文章编号：1007-4619 (2005)02-0123-08

用偏最小二乘法反演二类水体的水色要素

杨燕明[1,2]，刘贞文[1,2]，陈本清[1,2]，唐军武[3]

(1. 国家海洋局 第三海洋研究所，福建 厦门 361005；
2. 海洋大气化学与全球变化国家海洋局 重点实验室，福建 厦门 361005；
3. 国家卫星海洋应用中心，北京 100081)

摘 要： 简要介绍了偏最小二乘法的原理、算法及优点。将该方法应用于黄海和南海二类水体光谱的水色要素反演，交叉检验结果表明反演精度高，预报相对误差不超过38%。该方法应用于加有5%随机噪声的人工合成光谱的水色要素反演，结果表明模型的稳健性强，预报相对误差不超过5%。研究结果表明，偏最小二乘法适合于处理变量多样本数又少的问题，适合于从二类水体光谱中提取水色要素信息。

关键词： 二类水体光谱；水色要素反演；偏最小二乘法

中图分类号： TP79 **文献标识码：** A

1 引 言

研究二类水体光谱与水色因子之间的关系具有重要的现实意义。中国近岸水体除南沙、台湾以东、海南岛附近等区域外，基本上都属于二类水体，特别是大陆架海域。随着中国海洋系列卫星的发射，研究二类水体光谱与水色因子之间的关系迫在眉睫。与一类水体的光学性质不同，二类水体的光学性质除受浮游植物及其降解物色素的影响外，还受黄色物质、悬浮物等影响，用一类水体的算法得到二类水体的物质含量是不可靠的。另一方面，二类水体大气校正波段(近红外波段)的离水辐亮度不为零，水体信号与气溶胶散射信号难以区分。二类水体的水色要素的反演一直是海洋光学遥感的难题[1]。二类水体的特点决定了其反演模型必然具有很强的局域性[1]。

超光谱水色遥感是长远的发展趋势，如美国海军提出的10nm分辨率、1024波段的水色遥感器[2]。就目前而言，随着新一代水色遥感器如中分辨率成像光谱仪(MODIS)和中国CMODIS的发射，光谱通道显著增加。如何充分利用这些光谱的信息，同时去除光谱波段之间的相关性，反演水色要素浓度，便有了十分现实的意义。在统计方法中，主成分分析能将所有光谱通道的反射率转换为几个正交的主成分，并已成功地应用于水色遥感领域[3,4]。

与主成分分析不同，偏最小二乘法(PLS)除了考虑光谱反射率矩阵外，还考虑了水色要素矩阵的信息[5]，因此更适合二类水体的信息提取。偏最小二乘法求得的模型的残差平方和与主成分回归相比差别不大，但PLS求得模型的预报残差平方和(PRESS)较小，因而具有较高的预报稳定性，得到的模型参数与实际问题的经验规律比较一致。另外，PLS方法比较适合于处理变量多而样本数又少的问题。在海水光谱测量中，要获得一个准确的水色要素和光谱反射率样本需要花很多的经费和精力，变量的个数(波段数)远大于样本的个数，但每个样本可给出的信息又很大。如何从少量样本中提取水色要素信息，是本文试图解决的问题。

2 偏最小二乘法简介

2.1 偏最小二乘法的基本原理

为了叙述上的方便，我们引入矩阵 X 和矩阵 Y。矩阵 X 描述 I 个样本数和 J 个光谱反射率波段，即 $I \times J$；矩阵 Y 描述 I 个样本数和 K 个水色因子，即 $I \times K$。主成分回归法在矩阵 X 的本征矢量或因子数的测试中，所处理的仅为 X 矩阵，而对矩阵 Y 中的信

