

投影追踪回归法在月径流长期预报中的应用

李杰友¹,夏红梅²,熊学农¹,林镜榆¹

(1. 河海大学水文水资源及环境学院,江苏 南京 210098;2. 广东省南水水电厂,广东 乳源 512700)

摘要 结合长期水文预报的特点,将投影追踪回归(Projection Pursuit Regression,缩写为 PPR)分析方法应用于月径流长期预报研究,探讨水文气象数据分析,为长期水文预报提供新的有效途径,此外,解决一些线性回归方法难以解决的问题.

关键词 投影追踪回归;月径流;长期预报;数据分析

中图分类号 P338+.2 **文献标识码** A **文章编号** 1000-1980(2001)01-0046-04

1 问题的提出

设有回归问题 $y = f(x) + \epsilon$, 其中 y 是一维随机变量, x 是 d 维随机或非随机的向量, ϵ 是随机误差. 线性回归是用 x 的线性函数 $a'x$ 作为回归函数 $f(x)$ 的一阶逼近, 这当然会产生模型误差. 因为在许多的实际问题中, $f(x)$ 往往是非线性的, 勉强用线性逼近, 误差太大, 难于获得好的效果. 投影追踪回归分析方法是采用一系列‘岭函数(ridge function)’来逼近回归函数 $f(x)$. 从理论上讲, PPR 法具有以下优点: (a) 采用投影的方法, 将高维问题转化为一维问题, 从而克服了‘维数灾难(curse of dimensionality)’; (b) 能处理离散化实变量时间系列非正态分布的回归问题; (c) 能有效处理离散化实变量时间系列的离散群点, 使回归方程受离散群点的影响减小而保持稳定; (d) 回归方程以非线性方程的形式出现, 不需经逐段线性化来逼近非线性问题.

2 投影追踪回归方法原理

2.1 回归函数 $f(x)$ 的投影追踪逼近

Friedman 和 Stuetzle 在文献[1]中提出的 PPR 法是用‘岭函数(ridge function)’的和

$$f_k(x) = \sum_{i=1}^k g_i(a_i'x) \tag{1}$$

来逼近 $f(x)$, 然后用观测值 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 来估计 $f_k(x)$.

若 x 是随机向量, 文献[1]逼近 $f(x)$ 采用的准则是

$$E[f(x) - f_k(x)]^2 \rightarrow 0 \quad (k \rightarrow \infty) \tag{2}$$

若 x 是非随机的, 则需要选择一个 R^d 上的概率测度来作为 X 的‘分布’.

寻找 f_k 的办法是逐次找出最优的 $g_1(a_1'x), g_2(a_2'x), \dots, g_k(a_k'x)$. 若选定方向 a , 按逼近 $f(x)$ 的准则(式(2)), $g_a(a'x) = E(f(x) | a'a'x)$ 显然是使

$$E[f(x) - g_a(a'x)]^2 = \min_a \tag{3}$$

的解. 因此, 第一个最优方向 a_1 应该是 $E[f(x) - g_a(a'x)]^2 = \min_a$ 的解, 而相应的 $g_1(a_1'x) = g_{a_1}(a_1'x)$. 令 $r_1(x) = f(x) - g_1(a_1'x)$, 并用 $r_1(x)$ 代替 $f(x)$, 重复上述过程, 即可得到 $g_2(a_2'x), g_3(a_3'x), \dots, g_k(a_k'x)$ 以及式(1)中的 $f_k(x)$.

Huber 在文献[2]中证明了这样求出的 $f_k(x)$ 是 L_1 (一维) 收敛到 $f(x)$ 的, 即

$$E(|f(x) - f_k(x)|) \rightarrow 0 \quad (k \rightarrow \infty) \tag{4}$$

2.2 回归函数 $f(x)$ 的估计

设 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 是观测值, PPR 法估计回归函数 $f(x)$ 的步骤是:

- a. 对给定的投影方向 a , 用 $(a'x_i, y_i)$ 作一维的非参数回归, 得到 $\hat{g}_a(a'x)$;
- b. 求 $\hat{a}_1$, 使 $\sum_{i=1}^n (y_i - \hat{g}_a(a'x_i)) = \min_a$, 得到 $\hat{g}_1(\hat{a}_1'x) = \hat{g}_{\hat{a}_1}(\hat{a}_1'x)$;
- c. 令 $r_i = y_i - \hat{g}_1(\hat{a}_1'x_i)$ ($i = 1, 2, \dots, n$, 下同), 用 r_i 代替 y_i , 重复 a, b 过程得到 $\hat{g}_2(\hat{a}_2'x)$;
- d. 反复进行 c 中的迭代, 即可得到 $\hat{g}_3(\hat{a}_3'x), \dots, \hat{g}_k(\hat{a}_k'x)$, 直到第 k 步, 残差平方和 $\sum_{i=1}^n r_i^2$ 没有明显地减少为止.

