

目 录

A 气候数据分析	1
A.1 普通相关系数	1
A.2 秩相关系数	12
A.3 谐波分析	14
A.4 功率谱分析	22
A.5 交叉谱分析	27
A.6 时间序列滤波分析	32
A.7 EOF分析	42
A.7.1 REOF分析	52
A.8 SVD分析	52

图 目 录

A.1 北京夏季气温和降水的相关系数	8
A.2 ENSO指数与全球温度的滞后相关	9
A.3 虚假的相关信息	11
A.4 沙尘暴日数与850hPa高度高频变率方差的相关场	12
A.5 北京月平均气温的谐波分析	15
A.6 北京夏季平均气温的谐波分析	21
A.7 北京夏季气温各谐波的方差贡献	21
A.8 500hPa高度场超长波分析	23
A.9 Nino3区海温的功率谱分析	29
A.10 全球副高指数和NinoC区SST交叉谱分析	32
A.11 北京夏季气温简单滑动平均	33
A.12 1-2-1加权滑动平均的频率响应函数	35
A.13 9点二项式权重系数	36
A.14 高斯滤波的频率响应函数	39
A.15 不同参数 n 的权重系数的频率响应函数	40
A.16 北半球1月份MSLP距平EOF分析第一模态	47
A.17 不同资料处理方式EOF结果的差别	49
A.18 北半球春季NDVI和气温SVD分析结果	55

表 目 录

A.1 Z 统计表	5
A.2 相关系数检验表	6
A.3 有效自由度估计	7
A.4 北京气温年循环的谐波分析	16
A.5 北京夏季气温谐波F检验值	20
A.6 不同落后步长对功率谱分析结果的影响	28
A.7 滑动平均计算实例	33
A.8 二项式权重系数	37
A.9 高斯滤波权重系数	38

A

气候数据分析

A.1 普通相关系数

计算方法

为了度量两个变量之间的线性关联程度，常用的指标是普通相关系数，即Pearson相关系数。任意两个变量 x 和 y ，样本(时间长度)为 n ，其标准差分别为 σ_x 和 σ_y 则它们之间的Pearson相关系数的计算公式是：

$$r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right) \quad (\text{A.1})$$

另外一种计算公式是：

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (\text{A.2})$$

r 可以客观的度量两个因子之间的关联程度。 $|r| \leq 1$ ，正相关表示两者之间变化性质相同，负相关表示两个变量之间变化性质相反。例如图A.1是北京夏季降水和气温的时间序列，通常夏季降水如果偏多，则地面接受的太阳辐射减少，而且由于地表和土壤水份蒸发也消耗大量热量，也会降低地面气温。1951 – 2002年共52年二者的相关系数是-0.433，也就很好的说明这一点。

协方差与方差

两个变量之间的协方差定义为:

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

显然一个变量与自己的协方差, 就是其方差:

$$S_{xx} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \sigma_x^2$$

通常大家常用简单的回归方程来表示两个变率之间的关系, 某种程度上是与相关系数等价的。如果用北京夏季降水量的变化来拟合夏季气温的年际波动, 即

$$y = a + bx \quad (\text{A.3})$$

这里 a 是截距, b 是回归系数, y 是气温, x 是降水量。利用最小二乘法, 可以得到:

$$b = \frac{\sum_{i=1}^n y_i x_i - (\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)/n}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n} \quad (\text{A.4})$$

$$= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})} \quad (\text{A.5})$$

$$= \frac{S_{xy}}{S_{xx}} \quad (\text{A.6})$$

截距 a 可以由下式计算得到:

$$\begin{aligned} a &= (\sum_{i=1}^n y_i)/n - b(\sum_{i=1}^n x_i)/n \\ &= \bar{y} - b\bar{x} \end{aligned}$$

S_{xy} 表示 xy 之间的协方差, S_{xx} 表示 x 与自己的协方差, 即方差。根据相关系数

的定义(A.2式), 可以得到:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (\text{A.7})$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (\text{A.8})$$

$$= \frac{S_{xy}}{\sigma_x \sigma_y} \quad (\text{A.9})$$

由A.6式和A.9式, 可以得到:

$$b = r \sqrt{\frac{S_{yy}}{S_{xx}}} \quad (\text{A.10})$$

$$= r \frac{\sigma_y}{\sigma_x} \quad (\text{A.11})$$

可见回归系数 b 只与温度和降水的相关系数以及各自的标准差有关。如果进行回归分析之前, 将温度和降水时间序列已经标准化, 即 $\sigma_x = \sigma_y = 1$; $\bar{x} = \bar{y} = 0$; 那么温度和降水的关系简化为: $y = rx$ 。由相关系数, 可以很方便的估计出北京夏季降水量和温度两个变量之间的线性关系系数(b):

$$b = r \frac{\sigma_y}{\sigma_x} = -0.433 \times \frac{0.8679}{190.120} = -0.00198^\circ\text{C}/\text{mm}$$

因此当降水量每增加100mm, 气温大致下降0.2°C。

那么用降水解释的气温方差是多少呢? 根据方程(A.3), 将降水量代入方程可以计算出相应的温度时间序列。最后得到这个计算出来的温度序列的方差是0.1414(°C)², 而实际观测的气温序列(即图A.1(a))的方差是0.7533(°C)²; 因此, 降水的变化可以解释气温年际变化方差的值是0.1414/0.7533 × 100% = 18.8%。实际上可以证明, 解释的方差比例可以简单地相关系数估计出来:

$$\frac{\sigma_y^2}{\sigma_y^2} = r^2$$

即 $r^2 \times 100\% = (-0.433)^2 \times 100\% = 18.8\%$ 。上面的例子说明可以直接用相关系数来判断两个变量之间相关、置信度以及他们之间的方差解释率。

显著性检验

为检验相关系数有别于 0 的置信度，可以用 Z 统计量检验，即：

$$Z = \frac{\sqrt{n-3}}{2} \ln \frac{1+r}{1-r} \quad (\text{A.12})$$

上面的计算中北京夏季降水量和温度的相关系数是 -0.433 ， $n = 52$ ，代入公式(A.12)，得到：

$$Z = \frac{\sqrt{52-3}}{2} \ln \frac{1+0.433}{1-0.433} = 3.245$$

再查 Z 分布表，即累积标准化正态分布统计表。以方便大家查询在表A.1中列出。一般统计书中也有此表。对于单侧0.01显著水平检验，由于 Z 统计量两侧是对称的， $1 - 0.01 = 0.99$ ，在表A.1中找到对应于0.99累积概率的 Z 值，得到 $Z_{0.01} = 2.33$ ，该值小于计算得到的 Z 值。因此，上述相关系数是显著的。

另外，也可以用 t 统计量检验。即：

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (\text{A.13})$$

得到

$$t = \frac{0.433\sqrt{52-2}}{\sqrt{1-0.433^2}} = 3.397$$

查 t 分布表，对于双侧0.005显著水平检验，样本数为52时， $t_{0.005} = 2.678$ ，小于计算得到的 t 值。也可得到相关系数是显著的结论。

在样本数比较少的情况下(n 低于30)，应该用 t 统计量检验。大样本情况下， t 统计量分布逼近正态分布。实际应用中为方便检验，常常根据(A.12)式或者(A.13)式将不同自由度和置信度情况下的相关系数阈值事先计算出来，编成检验表，需要时直接查表比较就可以了。表A.2给出了不同置信度时，根据双测 t 统计量检验计算的相关系数阈值。分析中得到的相关系数的绝对值高于阈值时，相关系数才是显著的。

自由度的估计

简单估计：随机样本数减2，即 $\nu = n - 2$ 实际上气候变量的一个突出特点就是具有红噪声谱，即不同时间的数据之间不是完全独立的(不是随机的)。气候变量某

表 A.1: 累积标准化正态分布函数(Z 统计量)

$$(F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{z^2}{2}} dz)$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

表 A.2: 不同置信度相关系数阈值(双侧 t 统计量检验).

自由度	90%	95%	99%	99.9%	自由度	90%	95%	99%	99.9%
1	0.920	0.954	0.986	0.997	36	0.271	0.320	0.412	0.511
2	0.833	0.891	0.956	0.987	37	0.267	0.316	0.407	0.505
3	0.758	0.829	0.919	0.970	38	0.264	0.312	0.402	0.499
4	0.697	0.774	0.880	0.948	39	0.260	0.308	0.397	0.494
5	0.646	0.727	0.843	0.924	40	0.257	0.304	0.392	0.488
6	0.605	0.685	0.808	0.899	41	0.254	0.300	0.388	0.483
7	0.570	0.650	0.775	0.875	42	0.251	0.297	0.384	0.478
8	0.540	0.619	0.746	0.851	43	0.248	0.294	0.379	0.473
9	0.514	0.592	0.719	0.828	44	0.245	0.290	0.375	0.468
10	0.491	0.567	0.695	0.807	45	0.243	0.287	0.372	0.464
11	0.471	0.546	0.672	0.786	46	0.240	0.284	0.368	0.459
12	0.453	0.526	0.652	0.767	47	0.238	0.281	0.364	0.455
13	0.437	0.509	0.633	0.749	48	0.235	0.278	0.361	0.451
14	0.423	0.493	0.615	0.732	49	0.233	0.276	0.357	0.446
15	0.410	0.478	0.599	0.715	50	0.230	0.273	0.354	0.442
16	0.398	0.465	0.584	0.700	55	0.220	0.261	0.338	0.424
17	0.387	0.453	0.570	0.686	60	0.211	0.250	0.325	0.407
18	0.377	0.441	0.557	0.672	65	0.203	0.240	0.312	0.393
19	0.367	0.431	0.545	0.659	70	0.195	0.232	0.302	0.379
20	0.358	0.421	0.533	0.647	75	0.189	0.224	0.292	0.367
21	0.350	0.411	0.522	0.635	80	0.183	0.217	0.283	0.357
22	0.343	0.403	0.512	0.624	85	0.177	0.211	0.275	0.347
23	0.336	0.395	0.503	0.613	90	0.173	0.205	0.267	0.337
24	0.329	0.387	0.493	0.603	95	0.168	0.200	0.260	0.329
25	0.322	0.380	0.485	0.594	100	0.164	0.195	0.254	0.321
26	0.316	0.373	0.476	0.585	110	0.156	0.186	0.242	0.307
27	0.311	0.366	0.469	0.576	120	0.150	0.178	0.232	0.294
28	0.305	0.360	0.461	0.567	130	0.144	0.171	0.223	0.283
29	0.300	0.354	0.454	0.559	140	0.139	0.165	0.215	0.273
30	0.295	0.349	0.447	0.552	150	0.134	0.159	0.208	0.264
31	0.291	0.343	0.441	0.544	160	0.130	0.154	0.202	0.256
32	0.286	0.338	0.434	0.537	170	0.126	0.150	0.196	0.249
33	0.282	0.333	0.428	0.530	180	0.122	0.145	0.190	0.242
34	0.278	0.329	0.423	0.523	190	0.119	0.142	0.185	0.236
35	0.274	0.324	0.417	0.517	200	0.116	0.138	0.181	0.230

表 A.3: 有效自由度(ν)与样本数(n)的比值随自相关大小($r(\Delta t)$)的变化

	$r(\Delta t)$									
	<0.16	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
ν/n	1	0.81	0.60	0.46	0.35	0.26	0.18	0.11	0.05	据A.15式
ν/n	1	0.92	0.84	0.72	0.60	0.47	0.34	0.22	0.11	据A.16式

一时刻的状况对后面的状况是有影响的。因此，序列的有效自由度要比 $n - 2$ 要小。这会影响对相关系数信度的估计和假设结论的判断。很多气候变量有很强的持续性或者很高的自相关，例如海温。因此进行相关系数的显著性检验时，需要首先对时间序列的有效自由度进行估计。

估计有效自由度的方法有很多。红噪声时间序列的自相关系数随落后时间步长减少，自相关系数越大则独立样本数(有效自由度)越小。Dawdy和Matalas(1964)曾给出有效自由度(ν)与样本数(时间序列长度 n)之间的计算公式：

$$\frac{\nu}{n} = \frac{[1 - r_x(\Delta t)r_y(\Delta t)]}{[1 + r_x(\Delta t)r_y(\Delta t)]} \quad (\text{A.14})$$

Leith(1973)指出有效自由度(ν)与样本数(时间序列长度 n)之间有如下关系：

$$\frac{\nu}{n} = -\frac{1}{2}\ln[r(\Delta t)] \quad (\text{A.15})$$

Bretherton等(1999)给出的另外一种计算方法是：

$$\frac{\nu}{n} = \frac{[1 - r(\Delta t)^2]}{[1 + r(\Delta t)^2]} \quad (\text{A.16})$$

表A.3中列出了上述两种方法对在不同自相关情况下有效自由度与样本数之间的估计值。实际计算中可以取 x_i 和 y_i 各自有效自由度的平均值，或者更严格一些取其中的极大值。如对北京夏季降水和气温时间序列分别估计其有效自由度。因为气温落后步长为1时的自相关 $r(\Delta t) = 0.2973$ ；根据式A.15估计的气温的有效自由度是：

$$\nu_T = -n\frac{1}{2}\ln[r(\Delta t)] = -52\ln(0.2973) = 32$$

因为降水时间序列的 $r(\Delta t) = 0.043$ ；所以 $\nu_P = n = 52$ 。那么判断降水与温度之间相关系数时的有效自由度可以取 $\nu = (32 + 52)/2 = 42$ 。查表A.2可知 -0.433 的

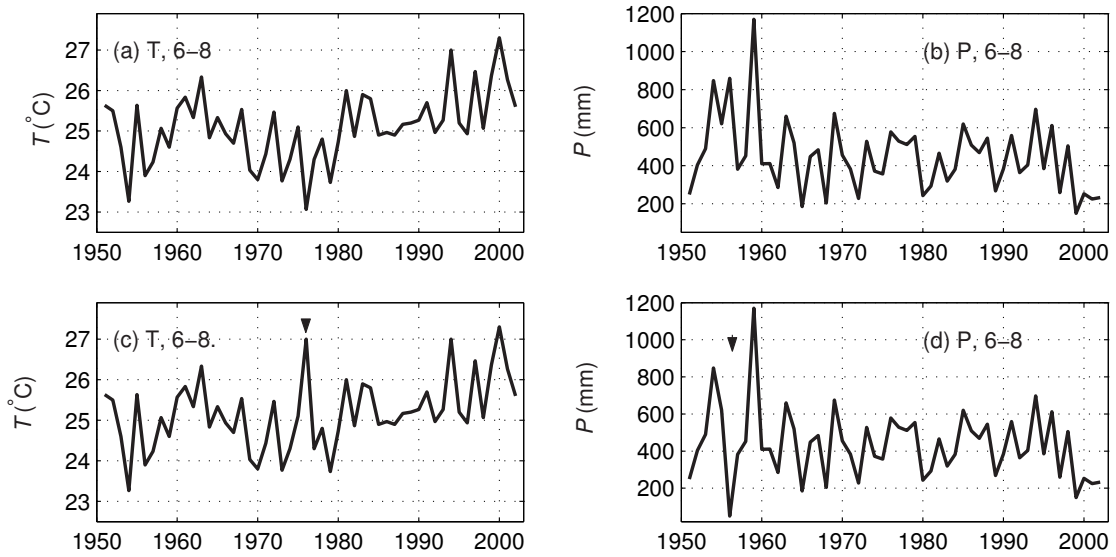


图 A.1: 北京夏季气温(a)和降水(b)的时间系数。(c)和(d)分别包含一个野值(outlier, 图中用箭头标出)。Pearson相关系数对野值很敏感, $r(a, b) = -0.433$; $r(c, d) = -0.255$