收稿日期：2003-06-24；**修订日期：**2003-11-09

基金项目：国家海洋局重点青年基金(项目编号 Y00801)和中央级科研院所长基金资助。

作者简介：杨燕明(1966—)，男，博士后，研究员。主要从事海洋遥感、海洋声学研究。发表论文十余篇。

息并未考虑。事实上,Y 中也可能包含非常有用的信息。所以一种很自然的想法是,在矩阵 X 因子的测试中应同时考虑矩阵 Y 的作用。因此,偏最小二乘法和主成分分析相似,差别在于描述变量 X 中因子的同时也描述变量 Y。为此,首先将 X 矩阵作双线性分解,即

$$X = TP^{\mathrm{T}} + F \tag{1}$$

式(11)中矩阵 T 含有两两正交的得分矢量 t,即为矩阵 X 中变量的线性组合。F 为运用偏最小二乘法去拟合 X 所引进的误差,简称残差阵。P^{T} 代表 P 的转置,X 的负载。由于需要用到矩阵 Y 的信息,矩阵 Y 也作双线性分解,即

$$Y = UQ^{\mathrm{T}} + E \tag{2}$$

式(2)中各项的含义与式(1)的相似。PLS 要求 X 分解得到的得分变量 t 与 Y 分解得到的得分变量 u 为最大重叠或相关性最大,因此有

$$u = vt + e \tag{3}$$

式(3)中 e 为残差矢量,系数 v 根据最小二乘法确定。为了使得分矩阵 T 既可描述 X 矩阵,同时又可描述 Y 矩阵,则需要采取折中的方案,即将 T 进行坐标旋转。显然,坐标旋转后的 T 得分矩阵对于 X 矩阵的表达已不再是最优的状况,得分变量 t 的方差不再是最大,不再两两正交。

2.2 噪声滤去原理

在实际建模过程中,由于矩阵 X 的波段之间存在着相关性,同时还包含有噪声,所以在建立 *PLS* 模型时取 X 矩阵分解后的得分变量个数 h,而 h 一般小于实际回归的波段变量数 J,使得一些包含有噪声的得分变量被删除,因而具有噪声滤去作用。实际上,*PLS* 通过删除两种变量来滤除噪声:第一种是把那些与水色要素变量相关系数很小的,认为包含有很多噪声的波段变量删去;第二种若某一波段变量与目标变量虽然有一定的相关性,但这个变量又与另一个变量有较大的相关性,而另一个变量与目标变量的相关性较强,则这个变量若用于建模也会给模型引入噪声。因为这个变量中包含的与目标变量相关的成分已与另一个变量重叠,那些有用的成分完全可由另一个变量代替,同时这个变量还包含一定成分的噪声,这个变量应该删去。那么哪些变量应该保留,哪些变量应该删去? 在 *PLS* 中的做法是:首先把所有这些变量经线性组合变换成两两正交的矢量,这样这些正交的矢量就没有信息重叠部分了,然后再把这些正交矢量按目标变量的相关性大小排序,与目标变量相关性大的矢量排在前面,逐次小的排在后面,这样最后面的矢量与目标变量的相关性最小,那些小到一定程度的矢量,被认为包含较多的噪声成分,应该删去。

2.3 *PLS2* 算法

这里仅介绍 Y 中的变量个数大于 1 的 PLS2 算法:

(1)$u = Y$ 中的第一列

(2)$w = X^{\mathrm{T}}u/(u^{\mathrm{T}}u)$

(3)$w = w/\|w\|$,归一化

(4)$c = Xw$

(5)$c = Y^{\mathrm{T}}t/(t^{\mathrm{T}}t)$

(6)$c = c/\|c\|$

(7)$u = Yc/(c^{\mathrm{T}}c)$

(8)如果收敛则转到步骤 9,否则转到步骤 2

(9)X 负载矢量:$p = X^{\mathrm{T}}t/(t^{\mathrm{T}}t)$

(10)Y 负载矢量:$q = Y^{\mathrm{T}}u/(u^{\mathrm{T}}u)$

(11)回归(t 对 u):$v = (u^{\mathrm{T}}u)/(t^{\mathrm{T}}t)$

(12)残差:$X = X - tp^{\mathrm{T}}$,$Y = Y - vtc^{\mathrm{T}}$

(13)计算下一维,转到步骤 1,直至求得所需的维数。

3 偏最小二乘法应用于二类水体光谱信息提取

在二类水体的光谱测量中,通常使用多达 256 个通道的地物光谱仪,每条光谱曲线提供 256 个波段的信息,即矩阵 X 有 256 个变量。为了提高偏最小二乘法的计算效率,根据已有的研究结果,光谱曲线的波长范围取 400—800nm,每个波段间隔 5—8nm,这样矩阵 X 变量个数在 50—80 个之间。