通过上述步骤就得到了回归函数 $f(x)$ 的 PPR 法估计 $\hat{f}_k(x) = \sum_{i=1}^k \hat{g}_i'(\hat{a}_i'x)$.

2.3 计算要点

- a. 根据实际问题构造一个初始的回归模型, 例如 $f_0(x) = c$ (常数).
- b. 寻找一个方向 a , 使得当前残差 $r_i = y_i - f(x_i)$ 与投影 $z = a^Tx$ 有尽可能大的回归依赖关系, 得到平滑函数 $g_a(z)$.
- c. 将回归模型更新为 $f(x) \triangleq f(x) + g_a(a^Tx)$.
- d. 重复 b, c 过程, 直到回归效果不能明显改进为止. 经过 m 步迭代后, 回归函数 $f(x)$ 可近似表示为

$$f(x) \doteq f_0(x) + \sum_{j=1}^m g_j(a_j^Tx)$$

上述计算中, 关键是要设计好一维的非参数回归的算法, 既要算得快, 又要有好的效果. 第 b 步是最重要的一步, 据文献[1], 第 b 步的细算法为:

- (a) 得到当前的投影方向 a ;
 - (b) 将 $z_i = a^Tx_i$ 按升序排序后仍记为 z_i , 相应的当前残差记为 r_i ;
 - (c) 将 r_i 做滑动中位数平滑 $r_i = \text{med} \{r_{i-1}, r_i, r_{i+1}\}$ (med 表示中位数);
 - (d) 估计每个 z_i 的响应方差 σ_i^2 , 方法是将 z_i 的左、右邻共 $2k$ 对 (z_i, r_i) 作局部的线性拟合, 并求出其残差平方和, 除以 $2k$ 得到 σ_i^2 ;
 - (e) 对 $\{r_i^2\}$ 作固定带宽的滑动平均 $\sigma_i'^2 = (\sigma_{i-k}^2 + \dots + \sigma_{i-1}^2 + \sigma_i^2 + \sigma_{i+1}^2 + \dots + \sigma_{i+k}^2) / (2k)$;
 - (f) 对 $\{r_i\}$ 作局部线性拟合, 其带宽由 $\sigma_i'^2$ 确定, 得到平滑值 $g_a(z_i)$;
 - (g) 将 r_i 与平滑值 $g_a(z_i)$ 的残差平方和作为 a 方向的投影目标函数 $Q(a)$.
- 重复以上过程, 直至找到使 $Q(a)$ 达到最小的投影方向 a . 在此过程中, 需要一个使得 $|a| = 1$ 且 $Q(a)$ 最小的过程, 文献[1]建议使用 Rosenbrock 方法.

3 应用实例

3.1 研究对象及研究区水文气象特征

年前月径流的长期预报, 即上一年末对下一年全年的月径流量作出预报, 对生产单位制定生产计划、防洪抗旱、水资源的规划管理和综合利用有着十分重要的意义. 本文将福建省水口电厂月径流量作为研究对象.

水口电厂位于福建省闽江干流, 坝址以上流域面积为 $52\,438\text{ km}^2$, 属亚热带季风气候区, 气候温和、雨量充沛. 全年大致可分为四个时期 3~4 月为春雨期, 5~6 月为梅雨期, 7~9 月为台风雷雨期, 10 月至次年 2 月为旱季. 其中梅雨期为主要雨季, 该时期降水量约占全年降水量的 36%. 造成本流域大范围降雨的天气系统主要有锋面雨和台风雨两大类. 锋面雨属温带天气系统, 台风雨属热带天气系统, 两者成因不同, 其降雨特性亦有差异. 坝址以上流域处在西太平洋高压脊的西北缘, 上空盛行西南气流, 气流将低纬度海洋暖湿空气带送进本流域, 此时北方冷空气强度虽比冬季减弱, 但活动仍频繁与活跃. 当冷空气向南侵袭时, 冷暖气团在流域上空对峙, 形成静止锋, 引起本流域持久和大范围的降雨. 台风雨则是一个深厚的热带气旋, 本流域发生的台风暴雨一般历时仅为 24 h, 暴雨峰往往集中在台风登陆前 12 h 前后, 台风登陆后随着强度减弱, 降水亦逐渐减弱或停止.