相关系数可以达到99%置信度。如果严格一些取 $\nu = 32$, 则置信度略低于99%。如果用公式A.16计算, 则气温序列的有效自由度为

$$\nu_T = n \frac{[1 - r(\Delta t)^2]}{[1 + r(\Delta t)^2]} = 52 \times \frac{[1 - 0.2973^2]}{[1 + 0.2973^2]} = 43$$

。

滞后相关

如果两个时间序列包含有行波, 或者任何随时间传播的信号, 简单地计算他们之间的相关系数是不恰当的, 必须要考虑到他们之间的位相差。如果信号正好是同位相的话, 则计算的相关系数接近+1.0; 如果位相差为180度, 则相关系数接近-1.0; 当位相差为90度时, 则相关系数接近 0.0。为解决上述问题, 常常需要计算滞后相关系数, 即考虑信号的时间差。反过来, 也可以通过计算时间序列之间的不同时间滞后相关, 可以发现他们之间的可能的位相差。气候系统中很多的要

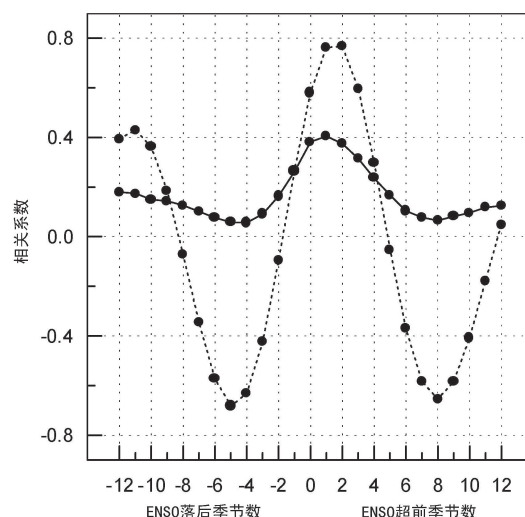


图 A.2: 全球平均温度与ENSO指数的交叉相关系数, 实线为根据原始数据计算, 虚线为带通滤波后计算结果。1880 – 1997年资料。据龚道溢和王绍武, 1999

素之间的关系和联系, 都是有时间滞后的。下面给出一个ENSO与对全球温度关系的例子。

在全球气候系统中, ENSO占有重要地位, 其变化对全球气温有显著的影响。全球面积加权平均气温(取 80°N – 60°S 范围, 约占地球表面积的93%, 1880 – 1997年)与ENSO指数的同时相关系数为0.38。但是相关系数随时间滞后有明显变化, 当温度落后ENSO指数一个季度时, 相关系数最大, 达0.41。当然, 温度的变化不仅有突出的年际变化, 而且还有显著的年代际变化和长期趋势。而ENSO指数主要是2 – 7年左右的准周期变化。所以为了检测温度中的这种ENSO高频部分信号, 分别对温度和ENSO指数做带通滤波处理, 将包括从准两年振荡到7年左右周期的部分保留下来。滤波后的高频部分两者的同时相关系数达到了0.58, 而且是温度落后ENSO指数2个季相关最大, 提高到了0.77(图A.2)。其次, 热带地区与热带外地区也表现出一定的差别。ENSO指数与热带地区(20°N – 20°S)温度相关系数最高, 中、高纬度最大相关系数则明显下降。从最大相关的滞后时间看, 热带地区温度落后ENSO变化1个季, 而热带外地区则落后2 – 3个季度。因此, 这个例子较好地说明了ENSO对全球温度影响的传播及其时间滞后。

缺点

优点是定量描述变量之间的关系，简单直观，容易解释。

但是存在一些缺点，在实际应用中必须注意。图A.3给出了几种典型情况，来说明普通相关系数的局限性。(b)中说明当数据中包含两组或者更多组不同性质的数据时，不能混在一起同时处理，需要对不同的组分别对待；(c)说明如果包含非线性关系，用反映线性关系的相关系数是不合适的。实际应用中一些变量如风力与风速之间的关系就是非线性的；(d)说明Pearson相关系数对野值的影响是很敏感。

为了说明相关系数对野值的敏感性，我们可以用图A.1中北京夏季降水和气温序列来做例子进行分析，图中除了实际观测的原始值序列外，还人为地将气温和降水各一年的数据加一个很大的偏差值，图中分别以箭头标出。各加了一个异常值以后，两个时间序列的相关系数是 $r(c, d) = -0.255$ ，远低于原始值计算的相关 $r(a, b) = -0.433$ 。查表A.2可知，后者置信度达99%；而前者则达不到95%。

因此，进行相关系数分析之前需要首先对数据是否包含不同性质的数据(或多组数据)，是否有野值，是否包含非线性关系，是否有时间滞后现象等等特征做一些初步分析，再根据实际情况作出相应的处理。

线性化处理。

相关场的信度检验

气候数据不仅在时间上前后有联系，在空间上也有很大的连续性。即空间上也不是完全独立的。因此，空间场的相关是否显著还需要其他要求。目前缺乏严格检验方法。

通常计算北京夏季降水与大气环流场的相关。局地计算，判断其置信度。空间分布上连续。足够多。图A.4给出了一个相关场的例子。是内蒙古地区26个站平均春季沙尘暴日数与850hPa位势高度高频变率方差的相关系数。相关系数超过95%水平的区域用阴影标出。天气尺度变率方差定义为： $\overline{\Phi'^2}^{1/2}$ ， Φ' 为时间尺度小于7天的波动。显著区域位置集中，面积很大。可以看出天气尺度的波动是影响北方沙尘暴活动的一个重要因子。从显著区域的空间特征上还可以看到主要天气系统活动的路径呈南北走向。

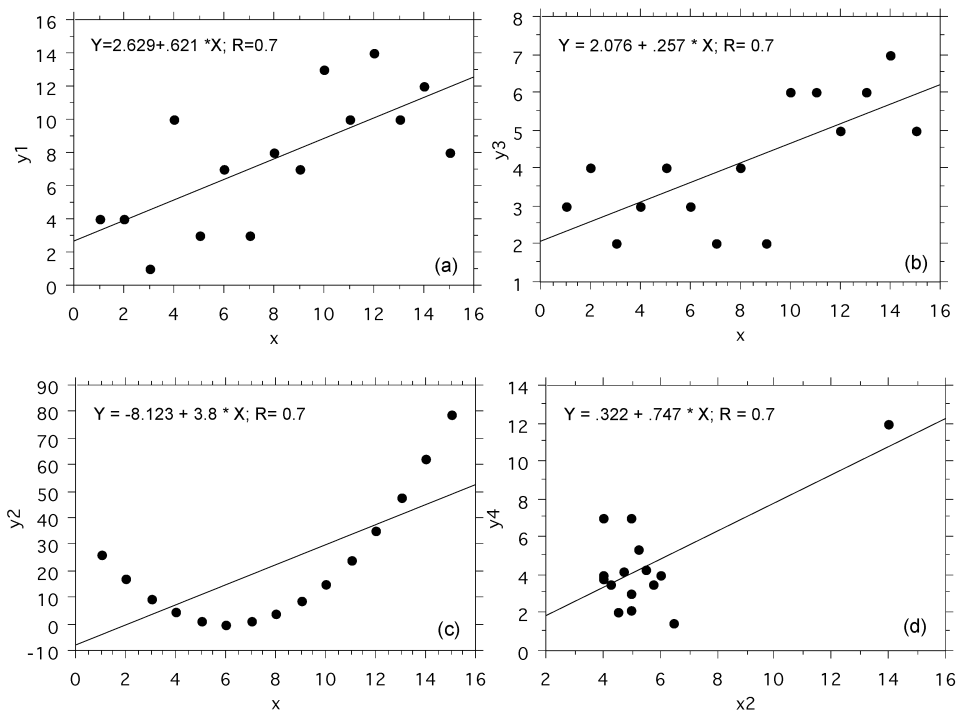


图 A.3: 截然不同意义的数据，得到同样的相关系数。(a)随机分布样本;(b)样本数据包含不同性质的2组数据;(c)非线性关系;(d)数据包含野值。据 D. L. Hartmann, 2002

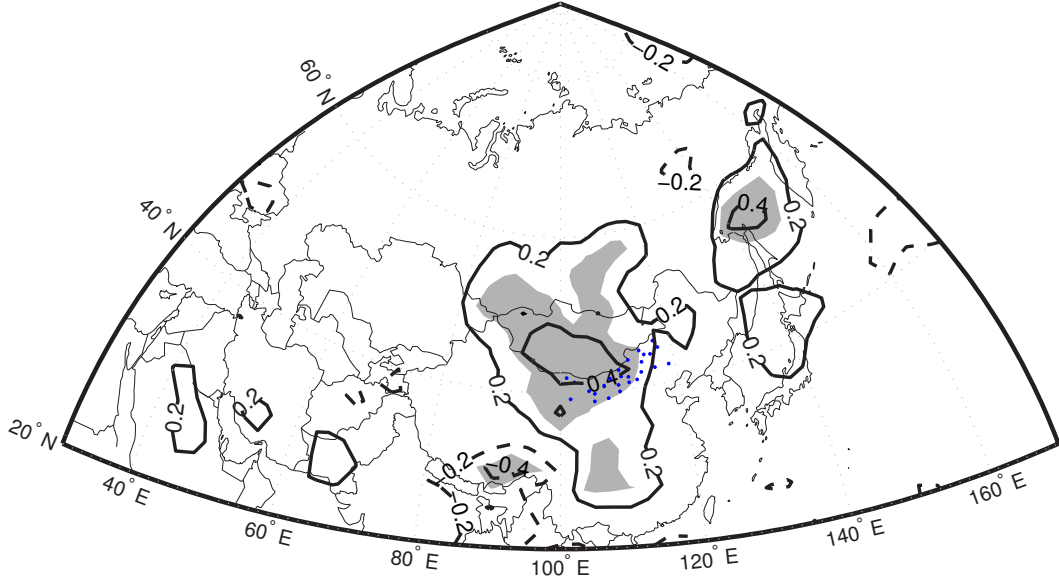


图 A.4: 内蒙26站(圆点标出) 平均春季沙尘暴日数与850hPa位势高度高频变率方差的相关系数。1962-2002年资料计算。

A.2 秩相关系数

计算方法

Spearman秩相关系数，鲁棒(robust)性比较好，对野值不很敏感。任意两个变量 x_i 和 y_i ，样本(时间长度)为 n ，分别将 x_i 和 y_i 按各自大小排序，各自序号分别记为 D_x^i 和 D_y^i ，那么序号之间的普通相关系数就是Spearman秩相关系数(r_s)。即：

$$r_s = \frac{\sum_{i=1}^n (D_x^i - \overline{D_x})(D_y^i - \overline{D_y})}{\sqrt{\sum_{i=1}^n (D_x^i - \overline{D_x})^2 \sum_{i=1}^n (D_y^i - \overline{D_y})^2}} \quad (\text{A.17})$$

可以证明：

$$r_s = 1 - \frac{6 \sum_{i=1}^n (D_x^i - D_y^i)^2}{n(n^2 - 1)} \quad (\text{A.18})$$

公式(A.18)因为简便，所以在实际分析中应用更为广泛一些。

信度检验

秩相关系数的显著性可以利用 Z 统计量来检验。

$$Z = \frac{r_s}{\sqrt{1/(n-1)}} \quad (\text{A.19})$$

同样以北京夏季降水和气温为例子(图A.1), 对于不包含野值的原序列(图A.1a,b), 秩相关系数 $r_s(a, b) = -0.470$;与普通相关系数 -0.433 接近。计算 Z 值:

$$Z = \frac{0.470}{\sqrt{1/(52-1)}} = 3.356$$

给定显著水平0.05, 查表A.1, 找到对应于 $1 - 0.05 = 0.95$ 的累积概率的 Z 值, 得到 $Z_{0.05}$ 的阈值是1.645; 该值低于计算的 Z 值。因此, 秩相关系数是显著的。

需要指出的是, 对于包含野值的序列(图A.1c,d), 前面一节中二者之间的普通相关系数通不过95%信度检验。如果计算其Spearman秩相关则 $r_s = -0.304$, 计算 Z 值:

$$Z = \frac{0.304}{\sqrt{1/(52-1)}} = 2.171$$

该值大于 $Z_{0.05} = 1.645$ 。因此虽然有野值的影响, 秩相关仍然达到95%置信度。由此可见秩相关对野值不如普通相关系数那样敏感。实际工作中, 如果怀疑资料序列中存在异常数据, 而又没有办法对这些异常数据进行排除或者订正的情况下, 用秩相关系数进行分析比用普通相关系数更为稳妥。

A.3 谐波分析

谐波分析又称调和分析(harmonic analysis)，指用三角函数来拟合数字信号或数字序列。根据拟合函数可以对不同的信号周期，位相及振幅的情况进行了解。

简单谐波估计

气候变量时间序列中包含多种时间尺度的变化，因此经常需要考虑如何反映不同时间尺度的变化。例如每天的观测气温多年积累下来，就会有一个连续的日平均气温序列，如果资料从1991年到2000年10就会有3652天。这个时间序列中包含了多种尺度的变化，其中最突出的是年循环，即365日或者是12个月的周期。夏季气温高，冬季气温低，从冬季到夏季再到冬季，是一个循环，每年的季节循环是非常规律的，可以近似用正弦或余弦函数来表示即

$$y_t = \cos(\alpha)$$

t 是时间，正弦函数和余弦函数只是位相上相差 90° ，有 $\cos(\alpha) = \sin(\alpha + \frac{\pi}{2})$ ，二者本质上是等价的。不过因为当 $\alpha = 0$ 时余弦函数等于1，对于 $\alpha \neq 0$ 的情况可以看成是有位相差，便于解释。因此，实际计算中多用余弦函数。图A.5是北京1951—2002年平均每月气温，如果要用余弦函数来拟合温度的变化，要考虑的问题包括：

- 将时间1到12月处理成弧度的变化，因为一个季节循环可以看出是一个完整的周期即 2π ，那么每个月相对应的弧度应该是 $\frac{2\pi t}{n} = \frac{2\pi t}{12}$ ， $t = 1 \dots 12$ ；
- 余弦函数平均值为0，最大值为+1，最小值为-1。所以必须对月平均温度进行处理，将其平均值提取出来使得其变化正负对称，得到平均值是 12.04°C 。余弦函数前还要乘上一个系数以保持其变化幅度与真实情况相符，7月气温最高比年平均气温高出 14.11°C ，1气温最低，比年平均低 16.065°C 。取二者绝对值的平均得到振幅为 $C = \frac{14.11+16.065}{2} = 15.088$
- 最后还要考虑到月平均气温最大值出现在7月份，所以要加上一个位相差，即 $\phi = \frac{2\pi \times 7}{n} = \frac{2\pi \times 7}{12} = \frac{7\pi}{6}$ 。

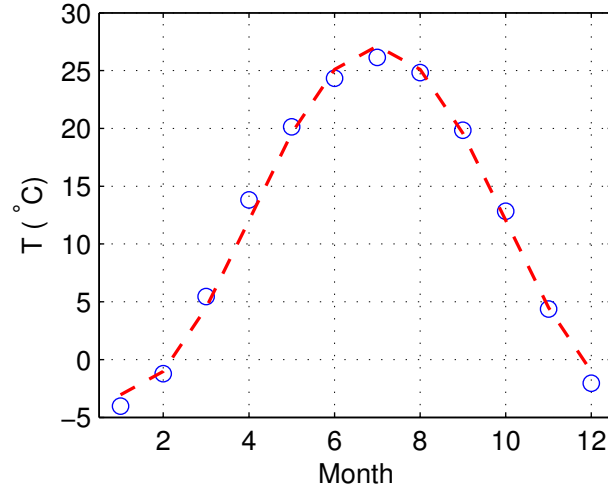


图 A.5: 北京月平均气温的谐波分析。圆点表示1951—2002年平均月平均气温，虚线为根据谐波分析得到的余弦函数计算的温度，即 $y_t = 12.04 + 15.088 \times \cos(\frac{2\pi t}{12} - \frac{7\pi}{6})$