3.1 数据预处理

样本数据的预处理分两个方面:一是线性化,二是标准化。通常对水色要素值取对数进行线性化,对光谱反射率进行线性化和标准化。标准化的目的是使样本点的分布结构有利于计算,并尽量避免数据的舍入误差。通过标准化,即变量与均值之差被标准偏差来除,矩阵 X 的列向量的均值为零,方差为 1。

3.2 *PRESS* 判据

普通检验需要二组样本,一是训练(拟合)样

本,二是检验(预报)样本。普通检验需要的样本数较多,当样本较少时就难于进行。目前 *PLS* 最常用的是 *PRESS* 判据:将 m 个样本中的 $m-1$ 个用作训练样本,剩下的一个样本作检验样本。第一次先将第一个样本留下作检验样本,用其余 $m-1$ 个样本建模,然后将检验样本代入模型,可求得一个估计值,记为 $y_{1,-1}$,第二次再将第二个样本留作检验样本,用其余样本建模,将第二个样本代入模型求得估计值 $y_{2,-2}$,如此循环 m 次,每次都留下一个样本作估计,这样可求出第 m 个预报残差值 $y_{m,-m}$,再将这 m 个残差值平方求和,即为 *PRESS*(prediction residual sum of squares):

$$PRESS = \sum_{i=1}^{m} (y_i - y_{i,-i})^2 \qquad (4)$$

PRESS 值越小,表示模型的预报能力越强。*PRESS* 值与普通的残差平方和有些类似,但从使用的角度看,*PRESS* 值非常有用,特别在变量筛选过程中,当往模型中引入变量时,普通的残差平方和(如主成分分析法)会不断减小,因此无法用普通的残差平方和来判断何时停止引入变量,而 *PRESS* 值则开始引入一些变量后会不断变小,当后面引入含噪声较大的变量时 *PRESS* 值会上升,因此 *PRESS* 值是一个较好的判据,我们可以选择 *PRESS* 值取极小值时的变量个数进行建模。

4　应用实例

4.1　南海水体的叶绿素反演

我们选择南海一个航次的现场测量数据进行处理。19 个站位的光谱反射率曲线见图 1。从光谱曲线可以看出,存在两类曲线,那些在 530nm 以后有反射峰的曲线都是二类水体的光谱曲线,其余的为一类水体光谱曲线。该航次的悬浮泥沙测量有问题,且具有悬浮泥沙数据的样本只有 12 个,为避免泥沙建模出现问题,故我们只选择叶绿素作为单目标变量。由于样本数目只有 19 个,我们只选择了 50 个光谱变量,即每隔 8nm 选择一个光谱值,同时删去了 800nm 以后的反射率数据。由于 650nm 以后的光谱反射率数据接近于零,取对数后超出系统的精度范围,因此实际上我们只选择 38 个光谱变量,这样矩阵 $\boldsymbol{X}$ 的维数是 19×38,矩阵 $\boldsymbol{Y}$ 的维数是 19×1。对这两个矩阵进行 *PLS* 方法计算后,我们得到了 *PLS* 因子解释的百分比方差(表 1,为节省篇幅只取了前 15 个因子),从表 1 可以看出,少数的几个因子已经解释了绝大多数的方差,说明光谱变量具有高度的相关性。