3.2 预报因子的引进

长期天气过程具有非绝热性和全球性的特点. 年内月径流量的多寡与大范围的环流特征及其维持的能源密切相关. 结合水口流域的水文气象特征, 本研究采用的气候、气象预报因子有 : (a) 北太平洋月平均海温场 (ST) ; (b) 东半球月平均 500 hPa 高度场 (H_5) ; (c) 北半球月平均 100 hPa 高场场 (H_1) ; (d) 全国 100 站月平均降水 (RR) ; (e) 全国 100 站月平均气温 (TT) ; (f) 大气海洋物理量 (MR), 主要包括环流指数、副高指数、极涡、青藏高原指数、印缅槽指数、海温指标、太阳黑子相对数、南方涛动指数等.

上述资料均来源于中央气象局公开发行的《气候监测公报》. 其中北太平洋月平均海温场资料地理范围为 $0^{\circ} \sim 50^{\circ}N, 180^{\circ}E \sim 80^{\circ}W$, 网格距为 $5^{\circ} \times 5^{\circ}$. 东半球月平均 500 hPa 高度场资料地理范围为 $10^{\circ}N \sim 85^{\circ}N, 5^{\circ}E \sim 180^{\circ}E$, 网格距为 $5^{\circ} \times 5^{\circ}$. 北半球月平均 100 hPa 高度场资料地理范围为 $10^{\circ}N \sim 80^{\circ}N, 10^{\circ}E \sim 0^{\circ}$, 网格距为 $10^{\circ} \times 10^{\circ}$. 预报因子和研究对象的时间均为每年的 1 ~ 12 月. 时域为 1964 年 1 月 ~ 1998 年 12 月 (其中, 1964 年 1 月 ~ 1994 年 12 月共 30 年作为历史拟合, 1995 年 1 月 ~ 1998 年 12 月作为独立样本试报).

年内月径流量的多寡与前期各类预报因子有着密切的联系, 为了客观、全面地反映上述预报因子在前期不同时刻对研究对象的不同影响, 分别将各类因子与研究对象进行 PPR 法分析, 从中挑选出最重要的第一投影因子作为代表该类因子的综合指标. 对各类因子的综合指标再通过 PPR 法对研究对象进行分析计算.

3.3 计算结果

3.3.1 影响因素

对各预报因子与预报对象进行 PPR 法分析计算, 获得的第一和第二个投影方向见表 1.

从表 1 中第一投影方向可以看出, 第 1 个因子 (北太平洋月平均海温) 对水口电厂月径流有重要影响. 第 4 个因子 (月平均降水) 的作用负的, 说明其对水口电厂月径流的影响是比较大的.

表 1 PPR 法分析计算的各预报因子投影方向						
Table 1 Calculated projecting direction of each forecasting factor						
投影方向	预 报 因 子					
	ST	H_5	H_1	RR	TT	MR
a_1	0.9534	0.0584	0.0	-0.1841	0.0	0.0032
a_2	0.2231	0.9081	0.0451	-0.2391	0.0701	0.0003

从表 1 中第二个投影方向可以看出, 第 2 个因子 (东半球月平均 500 hPa 高度场和第 4 个因子 (月平均降水) 对水口电厂月径流有影响).

3.3.2 试报结果

独立样本试报结果见表 2 (因篇幅所限, 历史拟合结果未列出). 据国家气候中心《气候预测评论 (1996)》一书中, 关于《国外长期预报业务及研究进展》一文介绍 ‘回归分析方法是目前国外长期预报业务中应用最多的一种统计方法, 例如印度国家气象局应用多元回归等方法制作长期预报, 准确率可达 70% 左右.’ 据此, 表中精度以相对误差小于或等于 30% 作为合格标准进行评定.

