一维的。回归问题就是利用样本 $\{(\vec{x}_i,y_i);\ i=1,2,\cdots,n\}$ 来估计条件期望

$$f(\vec{X})=E(Y|\vec{X})$$

目前最常用的回归模型是线性模型。在线性模型中，假设 $f(\vec{X})$ 是 $\vec{X}$ 的线性函数。然而在许多实际问题中 $f(\vec{X})$ 往往是非线性的，勉强应用线性逼近，误差太大，难以获得好的效果。从数据资料出发，寻求回归函数的一种较好的方法是非参数回归方法，它的基本思想是找到 $\vec{X}$ 点的邻域中的 $\vec{X}_i$ 点，把它们相应的 Y_i 值做平滑，从而得到 $f(\vec{X})$ 的估计值。但将这种方法用于多变量的高维空间时，却不能克服"维数祸根"的困难，即样本太小，无法达到使用这种方法的要求。

为了避免上述方法所遇到的矛盾，PP 回归采用了一系列岭函数的"和"来逼近回归函数，即

$$f(\vec{X})\sim\sum_{j=1}^{m}g_j(a_j^T\vec{X}) \tag{1}$$

式中 $g_j(a_j^T\vec{X})$ 表示第 j 个岭函数，$a_j^T\vec{X}$ 为岭函数的自变量，它是 p 维随机变量 $\vec{X}$ 在 a 方向上的投影。

这样一来，一方面可以用增大 m 的办法来减少模型误差；另一方面由于采用了投影手法，将高维问题转化为一维问题，从而克服了"维数祸根"。不难看出，当 $m=1$，$g_1(t)=ct$ 时，上式就成了线性回归。因此，PP 回归也是线性回归的一种推广。

实现 PP 回归的算法步骤[7]是

第一步：选定一个初始的回归模型，例如，$f_0(\vec{x})=$常数

第二步：重复过程

2.1 寻找一个投影方向 a，使得当前的残差 $r_i=y_i-f(\vec{x}_i)$，$i=1,2,\cdots,n$ 与投影 $Z=a^T\vec{x}$ 有尽可能大的回归依赖关系，得到平滑函数 $g_a(Z)$；

2.2：将回归模型更新为

$$f(\vec{x}):=f(\vec{x})+g_a(a^T\vec{x}) \tag{2}$$

2.3：重复 2.1 与 2.2 过程，直到回归模型不能得到明显改进为止。

经过 m 步叠代后，回归函数 $f(\vec{x})$ 可以近似为

$$f(\vec{x})\approx f_0(\vec{x})+\sum_{j=1}^{m}g_j(a_j^T\vec{x}) \tag{3}$$

不妨假定 $f_0(\vec{x})=0$，因此上式即是用岭函数有限项和逼近 $f(x)$。

a_j,g_j 的选取方法是，在 $j<m$ 给定后，寻找 a_m,g_m 使

$$r_m(x)=f(\vec{x})-\sum_{j=1}^{m-1}g_j(a_j^T\vec{x}) \tag{4}$$

的模 $\int r_m^2(x)dp$ 在 m 升到 $m+1$ 时，减少得最多．也即在计算中要使目标函数

$$Q=\sum_{i=1}^{n} r_{mi}^2=\sum_{i=1}^{n}[r_{m-1,i}-g_m(a_m^T\vec{x}_i)]^2 \tag{5}$$

达到最小．固定 a_m 以后，理想的 g_m 是条件期望

$$g(a^T\vec{x})=E(y|a^T\vec{x}) \tag{6}$$

假如对每个固定的 a，都能找到(6)式中的 g，目标函数 Q 就只是 a 的函数了，可用最优化方法解决．所以关键的一步是如何找到与 a 有关的 g．

按文献[2]，第 2.1 步的具体步骤如下：

2.1：复重过程

2.1.1　得到当前的投影方向 a；

2.1.2　将 $Z_i=a^T\vec{x}_i$，$i=1,2,\cdots,n$ 从小到大排列，排序后仍记为 Z_i，$i=1,\cdots,n$，相当的当前残差记为 r_i，$i=1,2,\cdots,n$；