这样就可以将气温的年循环表示为

$$y_t = \bar{y} + C \cos(\frac{2\pi t}{n} - \phi) \quad (\text{A.20})$$

即

$$y_t = 12.04 + 15.088 \times \cos(\frac{2\pi t}{12} - \frac{7\pi}{6}) \quad (\text{A.21})$$

上面的例子是直观的判断和计算，实际计算中对位相和振幅有客观准确的计算方法。根据A.20式，有

$$y_t = \bar{y} + C \cos(\frac{2\pi t}{n} - \phi) \quad (\text{A.22})$$

$$= \bar{y} + A \cos(\frac{2\pi t}{n}) + B \sin(\frac{2\pi t}{n}) \quad (\text{A.23})$$

其中 $A = C \cos(\phi)$, $B = C \sin(\phi)$, 显然 $C = \sqrt{A^2 + B^2}$ 。根据最小二乘法可以得到:

表 A.4: 北京气温年循环的谐波分析

t	y_t	$\cos(\frac{2\pi t}{12})$	$\sin(\frac{2\pi t}{12})$	$y_t \cos(\frac{2\pi t}{12})$	$y_t \sin(\frac{2\pi t}{12})$
1	-4.02	0.866	0.500	-3.486	-2.012
2	-1.22	0.500	0.866	-0.612	-1.059
3	5.45	0.000	1.000	0.000	5.450
4	13.81	-0.500	0.866	-6.903	11.956
5	20.14	-0.866	0.500	-17.444	10.071
6	24.34	-1.000	0.000	-24.338	0.000
7	26.15	-0.866	-0.500	-22.647	-13.075
8	24.83	-0.500	-0.866	-12.413	-21.501
9	19.84	-0.000	-1.000	-0.000	-19.844
10	12.84	0.500	-0.866	6.418	-11.117
11	4.37	0.866	-0.500	3.784	-2.185
12	-2.04	1.000	-0.000	-2.038	0.000
$\sum_{i=1}^n$	144.48	0	0	-79.679	-43.3158

$$A = \frac{2}{n} \sum_{i=1}^n y_t \cos(\frac{2\pi t}{n}) \quad (\text{A.24})$$

$$B = \frac{2}{n} \sum_{i=1}^n y_t \sin(\frac{2\pi t}{n}) \quad (\text{A.25})$$

得到A和B以后就可以由下式计算位相:

$$\phi = \begin{cases} \arctan(\frac{B}{A}), & A > 0 \\ \arctan(\frac{B}{A}) \pm \pi, & A < 0 \\ \frac{\pi}{2}, & A = 0 \end{cases}$$

利用这些公式可以计算北京气温季节循环谐波参数。根据表A.4中数据，可以得到


$$\begin{aligned}\bar{y} &= \frac{144.48}{12} = 12.04^{\circ}\text{C} \\ A &= \frac{2}{n} \sum_{i=1}^n y_t \cos\left(\frac{2\pi t}{n}\right) = \frac{2}{12} \times (-79.678) = -13.2798 \\ B &= \frac{2}{n} \sum_{i=1}^n y_t \sin\left(\frac{2\pi t}{n}\right) = \frac{2}{12} \times (-43.3158) = -7.2193\end{aligned}$$

可见振幅 $C = \sqrt{B^2 + A^2} = \sqrt{228.4714} = 15.1153$ ；再根据位相公式计算得到位相 $\phi = \arctan\left(\frac{B}{A}\right) \pm \pi = 0.5436 \pm \pi$ ；即 $\phi = 3.6395$ 或者 $\phi = -2.6437$ ；这样得到北京气温年循环的准确谐波函数是

$$\begin{aligned}y_t &= 12.04 + 15.1153 \times \cos\left(\frac{2\pi t}{12} - 3.6395\right) \\ \text{或者} &= 12.04 + 15.1153 \times \cos\left(\frac{2\pi t}{12} + 2.6437\right)\end{aligned}$$

这与前面简单方法估计出的方程A.21非常接近。

■练习：根据上述谐波分析方法，利用Matlab估计2001年北京日气温变化的年循环信号以及日气温距平值。

 **练习：**根据上述谐波分析方法，利用Matlab估计2001年北京日气温变化的年循环信号以及日气温距平值。

```
> load tmpbj2001.mat;
> n=1:365;
> figure;plot(n,tmpbj2001); grid on;hold on;
> xlabel('Day');ylabel('T ( ^\circC)');
> axis([0 length(n) min(tmpbj2001)-2 max(tmpbj2001)+2]);
> ybar=mean(tmpbj2001);
> A=2/length(n)*tmpbj2001*cos(2*pi*n(:)/length(n));
> B=2/length(n)*tmpbj2001*sin(2*pi*n(:)/length(n));
> C=sqrt(A^2+B^2);
> if A >0 phase=atan(B/A); end
> if A <0 phase=pi+atan(B/A); end
> if A == 0 phase=pi/2; end
> yhat=ybar+C*cos(2*pi*n/length(n)-phase);
> plot(n,yhat,'r--', 'LineWidth',1.5);
> figure;plot(n,tmpbj2001-yhat);grid on;
> xlabel('Day');ylabel('\Delta T ( ^\circC)');
```

高阶谐波估计

上面的例子和练习，都是只取一个周期循环的情况，如果序列中还有其他更短的周期，就需要用高阶谐波进行拟合。当然数学上任何一个信号或者数据序列，都可以表示成一系列的不同周期谐波的和。这种情况下，可以理解为将原始数据中的不同周期信号进行了分解。当然分解出来的各个谐波只是从数学的方法得到的，是否有物理有意义需要根据实际情况来判断，也与具体的研究对象和研究目的有关。这在后面的例子中能看到这一点。

对于一个数据序列 y_t (可以是时间序列也可以不是时间序列)，从数学上都可以表示成

$$y_t = \bar{y} + \sum_{k=1}^{n/2} [C_k \cos(\frac{2\pi kt}{n} - \phi_k)] \quad (\text{A.26})$$

$$= \bar{y} + \sum_{k=1}^{n/2} [A_k \cos(\frac{2\pi kt}{n}) + B_k \sin(\frac{2\pi kt}{n})] \quad (\text{A.27})$$

其中 A_k , B_k 及位相由下面公式计算:

$$A_k = \frac{2}{n} \sum_{t=1}^n y_t \cos(\frac{2\pi kt}{n}) \quad (\text{A.28})$$

$$B_k = \frac{2}{n} \sum_{t=1}^n y_t \sin(\frac{2\pi kt}{n}) \quad (\text{A.29})$$

$$\phi_k = \begin{cases} \arctan(\frac{B_k}{A_k}), & A_k > 0 \\ \arctan(\frac{B_k}{A_k}) \pm \pi, & A_k < 0 \\ \frac{\pi}{2}, & A_k = 0 \end{cases}$$

k 表示谐波的阶数，上面对北京2001年气温年循环的谐波分析中就是 $k = 1$ ；振幅 $C_k = \sqrt{A_k^2 + B_k^2}$ 。当然谐波数还与实际资料的长度和取样频率有关，如果上面的分析中取资料长度为2000年1月1日到2001年12月31日，要拟合年循环则谐波数就应该是 $k = 2$ 。

以北京夏季平均气温距平时间序列为例，图A.6给出了 $k = 1 \dots 5$ 的谐波结果，前5个谐波加起来与观测气温时间序列有很好的一致性。谐波拟合的好坏可以用各谐波的方差表示， $S_k = \frac{1}{2} C_k^2$ 即为振幅平方的一半。用前5个谐波来拟合历年气温的变化，总的方差解释率是

$$\sum_{i=1}^5 S_k = \frac{1}{2} \sum_{i=1}^5 C_k^2 = 0.3813$$

这个值与观测气温时间序列的方差(0.7533°C^2)的比值是 $0.3813/0.7533 = 50\%$ 。如果用前20个谐波来拟合则可达94%。

从这个例子可以看到，如果时间序列中有比较明显的正弦函数方式的周期变化的话，在相应周期的谐波函数中应该体现出来。所以谐波函数可以用来检测周

表 A.5: 北京夏季气温前20个谐波的F检验值

$k =$	1	2	3	4	5	6	7	8	9	10
F_k	6.272	0.967	5.171	0.779	1.544	0.504	0.001	0.328	0.233	0.369
$k =$	11	12	13	14	15	16	17	18	19	20
F_k	0.122	0.196	0.252	0.312	0.113	4.124	0.471	0.922	1.435	1.561

期。数学上每个谐波都对应一个周期，但是不是都有意义。周期是否显著需要进行显著性检验。构造统计量

$$F_k = \frac{1}{2} \frac{C_k^2/2}{(s^2 - C_k^2/2)/(n - 2 - 1)} \quad (\text{A.30})$$

式中 s^2 为观测序列的方差。该统计量服从第一自由度 $\nu_1 = 2$ ，第二自由度为 $\nu_2 = n - 2 - 1$ 的F分布。给定显著性水平查F分布表，如果计算的 F_k 超过F阈值，则说明第 k 谐波是显著的，其对应的周期也是显著。当然还可以知道该显著周期变化对应的位相。图A.7中给出了北京夏季气温前20个谐波的方差贡献率，最突出的峰值在 $k = 1, 3, 16$ 。计算各谐波的F值列于表A.5中，显然各个谐波的F检验值差别很大。查F分布表¹可知95%置信度的阈值是 $F_{0.05} = 3.19$ 。因此，这20个谐波中只有 $k = 1, 3, 16$ 三个是显著的，其他的都可能是随机噪音。其中当 $k = 1$ 时相应的周期是 $n = 52$ 年，很容易算出 $k = 3, 16$ 时对应的周期分别是 $\frac{n}{k} = \frac{52}{3} = 17.3$ 和 $\frac{52}{16} = 3.3$ 年。实际计算中不用计算每个谐波的周期显著性，通常只要分析最突出的几个峰值就可以了，其他的谐波可以看成是随机现象。

上面的谐波分析的例子都是对时间序列进行的，实际上其应用不仅仅局限于时间序列。气候研究中常涉及到的超长波分析就是对空间数据进行谐波分析。给定一个纬度，如 40°N ，沿 40°N 纬度从 180°E 到 180°W 读出每个经度上位势高度值，得到一个位势高度值序列，位势高度的沿纬圈的变化可以看成是一种波动，对这个序列进行谐波分析，如果取前4阶谐波则在中纬度地区相应的波长超过5000km，因此称超长波。图A.8是对2004年1月平均500hPa高度场的超长波分析结果($k = 1, \dots, 4$)。对日平均高度场进行谐波分析，谐波阶数取得更高，还可以分析长波和短波。

¹Matlab中可以用函数Fa=finv(0.95,2,49)直接得到 $\nu_1 = 2, \nu_2 = 49$ ，95%置信度下的阈值，Fa=3.19

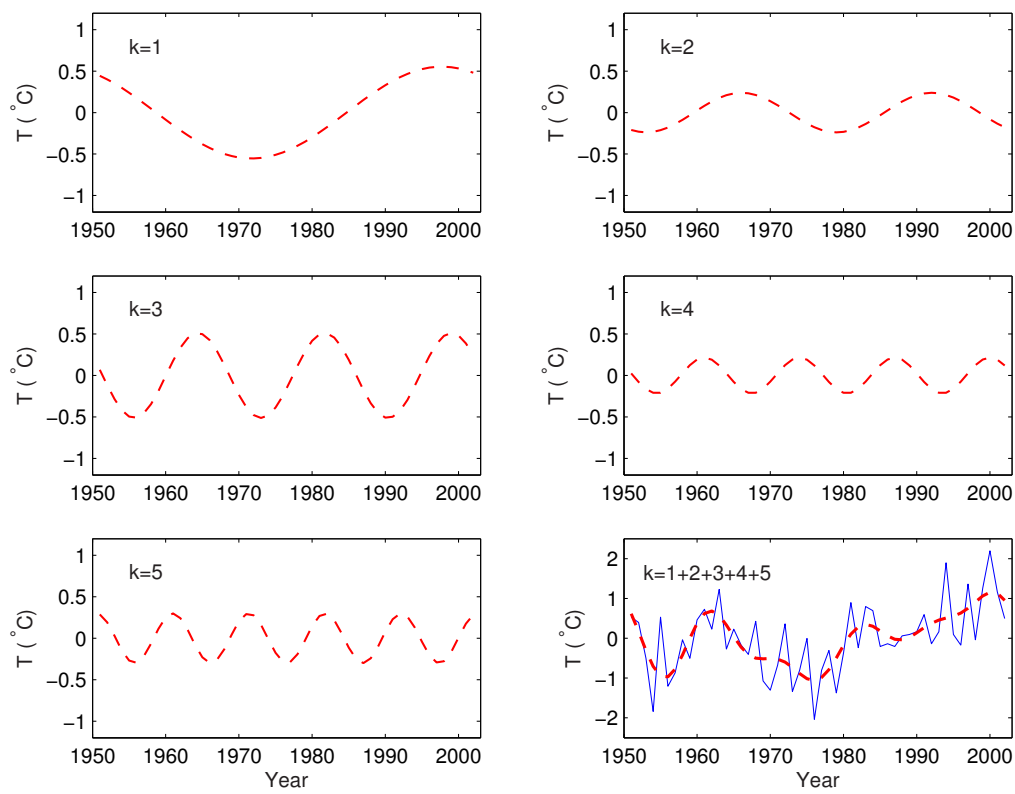


图 A.6: 北京夏季平均气温的谐波分析, 谐波阶数取 $k = 1, \dots, 5$, 虚线为谐波拟合结果, 实线为观测气温距平。

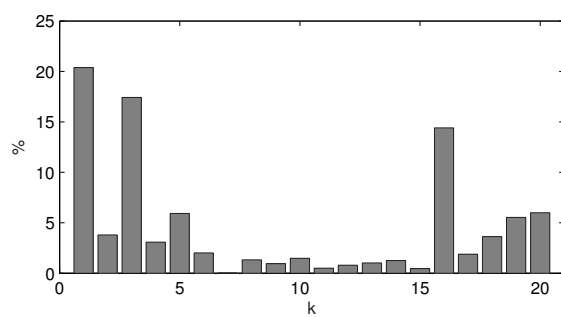


图 A.7: 北京夏季平均气温前20个谐波的方差贡献率。

根据谐波分析结果，还可以用显著周期进行外延做预测，如图A.6中可以选择 $k=1, 3, 16$ 这三个显著谐波对北京未来气温进行预测。此外，还可以对序列进行滤波。