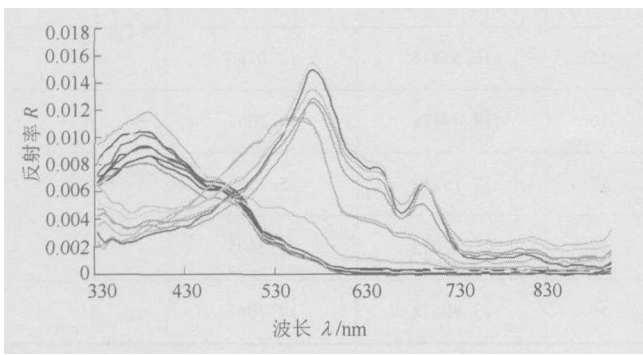

图 1　南海测量的光谱反射率曲线

Fig. 1　Spectral reflectance of South China Sea

至此,偏最小二乘法与主成分分析似乎没有区别。事实上,偏最小二乘法和主成分分析法在矩阵的分解过程中提取得分变量的优化判据是不同的,在主成分分析中从矩阵 $\boldsymbol{X}$ 抽取得到的得分变量的方差取极大值,而在偏最小二乘法中从矩阵 $\boldsymbol{X}$ 抽取得到的得分变量与目标变量的协方差取极大值。偏最小二乘法选取的因子数目既充分解释了矩阵 $\boldsymbol{X}$ 和矩阵 $\boldsymbol{Y}$ 的方差,又不发生过拟合现象。偏最小二乘法采用的是 *PRESS* 判据。从 *PRESS* 值随因子的变化图可以看出(图 2),*PRESS* 值在因子数为 7 时有一个极小值,因此我们取 7 个因子进行建模。事实上,从 *PLS* 解释的方差百分比可以看出(表 1),7 个因子已经解释了 99.9% 的方差。若增加因子数目,对已有的数据会拟合得更好,但预报能力即将下降(即过拟合)。

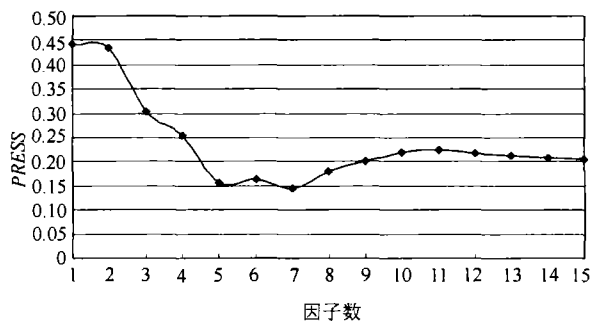

图 2　南海偏最小二乘法模型预报残差平方和随因子数的变化

Fig. 2　The relation of *PRESS* of South China Sea model and the number of latent variables

表1 南海偏最小二乘法模型因子解释的百分比方差

Table 1 Percent variation accounted for by *PLS* factors of South China Sea

因子	因子方差		因变量方差	
	当前方差	累计方差	当前方差	累计方差
1	78.3141	78.3141	85.8805	85.8805
2	6.7948	85.1089	7.9286	93.8091
3	13.4003	98.5092	1.3027	95.1118
4	1.2875	99.7968	2.2436	97.3554
5	0.085	99.8818	1.7875	99.1429
6	0.0174	99.8992	0.4206	99.5635
7	0.0326	99.9318	0.0635	99.6271
8	0.0181	99.9499	0.0936	99.7206
9	0.0174	99.9672	0.0848	99.8054
10	0.0092	99.9764	0.0816	99.887
11	0.0048	99.9812	0.0787	99.9657
12	0.0055	99.9867	0.0219	99.9876
13	0.0063	99.993	0.0101	99.9976
14	0.0026	99.9956	0.0016	99.9992
15	0.0018	99.9974	0.0006	99.9999

表2 南海叶绿素浓度预报值与实测值的比较

Table 2 Comparison of the predicted values and measured values of chl-a in South China Sea

样本	预报值	实测值	相对误差/%	样本	预报值	实测值	相对误差/%
1	0.06653	0.078403	15	10	0.096924	0.100579	4
2	0.098899	0.088998	-11	11	0.093701	0.082475	-14
3	0.117149	0.099593	-18	12	0.317457	0.324207	2
4	0.089791	0.088998	-1	13	0.319687	0.31785	-1
5	0.083634	0.093236	10	14	1.29312	1.2714	-2
6	0.092184	0.088998	-4	15	16.51716	12.0783	-37
7	0.07953	0.086879	8	16	18.49478	19.7067	6
8	0.099919	0.097474	-3	17	17.19913	15.2568	-13
9	0.129581	0.133497	3	18	21.32015	22.2495	4
				19	13.68618	19.7067	31