2.1.3　滑动中位数平滑

$$r_i\simeq \mathrm{med}\{r_{i-1},r_i,r_{i+1}\}$$

其中，med 表示取中位数；

2.1.4：对于每个 Z_i 用下述方法估计它的响应方差 σ_i^2，对于 Z_i 的 k 个左邻、k 个右邻，共 $2k$ 对 (Z_i,r_i)，作局部的线性拟合，并求出残差平方和，除以 $2k$，得到 σ_i^2；

2.1.5　对于 $\{\sigma_i^2\}$ 作固定带宽的滑动平均

$$\sigma_i'^2=(\sigma_{i-k}^2+\cdots+\sigma_{i-1}^2+\sigma_{i+1}^2+\cdots+\sigma_{i+k}^2)/(2k);$$

2.1.6　对于序列 $\{r_i\}$ 作局部线性拟合，带宽由 $\sigma_i'^2$ 所确定，得到 $g_a(Z_i)$，$i=1,\cdots,n$；

2.1.7　r_i 与平滑值 $g_a(Z_i)$ 的残差 $(i=1,\cdots,n)$ 的平方和作为 a 方向的投影目标函数 Q；

重复以上过程，直到能找到使 Q 达到最小的投影方向 a．其中，2.1.3 步是为了稳健性．2.1.4 步与 2.1.5 步是为了得到局部方差的平滑估计 $\sigma_i'^2$．通常并不把第 i 个值本身加入计算，以防止过度的拟合等．

三、旱涝预测应用举例

1. 研究对象

本文将全国 160 个站 30 年夏季 (6—8 月)的降水距平百分率进行了聚类分析，共得 18 个降水分区．本例将其中两个区的降水距平百分率区域平均值作为研究对象，它们分别为：(1) 上海、宁波、杭州三站组成的 I 区；(2) 汉口、九江、屯溪、安庆、常德、岳阳、南昌、贵溪、衢县、浦城十站组成的 II 区．

2. 可能因子

对旱涝有较大影响的可能因子，本例采用以下五个方面：

(1) 北太平洋副热带高压的强弱对我国夏季的降水影响较大，因此选用了北半球500hPa 副高面积指数作为第一类可能因子．时间取为每年的1—5 月；

(2) ENSO 现象是热带大尺度海气相互作用的一种现象，对我国天气气候变化有一定影响，本例将南方涛动指数 (SOI) 做为第二类可能因子，其时间取 I、II 两区降水的前期 SOI 值(上年 6 月至当年 5 月)；

(3) 对流层中层各大尺度波动的动能对降水变化是重要的．本例取北半球 500hPa 月平均动能[10]作为第三类可能因子，时间为每年 1—5 月；

(4) 太阳活动对大范围的大气和气候有一定影响，本例取瑞士苏黎世天文台发布的太阳黑子相对数月平均值作为第四类可能因子，时间取每年 12 个月(上年 6 月至当年 5 月)；

(5) 海洋是比陆地更为重要的大气热源，海洋的作用是一个重要的因素．本例取由 PP 主成分分析[9]得出的北太平洋海温场的第一至第六时间系数作为第五类可能因子，时间取每年 12 个月．

为了客观、全面地反映各类可能因子在前期不同时刻对研究对象的不同影响，我们进行了因子的预处理，即分别将每类可能因子的前期各月 (5 个月或 12 个月) 对研究对象 (I 区或 II 区降水)进行 PP 回归分析，取最重要的第一投影因子作为代表该类因子的综合指标，于是将上述五类可能因子共 106 个浓缩为 10 个综合因子．

3. 计算结果

以 I 区夏季 (6—8 月)降水距平百分率为研究对象，用 PP 回归进行搜索，获得第一个投影方向 α_1：

$$\alpha_1 = (0.0603, -0.1863, 0.0595, 0.0, 0.9788, 0.0, -0.0062, 0.0041, 0.0, 0.0)$$