A.4 功率谱分析

客观定量地检测时间序列中的显著周期，最常用的方法是功率谱。功率谱的计算常用的算法有两种。一是直接计算；二是落后自相关方法。

(1)直接计算方法。就是利用谐波分析方法，计算不同阶数的谐波振幅，即 C_k^2 ，振幅大表示能量强，因此也称功率谱(也称功率谱值或者是功率谱密度)。上面的例子中(图A.5)就是谐波直接计算的北京夏季气温的功率谱。对应的频率是

$$f = \frac{\omega_k}{2\pi} = \frac{k}{n}$$

单位是 a^{-1} 相应的周期是频率的倒数即

$$p = \frac{1}{f} = \frac{n}{k}$$

当 $k=1$ 时有最小的频率，也就是最长的周期， $p=n=52$ 年。随着谐波阶数 k 的增加对应的频率增加，周期变小。最多 k 能取多大呢？这是有限制的。显而易见表示一个完整的余弦波需要2年的资料，当 $k=\frac{n}{2}=26$ 时，刚好满足此条件。如果 $k=27$ 时一个完整的余弦波覆盖的资料不到2年，这不能满足描述一个完整的余弦波需要的最小资料数。实际计算中如果 $k>\frac{n}{2}$ 会出现假周期。 $k=\frac{n}{2}$ 时对应的频率就是最高频率，也称Nyquist频率，其对应的周期是最小周期，即 $\frac{n}{k}=\frac{n}{n/2}=2$ 年。因此直接计算方法能检测的频率在 $\frac{1}{n}-\frac{1}{2}$ 之间，即能检测的周期在 $2-n$ 之间。

通常可以用功率谱图来直观表示各周期的振幅或方差贡献，横坐标是频率(f)，或者周期(p)。因为频率和周期随 k 不是线性变化的，所以有时也用谐波阶数(k)做横坐标。纵坐标可以是 C_k^2 ，或者是标准化处理后的 C_k^2 ，图A.5中纵坐标就是将 C_k^2 处理成了方差贡献百分率。如果有时分析了很多谐波，而其中主要的能量集中在少数几个谐波上，则其对应的功率谱值很高，很多时候比其他谐波高几个数量级。因此为方便作图和比较常将纵坐标进行取对数处理。

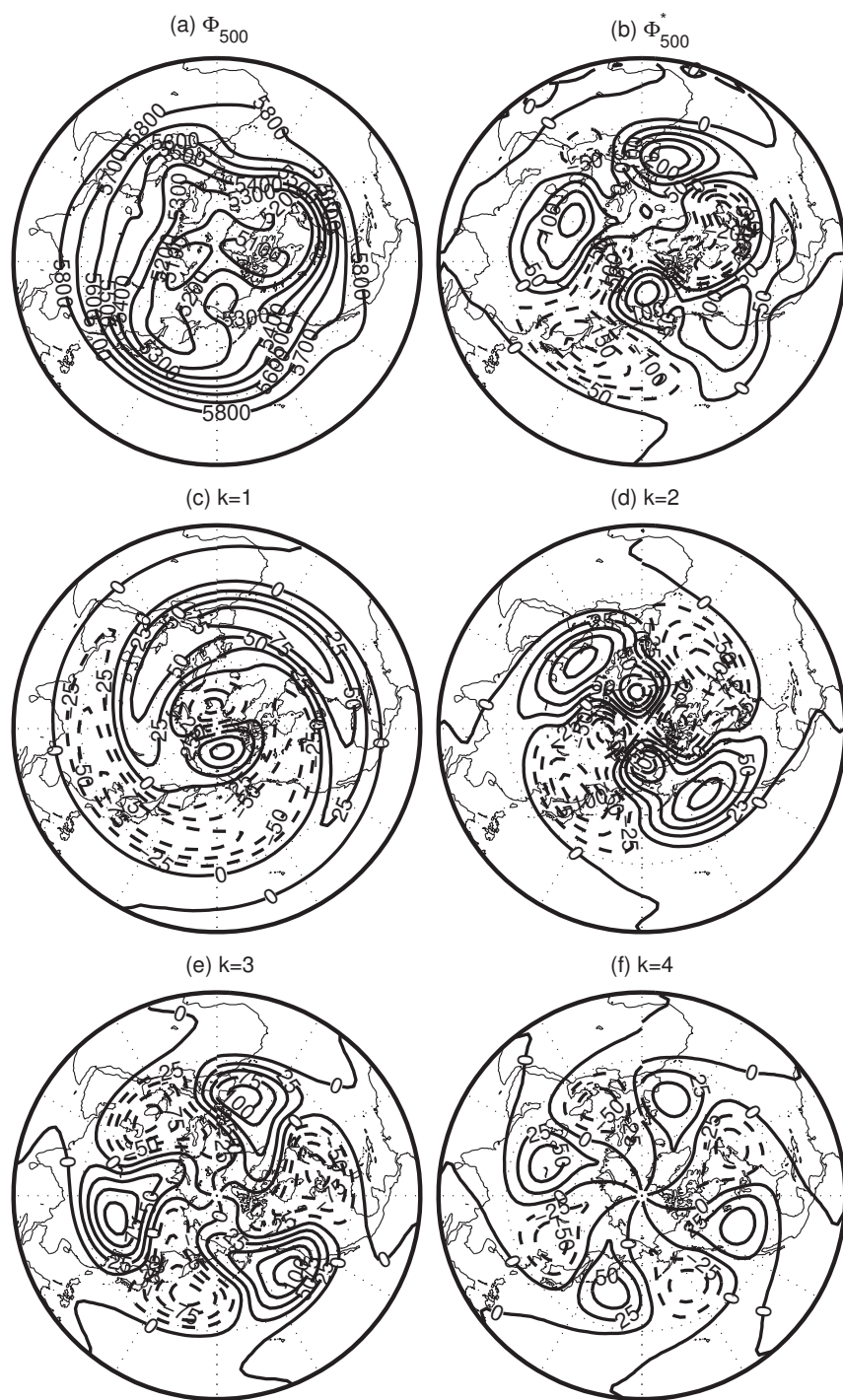


图 A.8: 2004年1月平均500hPa高度场的超长波分析。(a)为月平均高度场(Φ_{500}); (b)为对纬圈平均值的距平(即 Φ_{500}^*); (c)到(d)为对各纬圈高度值1—4阶谐波分析结果

(2)落后自相关方法。对一个时间序列先求其不同落后时间步长(τ)的自相关或者自协方差, 然后对自相关或者自协方差函数进行谐波分析, 以此来检测周期。如果一个时间序列有5年周期, 那么落后步长为5、10、15...时, 自相关或者自协方差函数就会周期性地出现极大值, 用谐波能很好的检测出来。落后步长长的话, 会提高对低频部分的检测分辨率, 但是计算落后相关使用的资料会变少而降低资料的自由度从而影响结果可靠性。如果落后步长取得太短又不利于低频周期的检测。落后步长取多少合适呢? 理论分析指出落后步长可以取 $\frac{n}{10} - \frac{n}{3}$, 实际计算中常常取 $\tau = \frac{n}{3}$ 左右。计算出来的功率谱值如果有峰值, 那么对应的周期可能明显, 是否显著还需要进行显著性检验。根据所分析的原时间序列的特性判断是红噪声谱还是白噪声谱, 然后分别用不同的方法进行检验。

落后自相关方法计算功率谱的步骤

- 决定最大落后步长($\tau = M$);
- 计算落后自相关系数

$$R(\tau) = \frac{1}{n - \tau} \sum_{t=1}^{n-\tau} \left(\frac{y_t - \bar{y}}{\sigma_y} \right) \left(\frac{y_{t+\tau} - \bar{y}}{\sigma_y} \right)$$

其中 $\tau = 0, \dots, M$;

- 计算功率谱粗谱密度

$$\hat{S}_k = \frac{B_k}{M} [R(0) + 2 \sum_{\tau=1}^{M-1} R(\tau) \cos(\frac{\pi k \tau}{M}) + R(M) \cos(\pi k)] \quad (\text{A.31})$$

其中系数 B_k 为

$$B_k = \begin{cases} 1, & k = 1, \dots, M - 1 \\ \frac{1}{2}, & k = 0, M \end{cases}$$

- 计算平滑功率谱密度值。上式中 \hat{S}_k 为功率谱粗谱密度值, 通常还有对其进行平滑处理来消除随机噪声的影响, 以便得到比较光滑和平稳的谱密度值。平滑的方法有很多种, 包括Barlett窗方法(矩形或者三角形窗), Hanning窗方法, Hamming窗方法等。这些方法使用不同的权重系数来对粗谱进行平

滑，结果也会有所不同，但是得到的谱密度值不会有本质的改变。得到平滑功率谱值 S_k 。

- 信度检验。上面得到的谱密度值如果有极大值，表示对应的频率和周期信号较强。是否显著需要进行检验。先要对原时间序列的噪声谱性质进行判断。给定置信度 $\alpha = 0.05$ ，有判据 $R_\alpha = \frac{-1+t_\alpha\sqrt{n-2}}{n-1}$ ，如果原时间序列 (y_t) 的一阶自相关系数 $R(1)$ 比 R_α 大，则为红噪声谱，反之为白噪声谱。 t_α 为 α 信度水平和 $\nu = n - 2$ 时的 t 分布值。以北京夏季平均气温时间序列为例，取 $\alpha = 0.05$ 和 $\nu = n - 2 = 50$ 时， $t_\alpha = 1.6759$ ，则 $R_\alpha = \frac{-1+1.6759\sqrt{50}}{n-1} = 0.2128$ ；而北京夏季平均气温的落后一年的自相关系数 $R(1) = 0.2935 > R_\alpha$ ，因此应该用红噪声谱检验。红噪声谱密度值的计算公式是

$$S_r = S_k \times \frac{1 - R(1)^2}{1 + R(1)^2 - 2R(1)\cos(\frac{\pi k}{M})}$$

为检验不同周期的显著性还需要计算出红色噪音谱的95%置信度的上限值

$$S_r^{0.05} = S_r \frac{\chi_{0.05}^2}{(2n - \frac{M}{2})/M} = S_k \times \frac{1 - R(1)^2}{1 + R(1)^2 - 2R(1)\cos(\frac{\pi k}{M})} \frac{\chi_{0.05}^2}{(2n - \frac{M}{2})/M}$$

如果原时间序列是白噪声谱，则95%置信度的白噪声上限为

$$S_r^{0.05} = S_k \frac{\chi_{0.05}^2}{(2n - \frac{M}{2})/M}$$

这里 $\chi_{0.05}^2$ 表示显著水平 $\alpha = 0.05$ 及自由度 $\nu = (2n - \frac{M}{2})/M$ 时的 χ^2 值(可以查表得到²⁾ 不管是红噪声谱检验还是白噪声谱检验，如果 S_k 超过 S_r 的95%信度上限阈值($S_r^{0.05}$)，则该周期是显著的。每个 k 对应的周期是 $P_k = \frac{2M}{k}$ ，当 k 等于最大落后步长时，对应的周期是2；当 $k = 1$ 时对应的周期是2倍最大落后步长，即所能表示的最长周期。如果功率谱值有多个显著极大值，表示有多个显著周期。相邻的多个周期中通常取中间的最突出的那个峰值。

■练习：利用落后自相关方法对北京夏季平均气温序列进行功率谱分析。并与图A.7结果进行比较。改变最大落后步长 $\tau_{max} = M$ ，看结果有什么变化。

²Matlab中， $\alpha = 0.05$ ，自由度 ν 的 χ^2 值可以由命令求得：Chia=chi2inv(0.95, ν)

■练习：利用落后自相关方法对北京夏季平均气温序列进行功率谱分析。并与图A.7结果进行比较。改变落后步长 τ ，看结果有什么变化。

```
function [num,Period, Frequency, Density, CL95]=spectrum(x,mLAG)
% function for power spectral analysis
% usage: [num,Period, Frequency, density, c195]=spectrum(x,mLAG)

xLEN=length(x); SER=x;N=xLEN;
mLAGWk=mLAG;mLEN=N;j=mLAG;j1=j+1;
%calculating auto-covariance coefficient
A=0.0;
C=0.;
for I=1:N A=A+SER(I);end % I
A=A/N;
for I=1:N SER(I)=SER(I)-A; C=C+SER(I).^2; end % I
C=C/N;
for L=1:J CC(L)=0.0;
for I=1:N-L CC(L)=CC(L)+SER(I)*SER(I+L); end %I
CC(L)=CC(L)./(N-L);
CC(L)=CC(L)/C;
end %L
C=1.0;
% estimating raw power spectra
SPE(1)=0.0;
for L=1:J-1 SPE(1)=SPE(1)+CC(L); end %L
SPE(1)=SPE(1)./(J+(C+CC(J))./(2*J));
for L=1:J-1 %
SPE(L+1)=0.;
for I=1:J-1 SPE(L+1)=SPE(L+1)+CC(I)*cos(pi*L*I/J); end % I
SPE(L+1)=2*SPE(L+1)./(J+C./J+(-1).^L*CC(J))./J;
end %
```

```
SPE(J1)=0.0;
for I=1:J-1 SPE(J1)=SPE(J1)+(-1).^I*CC(J); end %I
SPE(J1)=SPE(J1)/(J+(C+(-1).^J*CC(J))./(2*J));
%smoothing power spectra
PS(1)=.54*SPE(1)+.46*SPE(2);
for L=2:J PS(L)=.23*SPE(L-1)+.54*SPE(L)+.23*SPE(L+1); end %L
PS(J1)=.46*SPE(J)+.54*SPE(J1);
%statistical significance of PS
W=0.0;
for L=1:J-1 W=W+SPE(L+1); end %L
W=W/J+(SPE(1)+SPE(J1))./(2*J);
if (J > fix(N/2)) W=2.57*W; end
if(J == fix(N/2)) W=2.49*W; end
if(J < fix(N/2) & J > fix(N/3)) W=2.323*W; end
if (J == fix(N/3)) W=2.157*W; end
if (J < fix(N/3)) W=1.979*W; end
%the red noise examination
for L=1:J1
SK(L)=W*(1-CC(1).^2)/(1+CC(1).^2-2*CC(1)*cos(3.14159*(L-1)/J));
end % L
if (CC(1) > 0 & CC(1) >= CC(2) )
%the white noise examination
else
for L=1:J1 SK(L)=W; end %L
end % if
%calculating the length of cycle
T(1)=NaN;
for L=2:J1 T(L)=(2.0*J)/(L*1.0-1.0); end % L
num=1:J+1;num=num(:)-1; Period=T(:); Frequency=1./T(:);
Density=PS(:); CL95=SK(:);
```

注意事项

功率谱是对全局(即整个时间序列)频率或者周期的估计。如果某一时期某一周期比较明显,而其他时期该周期不明显,则功率谱分析可能反映不出来。对此可以分时间段来研究周期随时间的可能变化。例如图A.9是对不同时期Nino3区SST的功率谱分析结果。典型的ENSO变率集中在2—8年,但是不同时期分段分析结果可以发现,ENSO的典型年际变率也是随时间有明显的变化的。根据整个时间段(1856—2002)分析,明显的周期峰值包括5.5, 3.6, 及2.8年。比较其他的几个时段,可以看到在1856—1920及1951—2002年期间,3.6年左右的周期非常突出,但是在1921—1950期间 $0.02-0.03\text{月}^{-1}$ 左右的频率段功率谱是一个极小值。说明3.6年左右的周期明显减弱或者缺失。

此外,也可以用小波分析方法检查周期随时间的变化情况。

功率谱分析中对不同频率段的检测是不连续的,分辨率也是不同的。如图A.5中, $k=1$ 和2分别对应的周期是52年和26年,他们之间差了26年;而 $k=25$ 和26对应的周期分别是2.08和2年,之间只差0.08年。在直接功率谱计算方法中谐波阶数(k)通常使用的是整数,为提高分辨率也可以使用非整数。

落后自相关方法中,功率谱值对落后步长的变化也是比较敏感的。落后步长变化通常会影响到分析结果,特别是低频周期部分的结果(参考表A.6)。而对高频部分结果的影响比较小。通常可以根据相关知识对步长进行微调,以便使需要关注的周期附近具有较高的分辨率。例如,如果怀疑某一个变量可能受太阳活动影响,对其时间序列分析时就可以考虑调整落后步长,使得在11年左右的周期附近功率谱能有较高的分辨率,以利判断11年周期的谱值和显著性。