通过交叉检验,我们可以预报出叶绿素的浓度值。预报值与实测值的关系见表2。可以看出二者

符合得较好,最大相对误差不超 37%,误差直方图见图 3。值得一提的是,针对本例而言,尽管 *PLS* 模型是区域性的,但它对一类水体也是适用的。

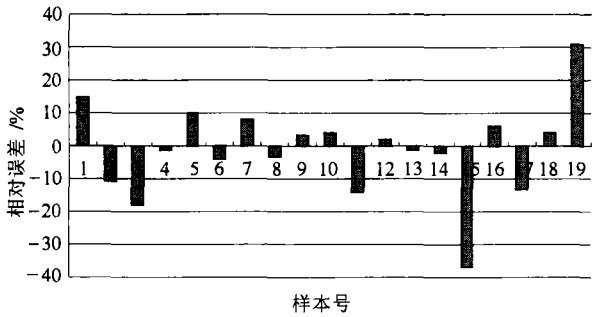

图 3　南海叶绿素反演的相对误差

Fig. 3　The relative error of the predicted chlorophyll-a in South China Sea

4.2　黄海二类水体的叶绿素和泥沙反演

黄海二类水体的样本数目有 14 个,目标变量有叶绿素和泥沙二种。由于样本数过少,为避免建模出现问题,我们只选择了 39 个光谱变量,即每隔 8nm 选择一个光谱值,同时删去了 800nm 以后的反射率数据。这样矩阵 X 的维数是 14×39,矩阵 Y 的维数是 14×2。对这两个矩阵进行 *PLS*2 方法计算后,我们同样得到了 *PLS* 因子解释的百分比方差(表 3)。与南海数据不同的是,39 个光谱变量只用了 13 个因子就全部解释了所有方差。那么建模需要几个因子呢? 我们从 *PRESS* 随因子数的变化关系图(图 4)可以看出,用 5 个因子来建模最理想。我们用 5 个因子模型通过交叉检验来预报叶绿素和泥沙的浓度,预报值和实测值分别见图 5 和图 6。

表 3　黄海偏最小二乘法模型因子解释的百分比方差

Table 3　Percent variation accounted for by *PLS* factors of China Yellow Sea

因子	因子方差		因变量方差	
	当前方差	累计方差	当前方差	累计方差
1	94.2713	94.2713	17.3772	17.3772
2	5.1397	99.411	4.7002	22.0774
3	0.3161	99.7271	39.8802	61.9576
4	0.1477	99.8748	16.1713	78.1289
5	0.0698	99.9446	2.7717	80.9006
6	0.0192	99.9638	5.6467	86.5474
7	0.0066	99.9704	7.7947	94.3421
8	0.0146	99.985	3.1308	97.4729
9	0.0057	99.9907	1.0336	98.5065
10	0.0044	99.9951	0.9106	99.4171
11	0.0025	99.9976	0.3596	99.7767
12	0.0013	99.9989	0.1951	99.9718
13	0.0011	100	0.0282	100
14	0	100	0	100

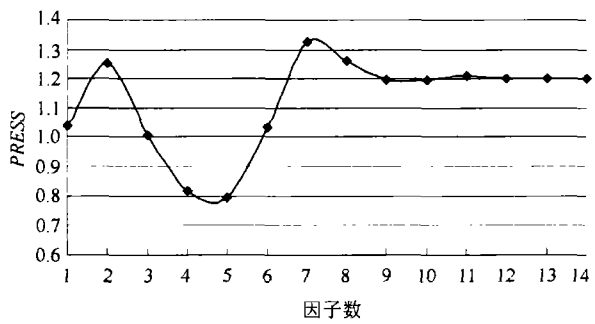

图 4　黄海偏最小二乘法模型预报残差平方和随因子数的变化

Fig. 4　The relation of *PRESS* of China Yellow Sea model and the number of latent variables

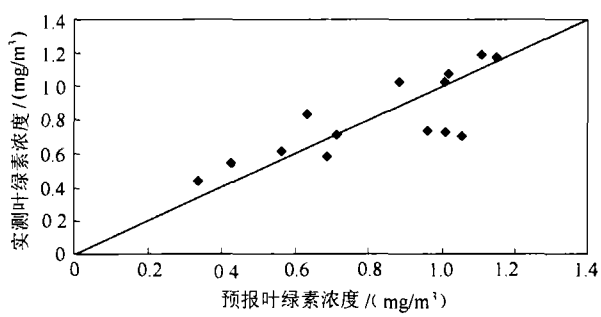

图 5　黄海偏最小二乘法模型反演叶绿素的结果与实测值的比较

Fig. 5　Comparison of the predicted values and measured values of chl-a in China Yellow Sea

表4 偏最小二乘法反演二类水体多成分引起的误差

Table 4 The relative error of the *PLS* predicted values of three components in Case 2 water from simulated spectra