可以看出第 5 个因子(即海温第一时间系数)对 I 区降水的重要影响，而第 2 个因子(南方涛动指数) 的作用是负的．这说明了 ENSO 现象对 I 区夏季降水的影响是比较大的．

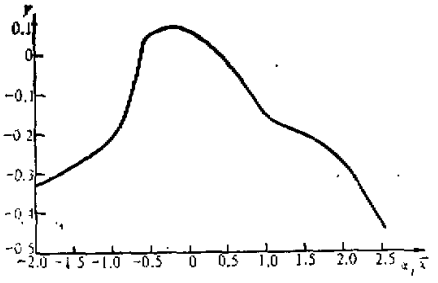

图 1 第一投影因子 $\alpha_1^T\vec{x}$ 与 y 的曲线图

图 1 为第一投影因子 $\alpha_1^T\vec{x}$ 与 I 区降水 y 的曲线图．图中 y 是经过光滑化的，记为 $y = g_{\alpha_1}(\alpha_1^T\vec{x})$．从图中可明显地看出，$Z_1 = \alpha_1^T\vec{x}$ 与 y 的关系是非线性的．当 Z_1 值中等时，I 区降水接近正常，而当 Z_1 值较大(或较小)时，I 区降水均偏小．

用 PP 回归继续搜索第二个投影方向 α_2：

$$\alpha_2 = (0.0106, -0.2862, 0.9191, 0.0327, -0.2423,$$
$$-0.0647, 0.0, -0.0654, -0.0702, 0.0003)$$

从中可见第 3 个因子(北半球 500hPa 平均动能)和 SOI 对 I 区降水有影响．

图 2 为第一步光滑后的余量 $r_1(x) = y - g_{\alpha_1}(\alpha_1^T\vec{x})$ 与第二个投影因子 $Z_2 = \alpha_2^T\vec{x}$ 的

函数图形。光滑后的 r，记为 $g_{\alpha_2}(\alpha_2^T\vec{x})$。

图中的曲线形状也是非线性的。

类似上述做法，还可找到其他投影因子，由于篇幅所限，不再一一例举。

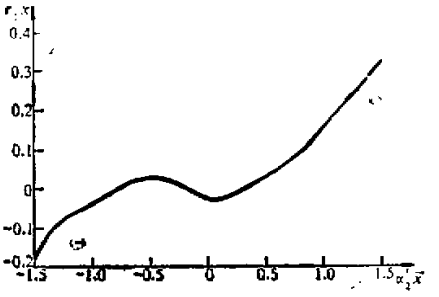

图 2　$\alpha_1^T\vec{x}$ 与 $r_1(x)$ 的曲线图

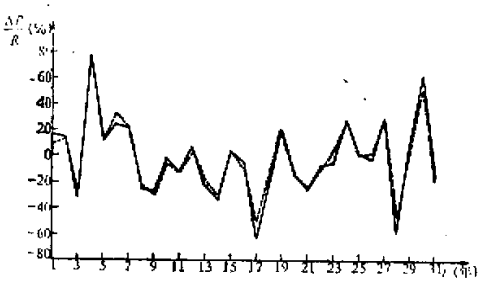

图 3　I 区夏季降水实况(实线)和用 PP 回归计算的拟合情况(虚线)及独立样本试报

图 3 是 I 区夏季 (6—8 月)降水距平百分率的历史拟合和独立样本试报的计算结果。

表 1 为用 PP 回归模型所得出的 I、II 两区夏季 (6—8 月) 降水距平百分率历史拟合 (30 年)的相关系数、平均绝对误差和独立样本 (1 年)的试报情况。

表 1　用 PP 回归做 I、II 两区夏季降水的历史拟合和试报情况

降水地区	拟合相关系数	拟合平均绝对误差(%)	独立样本实况(%)	独立样本试报(%)
I　区	0.98	15	−22.5	−24.5
II　区	0.95	28	−30.8	−29.5