资料准备时,最好首先对资料进行标准化处理。然后进行去掉其趋势。如果有趋势的话,会影响谱密度的分布特征,使大部分的能量集中在低频部分,降低对高频变化信号的检测能力。

A.5 交叉谱分析

又称互谱(Cross spectral analysis)。用途是计算两个时间序列是否有显著的周期,两者之间各种周期的关联程度,以及位相差。

表 A.6: 北京夏季气温的功率谱分析结果，两种不同落后步长分析结果的比较。

$M = 15$					$M = 16$				
k	周期	频率	谱值	95%阈值	k	周期	频率	谱值	95%阈值
1	30.000	0.033	0.168	0.228	1	32.000	0.031	0.163	0.212
2	15.000	0.067	0.119	0.212	2	16.000	0.063	0.119	0.198
3	10.000	0.100	0.069	0.190	3	10.667	0.094	0.071	0.180
4	7.500	0.133	0.035	0.167	4	8.000	0.125	0.038	0.160
5	6.000	0.167	0.029	0.146	5	6.400	0.156	0.026	0.141
6	5.000	0.200	0.024	0.128	6	5.333	0.188	0.027	0.124
7	4.286	0.233	0.015	0.113	7	4.571	0.219	0.016	0.110
8	3.750	0.267	0.029	0.100	8	4.000	0.250	0.015	0.098
9	3.333	0.300	0.083	0.091	9	3.556	0.281	0.046	0.089
10	3.000	0.333	0.105	0.083	10	3.200	0.313	0.095	0.081
11	2.727	0.367	0.088	0.078	11	2.909	0.344	0.094	0.075
12	2.500	0.400	0.054	0.074	12	2.667	0.375	0.078	0.071
13	2.308	0.433	0.022	0.071	13	2.462	0.406	0.042	0.068
14	2.143	0.467	0.022	0.069	14	2.286	0.438	0.019	0.065
15	2.000	0.500	0.030	0.069	15	2.133	0.469	0.012	0.064
					16	2.000	0.500	0.015	0.064

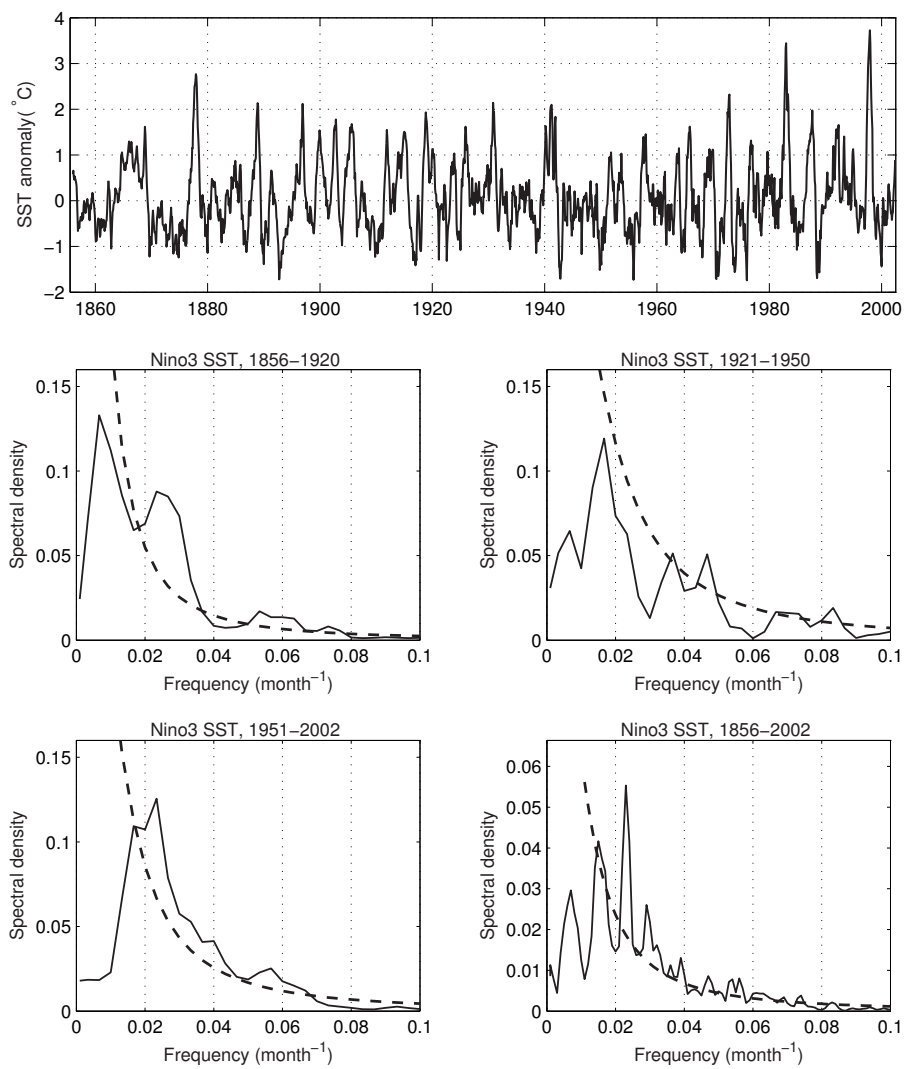


图 A.9: 不同时期Nino3区海温的功率谱分析。最上图是海温距平时间序列，下面四个图是几个时间段的功率谱值，其中的虚线是95%信度水平阈值线

对于两个时间序列 y_t 和 x_t 数据，进行交叉谱分析的步骤是

- 计算各自及相互之间的落后相关系数， $R_{xx}(\tau)$ ， $R_{yy}(\tau)$ ， $R_{xy}(\tau)$ ³及 $R_{yx}(\tau)$ ，其中 $\tau = 1, \dots, M$ 。最大落后步长 M 的取法与功率谱一致；
- 计算 x 和 y 各自的功率谱粗谱密度及平滑谱密度。计算公式为A.31。得到 $S_x(k)$ 和 $S_y(k)$
- 计算 x 和 y 的粗协谱。

$$\hat{P}_{xy}(k) = \frac{B_k}{M} \left\{ R_{xy}(0) + \sum_{\tau=1}^{M-1} [R_{xy}(\tau) + R_{yx}(\tau)] \cos\left(\frac{\pi k \tau}{M}\right) + R_{xy}(M) \cos(\pi k) \right\}$$

$k = 0, \dots, M$ ； 其中

$$B_k = \begin{cases} 1, & k = 1, \dots, M-1 \\ \frac{1}{2}, & k = 0, M \end{cases}$$

然后对粗协谱进行平滑处理以消除随机噪声。以下是一个常用的二项系数平滑公式

$$\begin{cases} P_{xy}(0) = \frac{1}{2} \hat{P}_{xy}(0) + \frac{1}{2} \hat{P}_{xy}(1); \\ P_{xy}(k) = \frac{1}{4} \hat{P}_{xy}(k-1) + \frac{1}{2} \hat{P}_{xy}(k) + \frac{1}{4} \hat{P}_{xy}(k+1); & k = 1, \dots, M-1 \\ P_{xy}(M) = \frac{1}{2} \hat{P}_{xy}(M-1) + \frac{1}{2} \hat{P}_{xy}(M) \end{cases}$$

其中 k 对应的周期是 $P_k = \frac{2M}{k}$

- 计算 x 和 y 的正交谱

$$\hat{Q}_{xy}(k) = \frac{1}{M} \sum_{\tau=1}^{M-1} [R_{xy}(\tau) - R_{yx}(\tau)] \sin\left(\frac{\pi k \tau}{M}\right)$$

$k = 1, \dots, M$ ； $\hat{Q}_{xy}(0) = \hat{Q}_{xy}(M) = 0$ 。同样需要对粗正交谱进行平滑以得到平稳的正交谱 $Q_{xy}(\tau)$

³ $R_{xy}(\tau) = \frac{1}{n-\tau} \sum_{t=1}^{n-\tau} \left(\frac{x_t - \bar{x}}{\sigma_x} \right) \left(\frac{y_{t+\tau} - \bar{y}}{\sigma_y} \right)$

$$\begin{cases} Q_{xy}(0) = \frac{1}{2}\hat{Q}_{xy}(0) + \frac{1}{2}\hat{Q}_{xy}(1) = \frac{1}{2}\hat{Q}_{xy}(1); \\ Q_{xy}(k) = \frac{1}{4}\hat{Q}_{xy}(k-1) + \frac{1}{2}\hat{Q}_{xy}(k) + \frac{1}{4}\hat{Q}_{xy}(k+1); \\ Q_{xy}(M) = \frac{1}{2}\hat{Q}_{xy}(M-1) + \frac{1}{2}\hat{Q}_{xy}(M) = \frac{1}{2}\hat{Q}_{xy}(M-1) \end{cases} \quad k = 1, \dots, M-1$$

- 计算凝聚谱。 x 和 y 的凝聚谱值 $CO_{xy}^2(k)$ 由下式计算

$$CO_{xy}^2(k) = \frac{P_{xy}^2(k) + Q_{xy}^2(k)}{S_x(k)S_y(k)}$$

$$k = 1, \dots, M$$

- 凝聚谱值($CO_{xy}^2(k)$)如果是极大值, 就表示 x 和 y 在该相同的频率($\frac{k}{2M}$, 即周期 $\frac{2M}{k}$)上都有较强的能量。是否显著需要检验。构造统计量

$$F = \frac{(\nu - 1)CO_{xy}^2(k)}{1 - CO_{xy}^2(k)}$$

$\nu = (2n - \frac{M}{2})/M$, 给定显著水平 α 查 F 分布表中第一自由度 $\nu_1 = 2$, 第二自由度 $\nu_2 = 2(\nu - 1)$ 的阈值 F_α ⁴, 如果凝聚谱值超过 F_α , 则对应的周期 $\frac{2M}{k}$ 是显著的。同时, x 和 y 中该周期对应的信号之间的位相差也可以求出, x 超前 y 的位相 $\phi(k)$ 是

$$\phi_{xy}(k) = \arctan\left(\frac{Q_{xy}(k)}{P_{xy}(k)}\right)$$

式中 $\phi_{xy}(k)$ 以弧度表示, 如果换算成时间, 则为 $\frac{\phi_{xy}(k)M}{k\pi k}$

程序: CROSPEC.FOR, 该程序根据“气象中的谱方法”(黄嘉佑, 李黄, 1984年, 气象出版社)一书有关内容编写。

实例: 全球副热带高压指数与赤道太平洋海温之间的交叉谱分析。借鉴中国气象局定义西太平洋副热带高压指数的做法, 定义全球副热带高压强度指数。50°S – 50°N之间所有500hPa位势高度超过5840gpm的格点, 与5840差值除10的累加和。即如果格点500hPa位势高度等于5850gpm, 则加1, 5860则加2, 其余以此类推。取5840gpm做为标准是考虑到如果标准太低, 月份之间差别不

⁴Matlab中, F_α 可以用命令求出。例如 $\alpha = 0.05$, $\nu_1 = 2$, $\nu_2 = 30$, 则 $F_\alpha = \text{finv}(0.95, 2, 30) = 3.32$

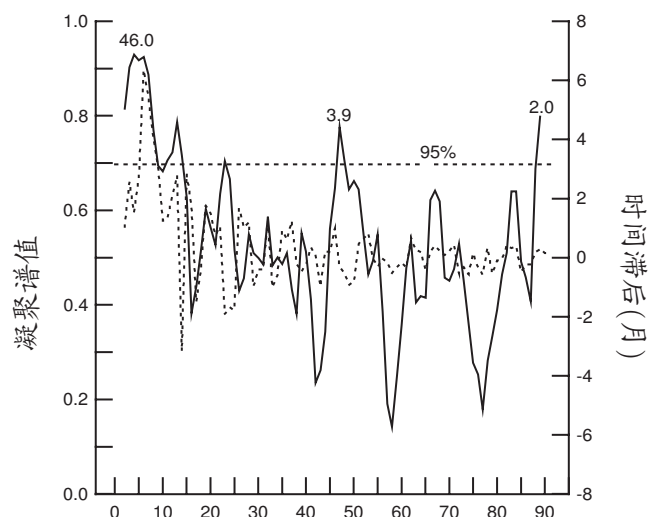


图 A.10: 全球副高指数和NinoC区SST交叉谱分析 (左纵坐标及实线为凝聚谱值, 右纵坐标及点线为位相差, 正值代表SST超前副高月份; 横坐标为落后步长(τ))

大, 反映不出月、季和年际变化; 如果标准取得太高则许多月份值为0, 失去了定义的意义。将各月标准化副高指数当做连续时间序列。副热带大气环流受赤道太平洋海温影响显著, 为了检查ENSO对全球副高影响的时间尺度及时间滞后关系, 对副高指数与NinoC区海表温度进行交叉谱分析。图A.10是交叉谱分析的凝聚谱值和位相差, 从达95%信度水平的46个月左右周期的位相看, 上述相关分析中的SST的数月超前关系是显著的, SST超前时间大约是6个月。

A.6 时间序列滤波分析

气候要素的时间序列中包含多种时间尺度变化, 而很多时候我们希望能保留某些频率的成分, 而去除掉其他的成分。如前面的谐波分析中, 我们就是保留了年循环的周期信号而去除了季节变化, 即相当于对噪声进行了消除。因此, 谐波分析是一种简单的滤波过程, 是一种低通滤波。当然我们可以通过谐波分析将我们需要的高频部分提取出来, 就是高通滤波。不过有时如果只有半年的资料, 就不能通过谐波来得到低频年变化信号。实际应用中这可以通过设计特殊的滤波器来实现。

表 A.7: 北京夏季气温简单滑动平均实例。简单3点滑动平均与1-2-1加权滑动平均。

年份	1951	1952	1953	1954	1955	1956	1957	1958	1959	...
T(°C)	25.6	25.5	24.6	23.3	25.6	23.9	24.2	25.1	24.6	...
3点滑动平均	$\times \frac{1}{3}$	$\times \frac{1}{3}$	$\times \frac{1}{3}$							
	—	25.2	24.5	24.5	24.3	24.6	24.4	24.6	25.1	...
1-2-1加权平均	$\times \frac{1}{4}$	$\times \frac{2}{4}$	$\times \frac{1}{4}$							
	—	25.3	24.5	24.2	24.6	24.4	24.4	24.7	25.0	...