样本	叶绿素			悬浮泥沙			黄色物质		
	设定值 /(mg/m³)	预报值 /(mg/m³)	相对误差/%	设定值 /m⁻¹	预报值 /m⁻¹	相对误差/%	设定值 /m⁻¹	预报值 /m⁻¹	相对误差/%
1	20	19.99	0	10	10.15	-2	1	1.00	0
2	20	20.38	-2	5	4.86	3	0.5	0.50	-1
3	20	20.09	0	2	2.03	-2	0.2	0.20	2
4	20	19.98	0	1	0.97	3	0.1	0.10	-2
5	20	20.07	0	0.1	0.10	-1	0.01	0.01	1
6	10	10.09	-1	10	10.13	-1	0.5	0.50	0
7	10	9.79	2	5	5.05	-1	0.2	0.20	0
8	10	10.02	0	2	2.05	-2	0.1	0.10	4
9	10	9.89	1	1	0.98	2	0.01	0.01	-2
10	10	9.80	2	0.1	0.10	0	1	1.01	-1
11	5	4.94	1	10	9.74	3	0.2	0.20	-2
12	5	4.94	1	5	5.06	-1	0.1	0.10	2
13	5	5.06	-1	2	2.03	-1	0.01	0.01	-2
14	5	4.97	1	1	1.02	-2	1	0.98	2
15	5	5.11	-2	0.1	0.10	1	0.5	0.50	-1
16	2	1.96	2	10	10.03	0	0.1	0.10	1
17	2	2.04	-2	5	4.95	1	0.01	0.01	4
18	2	2.01	-1	2	1.99	1	1	1.01	-1
19	2	2.02	-1	1	0.98	2	0.5	0.48	3
20	2	2.00	0	0.1	0.10	1	0.2	0.20	0
21	1	1.00	0	10	9.98	0	0.01	0.01	-3
22	1	1.03	-3	5	4.97	1	1	0.97	3
23	1	0.98	2	2	2.06	-3	0.5	0.53	-5
24	1	0.98	2	1	0.99	1	0.2	0.20	1
25	5	4.99	0	1.3	1.25	4	0.2	0.21	-7
26	2.5	2.52	-1	0.6	0.62	-3	0.05	0.05	2

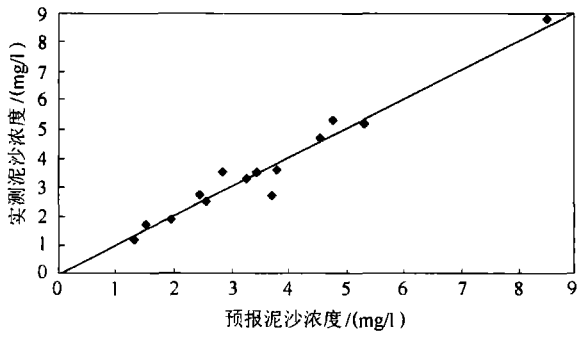

图6 黄海偏最小二乘法模型反演泥沙的结果与实测值的比较

Fig. 6 Comparison of the predicted values and measured values of total suspended matter io China Yellow Sea

4.3 二类水体合成光谱与多成分反演模型的稳定性

为了证明偏最小二乘法比多元线性回归具有更好的模型预报稳定性，我们根据文献[2]合成出26条反射率高光谱曲线，然后随机引入5%的光谱反射率误差。若模型的稳定性不佳，则光谱反射率的微小扰动将导致很大的反演误差。反演误差见表4。从表4的最后二行可以看出，与文献[6]相比，偏最小二乘法建立的模型有更小的反演误差，说明该模型具有更强的稳健性。