表 2 1995 ~ 1998 年月径流试报结果

Table 2 Monthly discharge forecasting results for 1995 ~ 1998

月 份	1995 年			1996 年			1997 年			1998 年		
	实测值	试报值	相对误差	实测值	试报值	相对误差	实测值	试报值	相对误差	实测值	试报值	相对误差
	$/(\text{m}^3 \cdot \text{s}^{-1})$	$/(\text{m}^3 \cdot \text{s}^{-1})$	$/\%$	$/(\text{m}^3 \cdot \text{s}^{-1})$	$/(\text{m}^3 \cdot \text{s}^{-1})$	$/\%$	$/(\text{m}^3 \cdot \text{s}^{-1})$	$/(\text{m}^3 \cdot \text{s}^{-1})$	$/\%$	$/(\text{m}^3 \cdot \text{s}^{-1})$	$/(\text{m}^3 \cdot \text{s}^{-1})$	$/\%$
1	886	846	4.5	549	447	0.4	555	562	-1.2	2065	1117	45.9*
2	1068	977	8.5	521	594	-14.0	924	977	-5.7	3948	977	75.3*
3	1688	1648	3.7	1358	1477	-8.8	1197	1648	-37.7*	4772	4899	-2.7
4	2275	1912	16.0	2387	2632	10.3	1926	1571	18.4	2022	2002	1.0
5	2602	2734	-5.1	1524	1729	13.5	1811	2027	-11.9	3040	2734	10.1
6	5607	5068	9.6	2713	3631	-33.8*	3199	3631	-13.5	8234	5925	28.0
7	3189	2359	26.0	1111	1765	58.9*	3984	2656	33.3*	2311	2953	-27.8
8	2842	1655	41.8*	1665	1471	11.7	3203	2311	27.8	1539	1038	32.6*
9	1061	982	7.4	957	946	1.1	2482	1799	27.5	1090	910	16.5
10	934	856	8.4	566	526	7.1	1556	994	36.1*	660	746	-13.0
11	593	521	12.1	479	494	-3.3	1309	926	29.3	669	606	9.4
12	516	421	18.4	446	543	-21.7	1442	638	55.8*	691	651	5.8

* 表示在该标准下不合格.

从表 2 中可见,1995 ~ 1998 年的合格率分别为 91.7%,83.3%,66.7%和 75.0%,平均为 79.2%,精度令人满意. 1996 ~ 1998 年由于受厄尔尼诺(EL Nino)现象的影响,世界性气候变化异常,受其影响,我国大部分地区发生洪涝和旱灾. 1998 年水口以上流域发生了历史上的特大洪水,其首场洪水的洪峰流量为 $8\,280\text{ m}^3/\text{s}$,三日平均流量为 $4\,023\text{ m}^3/\text{s}$,五日平均流量为 $3\,332\text{ m}^3/\text{s}$. 六月份一场洪水的洪峰流量为 $37\,500\text{ m}^3/\text{s}$ (历史上最大洪水),三日平均流量为 $24\,800\text{ m}^3/\text{s}$,五日平均流量为 $2\,166\text{ m}^3/\text{s}$. 两场洪水的洪峰流量分别为正常年份的 1.7 倍和 5.1 倍. 笔者认为这也是造成误差大的主要原因之一.

4 结 语

PPR 法是一种有特色的方法. 本文将其应用于月径流长期预报,从历史拟合和试报结果来看,预报的准确率比单纯的线性回归方法好,拟合相关系数要高(为了进行比较,使用上述同样预报对象和预报因子,应用线性回归方法进行计算,但未采用分级差别及门限回归,所得结果因篇幅所限未列出). 在福建省水口电厂、湖北省黄龙滩电厂、河北省潘家口水库月径流预报研究中发现,当研究对象与预报因子之间存在高度的非线性关系时,PPR 法的预报准确率与逐步回归方法相比并无显著提高,因为 PPR 法也无法处理这种高度的非线性关系.

参考文献：

[1] Friedman J H, Stuetzle W. Projection pursuit regression[J]. JASA76, 1981 :817 ~ 823.
[2] Huber P J. Projection pursuit. Ann. Statist [J]. 1985, 13 :435 ~ 475.
[3] 陈忠琰. 多元数据分析的 PP 方法 [j]. 数理统计与应用概率,1986(2) :103 ~ 123.
[4] 史久恩,陈忠琰. 旱涝研究新方法——投影追踪回归 [j]. 大气科学,1993(增刊),13 ~ 18.

Application of Projection Pursuit Regression to Monthly Discharge Long-Range Forecast

LI Jie-you¹, XIA Hong-mei², XIONG Xue-nong¹, LIN Jing-yu¹

- (1. College of Water Resources and Environment, Hohai Univ., Nanjing 210098, China ;
2. Nanshui Water Power Plant, Guangdong Province, Ruyuan 512700, China)

Abstract The projection pursuit regression analytical method is a nonparameter steady estimating method. Combined with the characteristics of long-range hydrologic forecast,PPR method is applied to monthly discharge long-range forecast. It is expected that the PPR method can provide a new and effective way for data analysis and long-range hydrologic forecast.

Key words : projection pursuit regression ; monthly discharge ; long-range forecast ; data analysis