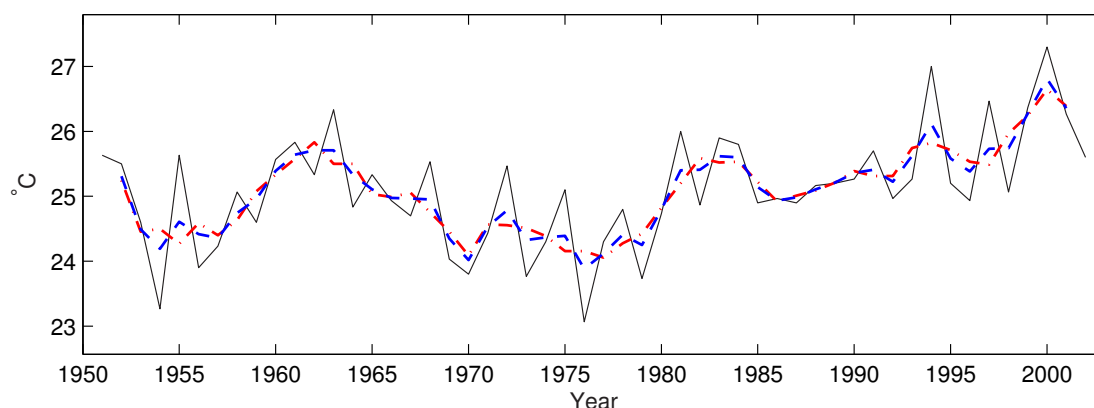


图 A.11: 北京夏季气温简单滑动平均。实线为原始序列值，点线为3点滑动平均，虚线为1-2-1加权滑动平均

例子：北京夏季气温序列。简单的3点滑动平均为等权重，相当于权重系数为 $\frac{1}{3}$ 。如中心点附近应该权重更高，简单3年1-2-1系数滑动平均(表A.7)。从图A.11中还可以看到，1-2-1加权平均比简单的3年滑动平均结果要光滑一些。通常加权系数越多得到的结果越平滑。

滤波的方法有很多种，按滤波的实现方式可以分为对称权重滤波和递归滤波。按用途可以分为低通滤波器，高通滤波器，带通滤波器，以及带阻滤波器。气候研究中最常用的滤波器都是对称权重的数字滤波器，因此滤波工作的一个重要内容是设计合适的数字滤波器，得到相应的滤波权重系数(w)。

频率响应

气候研究中最常用的滤波器都是对称权重的数字滤波器(w)，用权重系数对原时间序列 x 进行滤波可以得到所需要的时间序列即

$$y_t = \sum_{k=-L}^L w_k x_{t+k}$$

权重系数 $w_k = w_{-k}$ ；其长度是 $2L + 1$ 。

对于给定的频率 f ，时间序列在滤波前后可能的变化包括该频率的振幅强度和位相。如果权重系数是对称的，则位相是不变的。但是其振幅强度可能会发生改变，我们总是希望我们感兴趣的频率滤波后其振幅能不变，或者大部分能保留下来，而不感兴趣的频率的振幅等于或者接近零。如果用谐波来表示 y 和 x 的话，对应频率 f 的谐波振幅分别为 $C_y(f)$ 和 $C_x(f)$ ，那么其比值

$$H(f) = \frac{C_y(f)}{C_x(f)}$$

就是该频率的响应， H 称为频率响应函数。类似上述例子的对称权重系数滤波器，在信号处理中也称有限脉冲响应(FIR, finite impulse response)滤波器。其影响在时间上是有限的，如1-2-1滤波器，就只影响3个相邻的数据。这类滤波器的频率响应函数可以表达为

$$H(f) = w_0 + 2 \sum_{k=1}^L w_k \cos(2k\pi f \Delta t) \quad (\text{A.32})$$

f 为频率， w_k 是第 k 个权重系数， w_0 是中心权重系数， Δt 是资料时间步长，通常资料是等时间步长的，可以当成1；如是月分辨率的，其步长基本单位是1月，如果是年分辨率的资料， Δt 就是1年。这样A.32式可以简化为

$$H(f) = w_0 + 2 \sum_{k=1}^L w_k \cos(2k\pi f)$$

从1-2-1滑动平均的频率响应函数看(图A.12)，在 $f = 0$ 处即线性趋势上，滤波后完全不变，对高频部分 $f = 0.5$ 即周期为2的分量则完全消除了。该滤波器可以看成是一个低通滤波器。但是频率的衰减非常缓慢，对低频部分也有相当大的削弱，所以做为低通滤波器其滤波效果并不很理想。

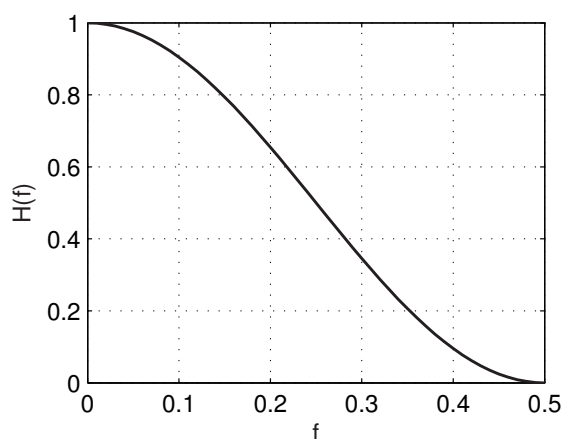


图 A.12: 1-2-1加权滑动平均的频率响应函数

几种简单滑动平均

这里介绍几种常用的简单滤波器，包括二项式系数滤波器，高斯滤波器。

二项式系数滤波器滤波器的权重系数由二项式系数求得⁵:

$$B_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad k = 0, 1, \dots, n$$

如果要计算3个二项系数，则 $n = 3 - 1 = 2$, $B_0 = \frac{2!}{0! \times 2!} = 1$; $B_1 = \frac{2!}{1! \times 1!} = 2$; $B_2 = \frac{2!}{2! \times 0!} = 1$;

显然这就是上面用到的1-2-1滤波器(相应的权重为1/4, 2/4, 1/4)。对于 $n = 4$ 时，系数为1, 4, 6, 4, 1; 换算成权重为1/16, 4/16, 6/16, 4/16和1/16。表A.8列出了 $n = 2, 4, 6, 8, 10$ 及12时的二项式系数。随着权重系数的增加，对高频部分的削弱更明显，图A.13中给出了9点二项式权重系数滤波器的频率响应函数，很明显，对高频部分($f > 0.3$)全部滤掉了，对低频部分则保留甚多，对 $f < 0.1$ 的部分保留率在60%以上。这与前面的1-2-1滤波器的响应函数相比，是更理想的低通滤波器。因此，对于年资料序列(如北京夏季平均气温序列)，如果想保留年10年以上时间尺度的长期变化，就可以用9点二项式滤波器进行滤波。

低通与高通是可以转换的，如果从北京夏季气温时间序列中减去9点二项式低通滤波的结果，得到的主要是年际尺度的变化，即相当于高通滤波处理结果。要

⁵计算二项式系数的Matlab命令为: `nchoosek(n,k)`，如 `nchoosek(2,0)=1`; `nchoosek(4,2)=6`

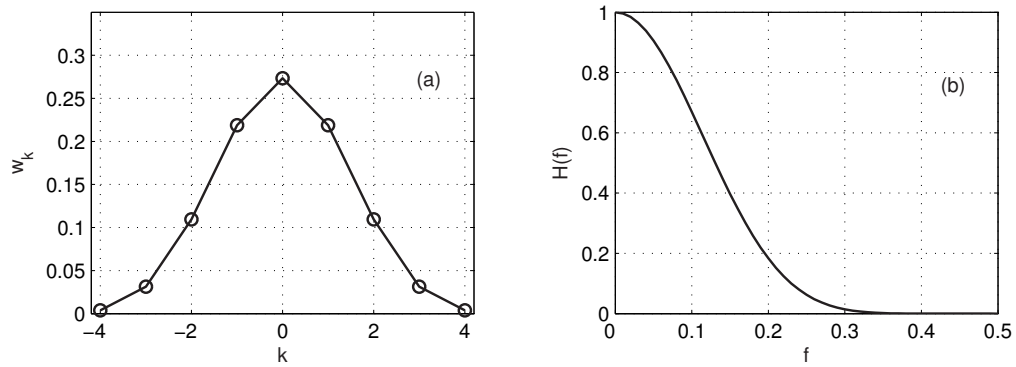


图 A.13: 9点二项式权重系数(a)及其频率响应函数(b)

注意的是滤波会损失数据，前面和后面各损失 $(2L + 1 - 1)/2 = L$ 个数据。共损失 $2L$ 个数据。为了得到与原序列相同的长度的滤波结果，常在原数据前后各添加 L 个数据，然后再滤波。添加的方式有很多，可以是添加平均值，或者是相邻的 L 个数据反向对称地添加。添加数据后得到的前 L 和后 L 个滤波结果是有一定误差的。

高斯滤波器(Gaussian filter)的权重系数遵从正态分布。表A.9给出了几种长度的高斯滤波器权重系数。图A.9中是相应的频率响应函数。可以根据其频率响应特征，挑选符合要求的权重。例如，想要检查北京夏季气温时间序列中的10年以上尺度的波动，则截断频率为 0.1年^{-1} 。判断标准是10年以上尺度的变化滤波后损失低于50%，即频率响应函数至少在50%以上。根据图A.9，权重系数的长度为9时，满足上述要求。因此，9点高斯常用来进行低通滤波，以消除年际波动变化。

设计滤波器

给定截断频率 f_s ，希望 $f > f_s$ 的部分完全去掉，而对 $f \leq f_s$ 的部分完全保留。这个理想的滤波器的频率响应函数为

$$H(f) = \begin{cases} 1, & 0 \leq f \leq f_s \\ 0, & f_s < f \leq \frac{1}{2} \end{cases}$$

表 A.8: 二项式权重系数

n	$k =$	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
2	B_k						1	2	1					
	w_k						0.250	0.500	0.250					
4	B_k					1	4	6	4	1				
	w_k					0.063	0.250	0.375	0.250	0.063				
6	B_k				1	6	15	20	15	6	1			
	w_k				0.016	0.094	0.234	0.313	0.234	0.094	0.016			
8	B_k			1	8	28	56	70	56	28	8	1		
	w_k			0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004		
10	B_k		1	10	45	120	210	252	210	120	45	10	1	
	w_k		0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001	
12	B_k	1	12	66	220	495	792	924	792	495	220	66	12	1
	w_k	0.000	0.003	0.016	0.054	0.121	0.193	0.226	0.193	0.121	0.054	0.016	0.003	0.000

表 A.9: 不同长度的高斯滤波权重系数

n	系数																
	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
5							.05	.24	.40	.24	.05						
7						.03	.10	.22	.29	.22	.10	.03					
9					.02	.06	.12	.19	.22	.19	.12	.06	.02				
11				.01	.04	.07	.12	.17	.18	.17	.12	.07	.04	.01			
13			.01	.02	.05	.08	.12	.14	.16	.14	.12	.08	.05	.02	.01		
15		.01	.02	.03	.06	.08	.11	.13	.13	.13	.11	.08	.06	.03	.02	.01	
17	.01	.01	.03	.04	.06	.08	.10	.11	.12	.11	.10	.08	.06	.04	.03	.01	.01

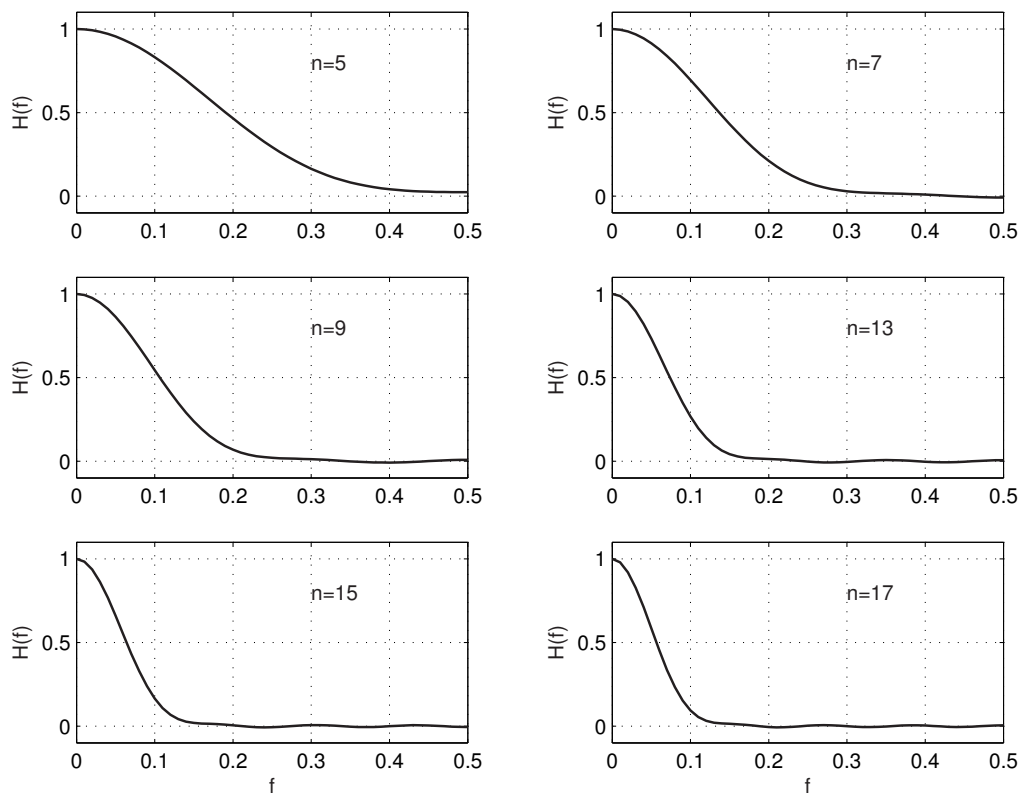


图 A.14: 不同权重系数个数高斯滤波的频率响应函数,

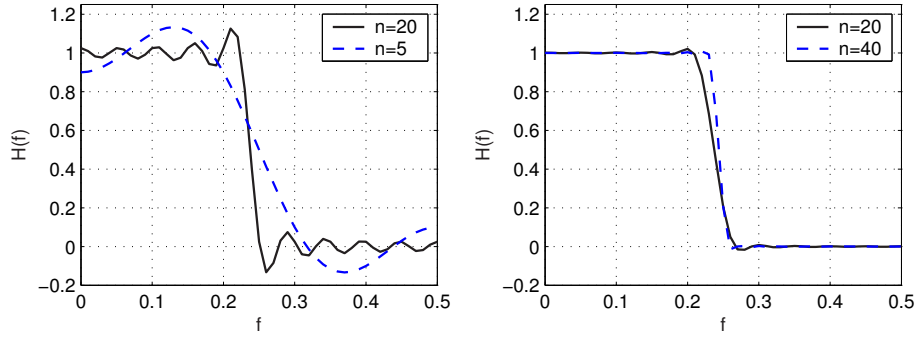


图 A.15: 不同参数 n 的权重系数的频率响应函数, 右边为Lanczos方法平滑后的频率响应函数

$$w_k = \int_0^{+\infty} H(f) \cos(2\pi f k) df$$

将 f 划分为等间隔的 n 等份, 则 $f_t = \frac{t}{2n}$ 计算上式积分;

$$w_k = \frac{1}{2n} \left[H(0) + 2 \sum_{t=1}^n H\left(\frac{t}{2n}\right) \cos(2\pi k \frac{t}{2n}) \right]$$

如果我们取截断频率 $f_s = 0.25$, 取 $n = 5$, 那么上式计算出的权重系数 $w_{-4} = 0$, $w_{-3} = -0.12$, $w_{-2} = 0$, $w_{-1} = 0.32$, $w_0 = 0.5$, $w_1 = w_{-1} = 0.32$, $w_2 = w_{-2} = 0$, $w_3 = w_{-3} = -0.12$, $w_4 = w_{-4} = 0$,

不过, 从其频率响应分布来看, 滤波效果与设想的还是有一定的差距。有两个方面的原因, 一个是 $n = 5$ 太小, 当 $n = 20$ 时, 效果就会改善很多(图A.15)。另外, 有很多起伏, 需要平滑处理。常用的一种平滑方法是Lanczos平滑法, 即相当于对权重系数加上一个平滑系数 $\frac{\sin(\pi k/n)}{\pi k/n}$, 平滑后的权重系数

$$\bar{w}_k = \frac{\sin(\frac{\pi k}{n})}{\frac{\pi k}{n}} w_k, \quad 1 \leq k \leq n$$