5 讨 论

(1)二类水体内部光学性质的复杂性[7]决定了

其成分反演算法具有区域性,可能需要利用高光谱信息。统计方法由于是经验算法,显然具有区域的特点,另一方面,高光谱信息的不同波段之间具有很强的相关性,统计方法如偏最小二乘法和主成分分析等就是可以减少这种相关性的方法之一。与主成分分析相比,偏最小二乘法除考虑光谱信息外,还考虑反演要素的信息。为了考虑反演要素信息,提取的因子之间不再完全正交(可以引入附加算法步骤使因子之间正交,但通常不需要如此)。与多元回归相比[8],偏最小二乘法适用于样本数少而变量多的场合,又不引起大的误差。在南海和黄海实测水体数据(样本数不超过 20 个)中,反演的相对误差均不超过 38%,南海叶绿素浓度的反演误差通常在 10% 左右。对合成光谱的反演误差,更是低于 5%,说明经过 *PRESS* 判据选择的因子模型,具备较强的预报能力,即模型的稳定性较强或模型较稳健。偏最小二乘法建立的模型具有较强的预报能力,并不意味着可以降低对水体光谱测量的要求。任何数据处理方法都不能从一堆垃圾数据中提取出有用的信息。为了有效地提取信息,光谱测量和水色要素测量结果都必需准确。关于如何准确测量水体光谱,文献[9,10]作了非常详细的叙述。比如,黄海水体的水色要素数据质量较差,模型的预报误差普遍较大(尽管不超过 38%)。

(2)数据的预处理对偏最小二乘法建模影响较大,特别是当水体成分浓度变化范围较大时必须线性化。无论是主成分分析还是偏最小二乘法,都是将数据投影变换后,考察因子与浓度之间的线性关系。如果关系是非线性的,很可能导致无法提取水色中蕴藏的信息。我们的经验是,对二类水体光谱而言,取对数是有效的线性化方法。通过对数变换,数据的分布比较均匀,可以较大地改善预报结果。如南海叶绿素数据最大值与最小值之间相差 300 多倍,进行对数变换后模型的预报精度显著提高。另一方面,建模时所选择的训练数据集在浓度范围上尽可能分布得均匀,以减小预报误差。如南海的叶绿素数据,预报误差较大的样本出现在样本稀少的浓度区间上。偏最小二乘法会将过于稀少的样本数据当作奇异数据处理,虽然我们可以通过"引入权重"的方式减少这些"奇异"样本的预报误差,但这毕竟是人为的方法。

(3)偏最小二乘法还能提供变量的载荷,光谱波段变量在某个因子上的载荷越大,说明该变量与那个因子越"相同",因此,载荷可视为波段变量与因子的相关性。当然,偏最小二乘法主要侧重于预报,尤其适合样本少变量多,而且这些变量又是高度相关的场合。如果光谱变量少,变量之间的相关性又小,样本又多,那么偏最小二乘法的效果与多元线性回归相似。理论上,如果得分变量的个数与光谱变量的个数相同,则偏最小二乘法的效果与多元线性回归的相同。如果叶绿素 - a 和总悬浮物之间具有一定的相关性[11],普通的统计方法可能导致定量反演的结果含糊不清,这时能处理目标变量之间协同变化的 *PLS*2 方法就是一种更好的选择。

6　结　论

在二类水体的水色因子信息提取中,高光谱水色遥感是长远的发展趋势。高光谱曲线的不同波段之间具有很强的相关性,偏最小二乘法能高效消除光谱波段之间的相关性,适合样本少而光谱变量多的回归场合。当只有一个目标变量如叶绿素需要回归时,*PLS* 通常能给出与主成分回归相同的定量预报结果,但需要更少的因子,因此偏最小二乘法所得到的因子更易于解释。作为一种经验算法,偏最小二乘法要求训练样本具有代表性,训练数据集在浓度范围上要尽可能分布得均匀,以减小预报误差。作为一种区域性的算法,偏最小二乘法可同时适用于一、二类两种水体的信息提取。偏最小二乘法在二类水体信息提取中的优势有待于今后大量现场测量数据的进一步验证。

参 考 文 献(References)

[1] Tang J W, Wang X M, Song Q J *et al*. 14th National Academic Meeting of Remote Sensing Technology[C]. 2003. 唐军武,王晓梅,宋庆君等. 第十四届全国遥感技术学术交流会[C], 2003, 10,青岛.

[2] Tang J W. The Simulation of Marine Optical Properties and Color Sensing Models [D]. Beijing: Institute of Remote Sensing Application, Chinese Academy of Sciences, 1999. [唐军武,海洋光学特性模拟与遥感模型[D]. 博士论文,1999.]

[3] Cao W X, Zhong Q Y, Yang Y Z. Principal Component Analysis for Ocean Color Remote Sensing in South China Sea[J]. *Journal of Remote Sensing*, 1999, **3**(2):112—115. [曹文熙,钟其英,杨跃忠. 南海水色遥感的主因子分析[J]. 遥感学报,1999, **3**(2):112—115.]