4. 与线性回归模型的计算结果比较

为了进行比较，本例使用上述同样的研究对象 (I 区和 II 区夏季降水距平百分率) 和 10 个综合因子，应用线性回归模型做计算，所得结果列于表 2。

对照表 1 和表 2 的相应部分可以看出，从历史拟合 (30 年)和独立样本 (1 年)的情况均说明 PP 回归模型的计算结果优于线性回归模型。

表 2 用线性回归做 I、II 两区夏季降水的历史拟合和试报情况

降水地区	拟合相关系数	拟合平均绝对误差(%)	独立样本实况(%)	独立样本试报(%)
I 区	0.67	51	−22.5	11.9
II 区	0.71	58	−30.8	55.9

此外，还计算了其他区的夏季降水和夏季气温等数据资料，都反映出 PP 回归的历史拟合和试报情况要比线性回归模型的好一些。

四、结论和讨论

通过本文将 PP 回归分析应用于旱涝研究可见，PP 回归比之常用的线性回归确有明显的优点。其最显著的特点是它能成功地在高维数据分析中克服由于高维空间中散布的数据点非常稀疏而引起的严重困难，因为 PP 回归对数据的分析是在低维子空间上进行的，对 1—3 维空间来说，数据点就足够密，可以发现数据在投影空间中的结构和特征，这一点对气象数据的样本容量通常不很大的情况，更为重要。

在实例研究中，可以看到 PP 回归能反映气象问题中的非线性结构。同时，通过 PP 回归分析的例子，还可以看出各个综合因子在不同时期对旱涝的影响是不同的，这对干旱、洪涝气候分析和预测都是有益的。

从 PP 回归与线性回归的对比可以看出，PP 回归模型的历史拟合及独立样本试报均比线性回归模型要好，这说明 PP 回归有可能较好地解决一些线性回归所解决不好的问题。

PP 回归方法虽然有许多优点，但它也有不足之处，即计算量比较大，计算时间几乎是随着维数的增加呈指数增加。通过其它一些应用研究表明，如将常用的线性回归方法和 PP 回归方法结合起来研究旱涝等气象问题，往往能产生更好的效果，而且还能减少计算量。

参 考 文 献

[1] Huber, P. J., 1985, Projection pursuit, *Ann. Statistics*, **13**, 435—525.

[2] Friedman, J. H. and W. Stuetzle, 1981, Projection pursuit regression, *J. Amer. Statistics Assoc.*, **76**, 817—823.

[3] 张尧庭、方开泰，1982，多元统计分析引论，科学出版社。

[4] 项静恬、史久恩等，1991，动态和静态数据处理，气象出版社。

[5] 俞善贤、陈孝源，1988，气象数据回归分析中的若干问题及其对策，气象学报，**46**，No. 3，327—332.

[6] 么枕生、丁裕国，1990，气候统计，气象出版社。

[7] 陈忠琏，1986，多元数据分析的 PP 方法，数理统计与应用概率，1 卷，No. 2，103—123

[8] 史久恩、周琴芳等，1983，北半球 500 毫巴月平均场球谐系数和物理量，气象出版社。

[9] Chang Hong, Shi Jiu-En and Chen Zhong-Lian, 1990, Projection pursuit principal component analysis and its application to meteorology, *Acta Meteorologica Sinica*, **4**, 254—263

A New Method for the Climate Research of Drought and Flood——Projection Pursuit Regression

Shi Jiuen
(Beijing Meteorological College, Beijing 100081)

Chen Zhonglian
(Graduate School, Chinese Academy of Sciences, Beijing)

Chang Hong
(Department of Mathematics, Colorado School of Mines, USA)

Abstract

Projection Pursuit (PP) Regression is a new statistical method that can deal with high-dimensional problems, especially high-dimensional and nonnormal problems. The principal idea of the PP regression is to project highdimensional data onto such a low-dimensional (1—3 dimensions) subspace that in an optimized way the configuration of data obtained in the projective subspace can reflect the structure and feature of the original high-dimensional data, and therefore the data can be analyzed and studied.

This paper applies the PP regression to the research of drought and flood. We hope that PP regression will offer a new tool for the data analysis of drought and flood, and solve some problems which are beyond the ability of linear models.

Key words: Drought and flood; Projection pursuit regression.