类似的方法可以设计高通, 带通, 带阻等滤波器。不过由上面的例子可以看出, 在应用中一个最大的缺点是要想获得理想的滤波效果, 往往要求较多的滤波系数, 而实际观测资料一般情况下往往较短, 滤波后序列前后还要各损失 $(n-1)$ 个

资料。参数 n 对结果影响很大，需要最优估计。应用起来很不方便。实际应用中，利用递归滤波器更为快捷有效。

递归滤波器有如下形式

$$y_n = \sum_{k=0}^K a_k x_{n-k} + \sum_{j=1}^J b_j y_{n-j}$$

这说明较低的阶数往往就能达到较为理想的结果。计算速度也很快。Matlab中提供多种递归滤波器的设计程序，其中Butterworth滤波器较为简单，效果也较好，因此常常使用。

调用命令是 $[b,a]=butter(ord,wa)$ ， ord 是滤波器的阶数，阶数高时滤波效果好。 wa 为归一化角频率，从0到1，1对应 π ，也就是Nyquist频率。下面给出几个例子具体说明任何使用。

■ 练习1: 对北京夏季气温实际序列，如果希望滤波保留10年以上尺度的变化，而滤掉10年以下的周期。

(1) 取10年为周期，则截断时间频率为0.1。因为资料的Nyquist频率是0.5；那么归一化频率 $wa = 0.1/0.5 = 0.2$ ；

(2) 选择阶数 $ord = 9$ ，利用 $[b,a]=butter(ord,wa)$ 得到系数 b 和 a 。

(3) 检查 b 和 a 确定的滤波器的频率响应函数。 $[hw,w]=freqz(b,a)$ ， hw 为角频率 w 的响应函数。可以作出频率响应图来检查。因为 w 是角频率，需要转换成时间频率($f = \frac{w}{2\pi}$)，即 $plot(w/2/\pi,hw)$

(4) 利用 b 和 a 进行滤波。 $yt=filtfilt(b,a,x)$ yt 就是所需要的低通滤波结果。原始序列减 yt 还可以得到高频变化部分。

■ 练习2: 对北京夏季气温实际序列，如果希望滤波保留5-10年尺度的变化，而滤掉其他时间尺度部分，同样利用Butterworth滤波器。

(1) 两个截断频率为 $wa = 0.1/0.5 = 0.2$ ；及 $wa = 0.2/0.5 = 0.4$

(2) 选择阶数 $ord = 9$ ，利用 $[b,a]=butter(9,[0.2 \quad 0.4])$ 得到系数 b 和 a 。

(3) 检查 b 和 a 确定的滤波器的频率响应函数。 $[hw,w]=freqz(b,a)$ ， $plot(w/2/\pi,hw)$

(4) 利用 b 和 a 进行滤波。 $yt=filtfilt(b,a,x)$ 就可以得到需要的带通滤波结果。

A.7 EOF分析

经验正交函数分析方法(empirical orthogonal function, 缩写为EOF), 也称特征向量分析(eigenvector analysis), 或者主成分分析(principal component analysis, 缩写PCA), 是一种分析矩阵数据中的结构特征, 提取主要数据特征量的一种方法。Lorenz在1950年代首次将其引入气象和气候研究, 现在在地学及其他学科中得到了非常广泛的应用。地学数据分析中通常特征向量对应的是空间样本, 所以也称空间特征向量或者空间模态; 主成分对应的是时间变化, 也称时间系数。因此地学中也将EOF分析称为时空分解。

原理与算法

- 选定要分析的数据, 进行数据预处理, 通常处理成距平的形式。得到一个数据矩阵 $X_{m \times n}$
- 计算 X 与其转置矩阵 X^T 的交叉积, 得到方阵

$$C_{m \times m} = \frac{1}{n} X \times X^T$$

如果 X 是已经处理成了距平的话, 则 C 称为协方差阵; 如果 X 已经标准化(即 C 中每行数据的平均值为0, 标准差为1), 则 C 称为相关系数阵

- 计算方阵 C 的特征根($\lambda_1, \dots, \lambda_m$)和特征向量 $V_{m \times m}$, 二者满足

$$C_{m \times m} \times V_{m \times m} = V_{m \times m} \times \Lambda_{m \times m}$$

其中 Λ 是 $m \times m$ 维对角阵, 即

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_m \end{bmatrix}$$

一般将特征根 λ 按从大到小顺序排列, 即 $\lambda_1 > \lambda_2 > \dots > \lambda_m$ 。因为数据 X 是真实的观测值, 所以 λ 应该大于或者等于0。每个非0的特征根对应

一列特征向量值，也称EOF。如 λ_1 对应的特征向量值称第一个EOF模态，也就是 V 的第一列即 $EOF_1 = V(:, 1)$ ；第 λ_k 对应的特征向量是 V 的第 k 列，即 $EOF_k = V(:, k)$ 。

- 计算主成分。将EOF投影到原始资料矩阵 X 上，就得到所有空间特征向量对应的时间系数(即主成分)，即

$$PC_{m \times n} = V_{m \times m}^T \times X_{m \times n}$$

其中 PC 中每行数据就是对应每个特征向量的时间系数。第一行 $PC(1, :)$ 就是第一个EOF的时间系数，其他类推。

上面是对数据矩阵 X 进行计算得到的EOF和主成分(PC)，因此利用EOF和PC也可以完全恢复原来的数据矩阵 X ，即

$$X = EOF \times PC$$

有时可以用前面最突出的几个EOF模态就可以拟合出矩阵 X 的主要特征。此外，EOF和PC都具有正交性的特点，可以证明 $\frac{1}{n}PC \times PC^T = \Lambda$ ；即不同的PC之间相关为0。 $E \times E^T = I$ 。I为对角单位矩阵，即对角线上值为1，其他元素都为0。这表明各个模态之间相关为0，是独立的。

由上面的计算过程可以看出，EOF分析的核心是计算矩阵 C 的特征根和特征向量。计算矩阵特征根和特征向量的方法很多，下面具体给出Matlab中进行EOF分析的两种不同的方法。具体步骤可参考下面两个框图中的实例。

方法1: 调用 $[EOF, E] = \text{eig}(C)$ ，其中EOF为计算得到的空间特征向量，E为特征根。然后计算主成分 $PC = EOF^T \times X$ 。需要指出的时，当数据量很大时，例如分析高分辨率的资料(如1km分辨率的NDVI资料)，空间范围很大维数 m 很容易超过数十万个点，则矩阵 C 的维数是个巨大量，需要占用大量内存，也会导致计算速度异常缓慢。而且很可能超出计算机的计算极限而死机。

方法2: 直接对矩阵 X 进行奇异值分解

$$X = U \sum V^T$$

其中 \sum 为奇异值对交阵(\sum 对角线上的元素为奇异值)，奇异值与特征根成倍数关系。

- 如果矩阵 $C = \frac{1}{n}XX^T$, C 的特征根为 λ , 则有 $\sum = \sqrt{n\lambda}$;
- 如果矩阵 $C = XX^T$, C 的特征根为 λ , 则有 $\sum = \sqrt{\lambda}$;

由于该方法是直接对矩阵 X 进行分解, 所以对内存的要求远小于方法1。计算速度很快。

两种方法对比练习。

显著性检验

可以证明

$$\sum_{i=1}^m \overline{X_i^2} = \sum_{k=1}^m \lambda_k = \sum_{k=1}^m \overline{PC_k^2}$$

这说明矩阵 X 的方差大小可以简单的用特征根的大小来表示。 λ 越高说明其对应的模态越重要, 对总方差的贡献越大。第 k 个模态对总的方差解释率为

$$\frac{\lambda_k}{\sum_{i=1}^m \lambda_i} \times 100\%$$

即使是随机数或者虚假数据, 放在一起进行EOF分析, 也可以将其分解成一系列的空间特征向量和主成分。因此, 实际资料分析中得到的空间模态是否是随机的, 需要进行统计检验。North等(1982)的研究指出, 在95%置信度水平下的特征根的误差

$$\Delta\lambda = \lambda \sqrt{\frac{2}{N^*}}$$

λ 是特征根, N^* 是数据的有效自由度, 这在前面相关系数分析中已经有介绍(见4页相关内容)。将 λ 按顺序依次检查, 标上误差范围。如果前后两个 λ 之间误差范围有重叠, 那么他们之间没有显著差别。

图A.16是对1949 – 2002年北半球1月平均海平面气压, 做距平处理处理及面积加权后进行EOF分析的结果。从特征根误差范围看, 第一和第二模态存在显著差别, 第二和第三模态之间也存在显著差别。但是第三特征根和第四及以后的特征根之间没有显著的差别。如果要分析主要的模态的话, 最好只选择前三个进行分析。

■练习：利用 $[E,V]=\text{eig}(C)$ 计算矩阵 X 的特征向量和主成分%

```
X=[2 6 1 5 2;  
    9 4 0 5 4];  
X(1,:)=X(1,:)-mean(X(1,:)); X(2,:)=X(2,:)-mean(X(2,:));
```

得到X的距平值: X=

```
-1.20    2.80   -2.20    1.80   -1.20  
 4.60   -0.40   -4.40    0.60   -0.40
```

%%% co-variance matrix

```
C=X*X'/5;
```

协方差阵C=

```
 3.76    0.92  
 0.92    8.24
```

```
[EOF,E]=eig(C); % V: eigenvectors; E: eigenvalues
```

```
PC=EOF'*X;
```

%% reverse the order

```
E=fliplr(flipud(E))
```

```
lambda=diag(E); % retain eigenvalues only
```

```
EOF=fliplr(EOF)
```

```
PC=flipud(PC)
```

得到EOF=

```
 0.19   -0.98  
 0.98    0.19
```

得到特征根E=

```
 8.42    0  
 0    3.58
```

得到主成分PC=

```
 4.28    0.15   -4.74    0.94   -0.62  
 2.07   -2.82    1.31   -1.65    1.10
```

%%check

```
EOF*EOF' % = I
```

检查EOF的正交性得到:

```
 1.00    0  
 0    1.00
```

```
PC*PC'/5 % = lambda
```

检查PC的正交性得到:

```
 8.42    0.00  
 0.00    3.58
```

```
EOF*PC % =X
```

可以完全恢复X的距平值:

```
-1.20    2.80   -2.20    1.80   -1.20  
 4.60   -0.40   -4.40    0.60   -0.40
```


■练习：利用 $[U, S, V] = \text{svd}(X)$ 计算矩阵 X 的特征向量和主成分

```
X=[2 6 1 5 2;  
    9 4 0 5 4];
```

```
X(1,:)=X(1,:)-mean(X(1,:));
```

```
X(2,:)=X(2,:)-mean(X(2,:));
```

X 的距平是:

-1.20	2.80	-2.20	1.80	-1.20
4.60	-0.40	-4.40	0.60	-0.40

```
[U,S,V]=svd(X);
```

得到 $U=$

0.19	0.98
0.98	-0.19

$S=$

6.49	0	0	0	0
0	4.23	0	0	0

$V=$

0.66	-0.49	0.56	0.09	-0.06
0.02	0.67	0.63	-0.32	0.22
-0.73	-0.31	0.53	0.25	-0.16
0.14	0.39	0.03	0.91	0.06
-0.10	-0.26	-0.02	0.06	0.96

```
EOF=U;
```

```
PC=S*V';
```

得到 $PC=$

4.28	0.15	-4.74	0.94	-0.62
-2.07	2.82	-1.31	1.65	-1.10

```
E=S.^2/5; %=lambda
```

E 的数值与上面得到的特征根完全一样即 $E=$:

8.42	0	0	0	0
0	3.58	0	0	0

```
EOF*PC % =X
```

可以完全恢复 X 的距平值:

-1.20	2.80	-2.20	1.80	-1.20
4.60	-0.40	-4.40	0.60	-0.40

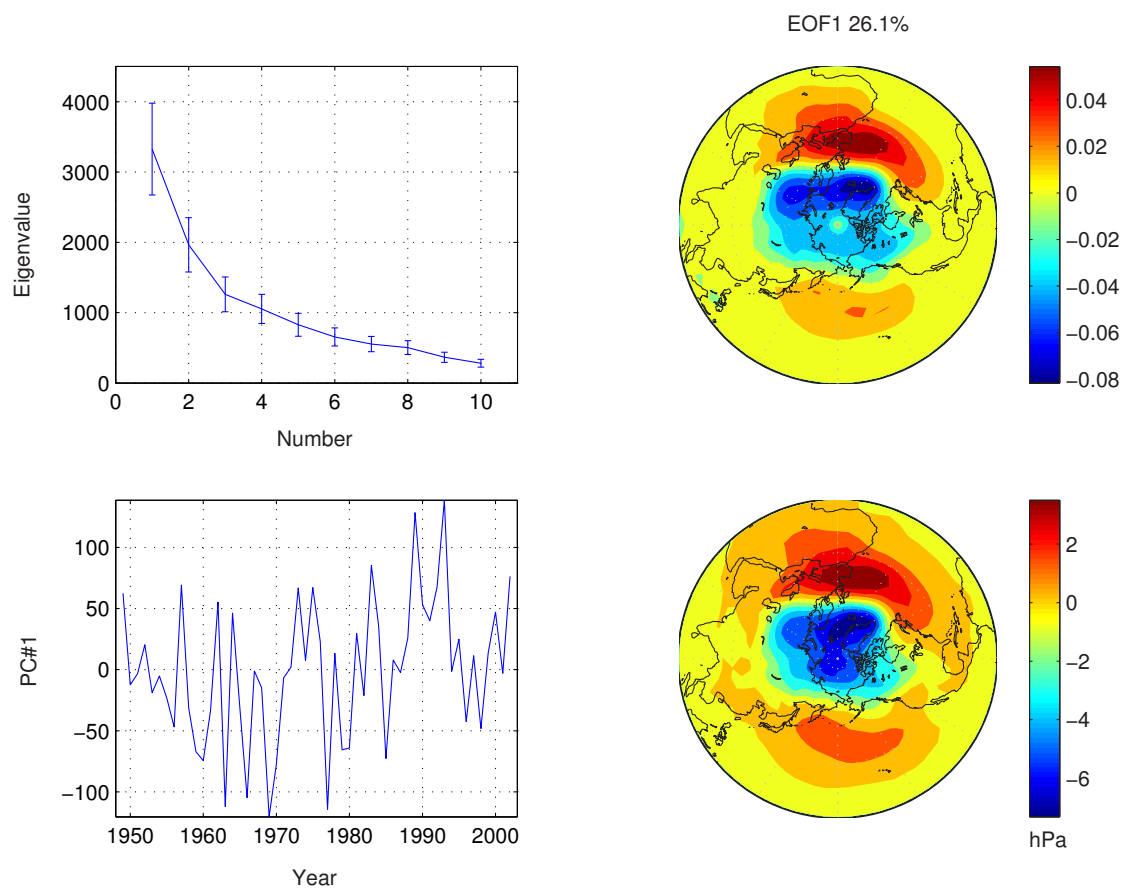


图 A.16: 北半球1月海平面气压EOF分析的第一特征向量. (a)为特征根及95%信度误差, (b)第一特征向量, (c)第一主成分, (d)第一主成分偏强 $+\sigma$ 时海平面气压的变化量(hPa). 1949 – 2002, NCEP/NCAR再分析资料

结果展示

通常情况下，主成分是有单位的，即反映的是矩阵 X 的单位，而空间特征向量是无量纲的。不过实际应用中常常对EOF分析得到的主成分和特征向量进行标准化处理得到新的 PC^* 和 EOF^*

$$PC^*(k) = \frac{PC(k)}{\sqrt{\lambda_k}}$$
$$EOF^*(k) = EOF(k) \sqrt{\lambda_k}$$