[4] Neumann A, Krawczyk H. Principal Component Inversion, IOCCG Training Course on Remote Sensing of Ocean Color[R]. Ahmedabad, India, February, 2001.

[5] Wang H W. Partial Least Squares Regression and Its Applications [M]. Beijing: National Defense Press, 2000. [王慧文. 偏最

小二乘回归方法及应用[M]. 北京:国防工业出版社,2000.]

[6] Tang J W, Tian G L. Ocean Color Analysis and an Algorithm for the Retrieval of Multiconstituents Based on Remote Sensing Reflectance [J]. *Journal of Remote Sensing*, 1997, **1**(4): 252—256. [唐军武,田国良. 水色光谱分析与多成分反演算法[J]. 遥感学报,1997,**1**(4):252—256.]

[7] Lahet F, Ouillon S, Forget P. A Three-component Model of Ocean Color and Its Application in the Ebro River Mouth Area [J]. *Remote Sensing of Environment*, 1999, **72**: 181—190.

[8] Sathyendranath S, Prieur L, Morel A. A Three-component Model of Ocean Color and Its Application to Remote Sensing of Phytoplankton Pigments in Coastal Waters[J]. *Int. J. Remote Sensing* 1989, **10** (8): 1373—1394.

[9] Tassan S. Local Algorithms Using SeaWiFS Data for the Retrieval of Phytoplankton Pigments, Suspended Sediment, and Yellow Substance in Coastal Waters[J]. *Applied Optics*, 1994, **33**(12): 2369—2378.

[10] Li T J, Tang J W, Chen Q L, *et al*. A Method for Measuring Water-Leaving Radiance Using Photometer[J]. *Journal of Tropical Oceanography*, 2001, **20** (4): 56—60. [李铜基,唐军武,陈清莲等. 光谱仪测量离水辐射亮度的方法[J]. 热带海洋学报,2001,**20**(4):56—60.]

[11] Gong C L, Fan W. Algorithm for Case2 Waters of Remote Sensing of Ocean Color[J]. *Marine Science Bulletin*, 2002, **21**(2): 77—83. [巩彩兰,樊伟. 海洋水色卫星遥感二类水体反演算法的国际研究进展[J]. 海洋通报,2002,**21**(2):77—83.]

Retrieval of Oceanic Color Constituents from Case Ⅱ Water Reflectance by Partial Least Squares Regression

YANG Yan-ming[1], LIU Zhen-wen[1], CHEN Ben-qing[1], TANG Jun-wu[2]

(1. *Third Institute of Oceanography, State Oceanic Administration, Xiamen* 361005, *China*;
2. *National Satellite Ocean Application Service, Beijing* 100081, *China*)

Abstract: It is generally recognized that Case 2 waters are more complex than Case 1 waters in their composition and optical properties. The standard algorithms (usually band ratio) in use today for chlorophyll retrieval from spectral data break down in Case 2 waters. Hyperspectral ocean color sensing may be necessary for Case 2 waters' constituents retrieval. However, hyperspectral data are usually highly correlated and statistical algorithms such as principal component inversion have been employed in ocean color sensing. In the present paper the principle, algorithm and advantage of another statistical algorithm-partial least squares regression (*PLS*) are briefly described. Then *PLS* is applied to the retrieval of oceanic color constituents from China Yellow Sea and South China Sea field reflectances, which are typical of Case 2 waters. Cross-validation of *PLS* analysis shows that the retrieval accuracy is good and the predicted relative error of chlorophyll-a is less than 37%. In order to check the robusticity of the *PLS* inversion model, *PLS* is also applied to the retrieval of oceanic color constituents from computed reflectances to which 5% noise is added randomly. The cross-validation results of *PLS* analysis on simulated data show that the model is robust and the predicted relative error of the three components (chlorophyll-a, Total Suspended Matter and Yellow Substance) is less than 5%. Pre-processing of data is essential for the constituents' concentration ranging over several magnitudes. As an empirical algorithm, the training data set for *PLS* should be typical that the data points distribute uniformly in the concentration range. It is suggested that *PLS* be suitable for the regression problems which have a few observations but a lot of spectra variables, e. g. the retrieval of oceanic color constituents from Case 2 water reflectance.

Keywords: case Ⅱ water spectra; oceanic color constituents' retrieval; partial least squares regression