或者是简单地将PC标准化，使得其平均值为0，标准差为1。再将它与原始资料矩阵 X 进行回归分析，这样就得到PC变化一个单位时，变量 X 对应的响应的空间特征及其强度。这样得到的回归系数的空间分布与空间特征向量的分布特征空间分布特征是相似的，但是回归系数可以看出相应的变化的数量大小。如图A.16(d)。

空间模态应该与主成分配合进行分析。二者符号是相对应的。

分析中保留的模态的数目，没有严格规定，还取决于分析目的。一般取满足North准则；或者有明确物理意义。

数据性质与预处理

(1) 误差

(2) 资料的处理。原始场，距平场，与标准化场

例子：我国160站夏季降水量的EOF分析(图A.17)

(3) 空间样本点。大范围的空间数据，特别需要注意资料空间代表性。非均匀

场与均匀分布场；空间抽样；面积加权。

北半球1月SLP例子

时空转换

有时空间样本 m 远大于时间序列长度 n ，计算 $m \times m$ 矩阵的特征根很困难，可以考虑对其进行时空转换。矩阵 $A = \frac{1}{n}XX'$ 和 $B = XX'$ 的特征根不同，但是特征向

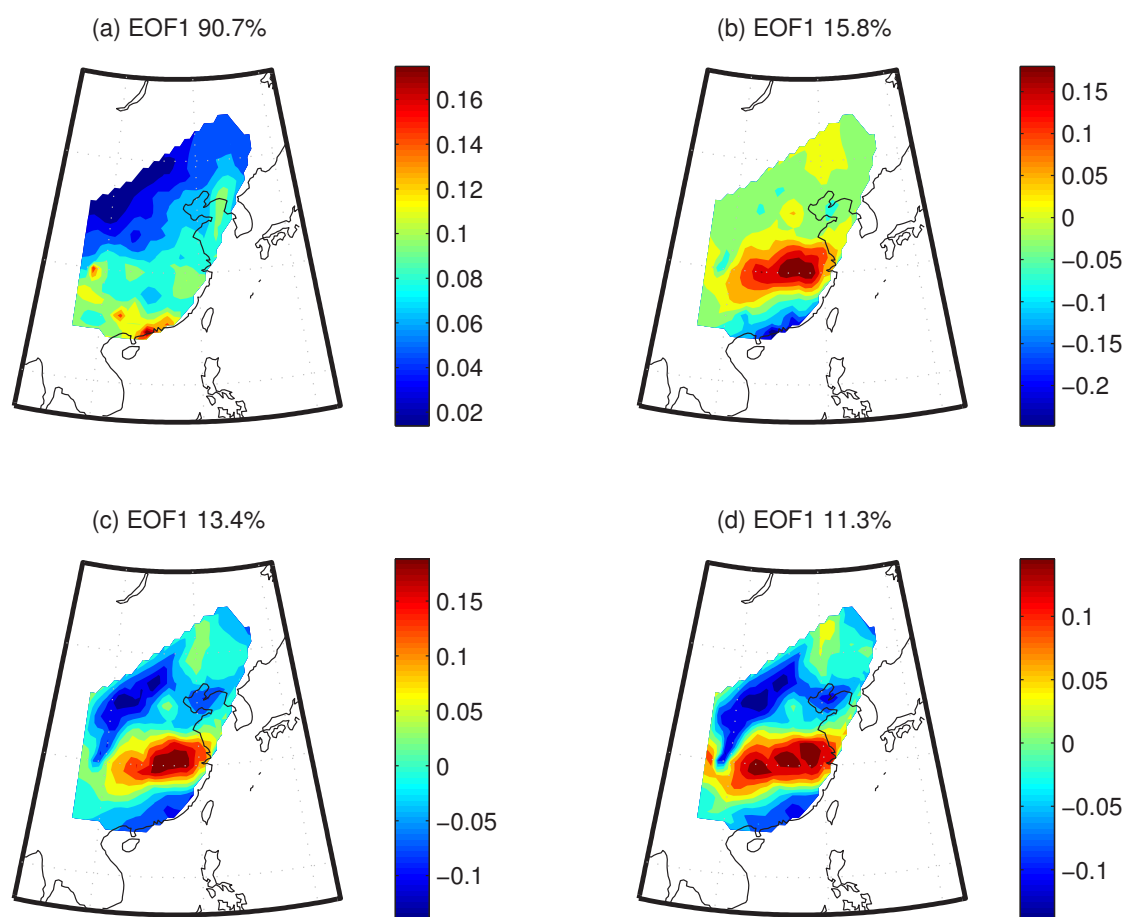


图 A.17: 我国东部地区夏季降水量EOF分析第一特征向量。(a)原始值, (b)距平值, (c)距平百分率, (d)标准化值. 1951 – 2002资料.

量是一样的。而可以证明 $C = X X'$ 和 $C^* = X' X$ 有相同的特征根, 但特征向量不同。因此, 通过时空转换可以求 $X' X$ 矩阵的特征根, 进而计算 $X X'$ 矩阵的特征向量。即有

$$C^* \times V^* = V^* \times \Lambda$$

V^* 是 C^* 的特征向量, Λ 是特征根对角矩阵。根据 V^* 是可以求出 C 的特征向量的, 首先计算 $V_a = X \times V^*$; 对 V_a 进行处理得到 C 的前 n 个特征向量 V_k

$$V_k = \frac{1}{\sqrt{\lambda_k}} V_a(:, k)$$

得到特征向量 V 后, 就可以计算相应的主成分

$$PC = V^T \times X$$

前面计算得到的 EOF 维数是 $m \times m$, 而通过时空转换得到的 EOF 维数只有 $m \times n$ 。即只能得到前 n 个特征向量。不过实际应用中对结果影响并不大, 因为通常我们只关心前几个最重要的模态。

下面是一个简单例子, 有一个矩阵 X , 维数是 5×2 , 先直接计算矩阵 $X X'$ 的5个特征向量, 然后再利用时空转换方法计算其前2个特征向量。

```
X=[ -1.20  4.60
    2.80 -0.40
   -2.20 -4.40
    1.80  0.60
   -1.20 -0.40]
[V1,E1]=eig(X*X'); %%
V1=fliplr(V1);%%
E1=fliplr(flipud(E1));%%
得到特征向量V1=
   -0.66    0.49   -0.45   -0.15   -0.32
   -0.02   -0.67   -0.14   -0.15   -0.72
    0.73    0.31   -0.56   -0.15   -0.17
```

```

-0.14    -0.39    -0.42    -0.58    0.57
 0.10     0.26     0.53    -0.77    -0.19

```

得到特征根E1=

```

42.11    0    0    0    0
 0    17.89    0    0    0
 0         0    0    0    0
 0         0    0    0    0
 0         0    0    0    0

```

如果进行时空转换的话，计算结果是：

```
[V2,E2]=eig(X'*X);%%
```

```
V2=fliplr(V2);%%
```

```
E2=fliplr(flipud(E2));%%
```

得到特征向量V2=

```

 0.19    -0.98
 0.98     0.19

```

得到特征根E2=

```

42.11    0
 0    17.89

```

可见E1和E2是一样的。再计算XX'矩阵的第一特征向量：

```
Va=X*V2; %%
```

```
V_k1=Va(:,1)/sqrt(E2(1,1));
```

得到：

```

 0.66
 0.02
-0.73
 0.14
-0.10

```

计算XX'矩阵的第二特征向量是：

```
V_k2=Va(:,2)/sqrt(E2(2,2));
```

得到：

```
0.49
```

-0.67
0.31
-0.39
0.26

可见，用时空转换方法得到的2个特征向量，与前面直接计算矩阵 XX' 和矩阵 $\frac{1}{n}XX'$ 的得到的前两个特征向量完全一致。当数据量很大时，如对全球1月份2.5°分辨率的再分析1000hPa高度场(Φ)进行EOF分析时，空间点的数量是 $m = 10512$ ，时间长度 $n = 54$ ，则矩阵 $C = \Phi\Phi^T$ 的维数是 10512×10512 ，而如果用时空变换方法，则矩阵 $C^* = \Phi^*\Phi$ 的维数是 54×54 ，很快就可以计算出前54个特征向量。高分辨率的遥感数据如NDVI，其空间维数远比气象数据大，但其长度通常只有20年左右，因此进行EOF分析时常需要借助时空变换手段。

A.7.1 REOF分析

算法：程序varimax.m。模态数目的选择。

A.8 SVD分析

EOF分析中一次只分析了一个变量 X ，地学中常常涉及多个要素场之间的关系。分析多个要素场关系的方法也有很多，包括混合EOF(combined empirical orthogonal function, 缩写CEOF)，奇异值分解(singular value decomposition, 缩写SVD)分析，典型相关(canonical correlation analysis, 缩写CCA)等。他们本质上是相同的。这里主要介绍SVD分析。需要指出的是这里SVD分析只是利用SVD方法检测两个要素场相关模态和分析的过程，SVD本身只是对矩阵运算求其奇异值及广义逆等，因此不要将二者混淆。

算法

- 两个矩阵 X 和 Y ，维数分别是 $m \times n$ 和 $p \times n$ 先计算他们的协方差阵 $C = \frac{1}{n}XY^T$ ， C 的维数是 $m \times p$
- 进行SVD分解得到

$$C = U \sum V^T$$

， U 是对应 X 的空间模态， V 是对应 Y 的空间模态， \sum 对角线为奇异值 γ

Matlab中命令为 $[U,S,V]=\text{svd}(C)$

- 主成分。 X 的主成分是 $A = U^T X$ ， Y 的主成分是 $B = V^T Y$
- 解释率。 γ 与 X 和 Y 的协方差平方成正比，因此解释率是

$$\frac{\gamma^2}{\sum \gamma^2} \times 100\%$$

结果解释：实例

分析1982－2000年春季北半球NDVI和气温之间的耦合关系。首先将每一个格点上的NDVI和温度都处理成对1982－2000年的距平，再相乘得到协方差阵，对协方差阵进行SVD分析，可以得到奇异值，每一个奇异值对应的NDVI和温度的模态，以及每一种模态的时间系数。结果见图A.18。

春季植被NDVI对温度的响应信号非常强。二者之间的协方差高度集中在最前面的几对模态中。第一到第七对模态，解释率分别为42.6，19.5，10.3，7.7，5.0，4.2和2.3%，这7对模态的总解释率高达91.6%，说明整体上来看春季NDVI与温度的关系是很密切的。二者之间最主要也是最重要的耦合关系已经包含在这前面几个模态之中了，这也表明我们只分析这几个模态就已经足够了。

其中最重要的第一对模态中心在西西伯利亚。第二对模态的主要特征是整个北美大陆表现为相同符号的变化，中心在美国的东北部地区。这前两对模态的空间尺度都很大，属于大陆尺度。第三及以后的各对模态尺度相对较小，都是区域性的。而且这些模态表现出NDVI与温度异常的高度一致性，正的温度中心对应NDVI的正中心，负的温度中心对应NDVI的负中心。通常最强的NDVI变化中心，也是温度异常的极值中心。

上述耦合模态都受大气环流变化的显著影响，最重要的第一模态与EU遥相关型有密切的联系，NDVI和温度的时间系数与EU的相关分别达到了0.72和0.78。第二模态与WP型关系最密切。第三模态与PNA关系最密切。第六模态NDVI和温度的时间系数与NAO的相关分别为0.52和0.58，都超过95%信度水平。第七模态NDVI和温度的时间系数与WA的相关分别为-0.52和-0.56，也都是显著的。有些模态同时受多个因子影响，如第五模态可能反映了包括PNA，SO，EA及NP等多种因素的影响。(详细内容可参考：龚道溢，史培军，何学兆. 北半球春季植被NDVI对温度变化响应的空间差异. 地理学报，2002, 57(5),505-514)

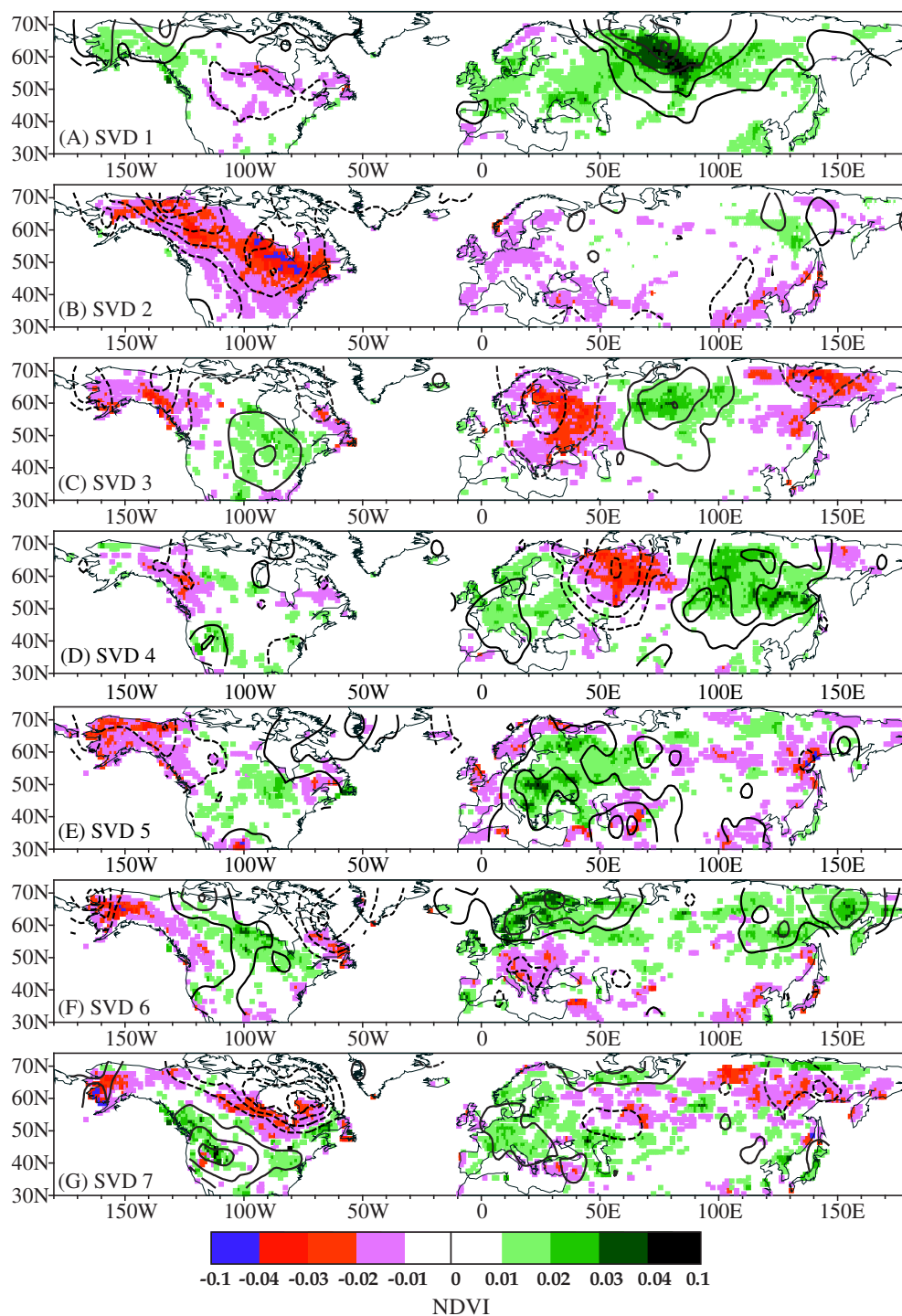


图 A.18: 北半球春季NDVI和气温SVD分析前七对相关模态. 1982 – 2000年资料.