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($\lambda_{1,\ldots,m}$)和特征向量$V_{m\times m}$，二者满足

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

  其中$\wedge$是$m \times m$维对角阵，即

$$\wedge = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_m \end{bmatrix}$$

  一般将特征根λ按从大到小顺序排列，即$\lambda_1 > \lambda_2 > \ldots > \lambda_m$。因为数据$X$是真实的观测值，所以$\lambda$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

$$PC_{m\times n} = V^T_{m\times m} \times X_{m\times n}$$

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$\frac{1}{n}PC \times PC^T = \wedge$；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]=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]=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]=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
```

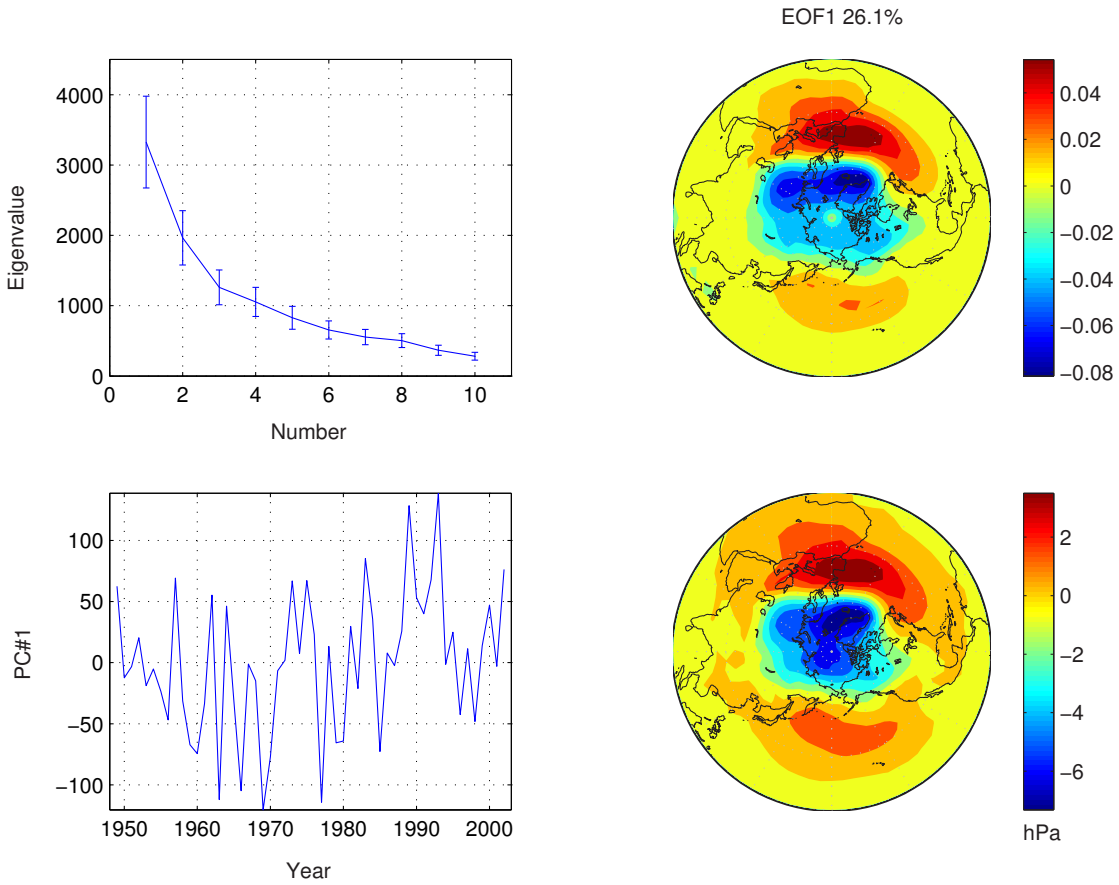

图 A.16: 北半球1月海平面气压EOF分析的第一特征向量. (a)为特征根及95%信度误差，(b)第一特征向量，(c)第一主成分，(d)第一主成分偏强$+\sigma$时海平面气压的变化量(hPa). 1949 – 2002，NCEP/NCAR再分析资料

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

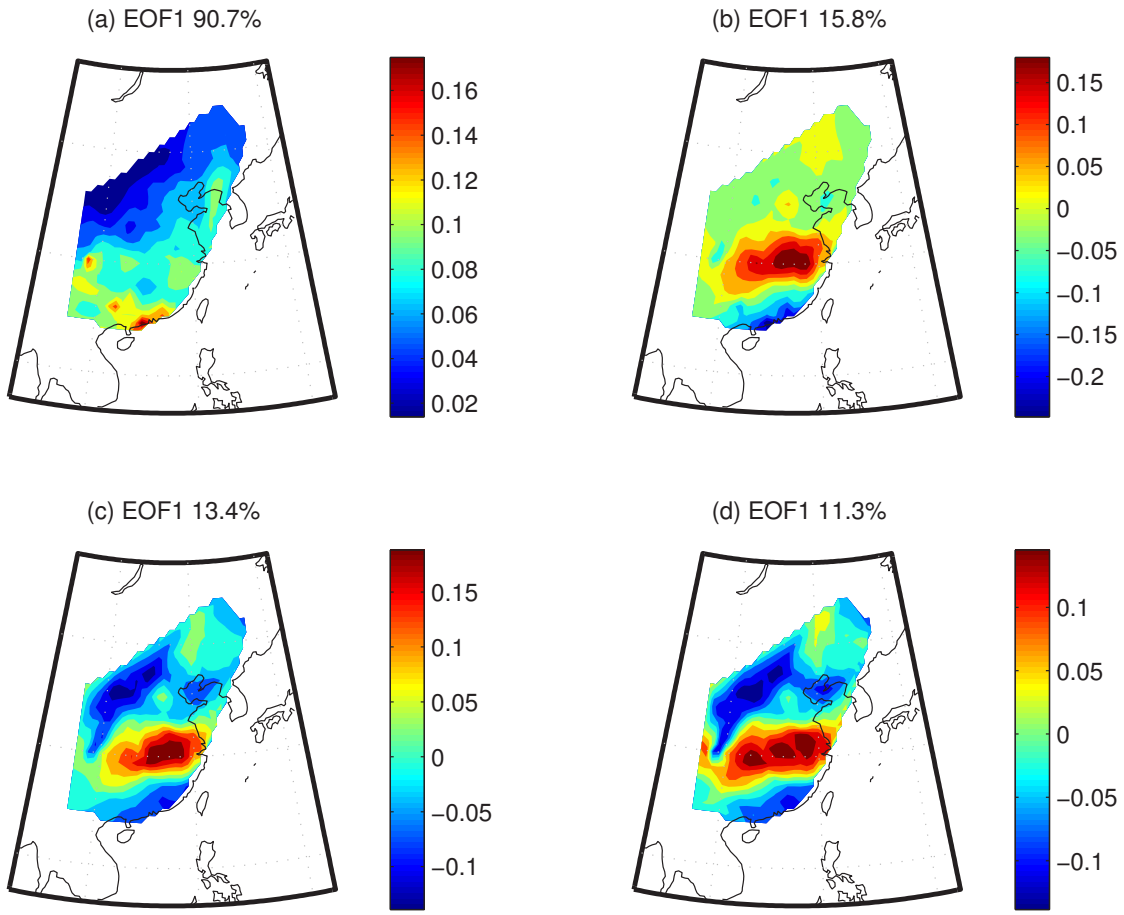

图 A.17: 我国东部地区夏季降水量EOF分析第一特征向量。(a)原始值，(b)距平值，(c)距平百分率，(d)标准化值. $1951-2002$资料.

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XX'矩阵的特征向量。即有

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

V^*是C^*的特征向量，$\wedge$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XX'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的。这里主要介绍SVD分析。需要指出的是这里SVD分析只的是利用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]=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)

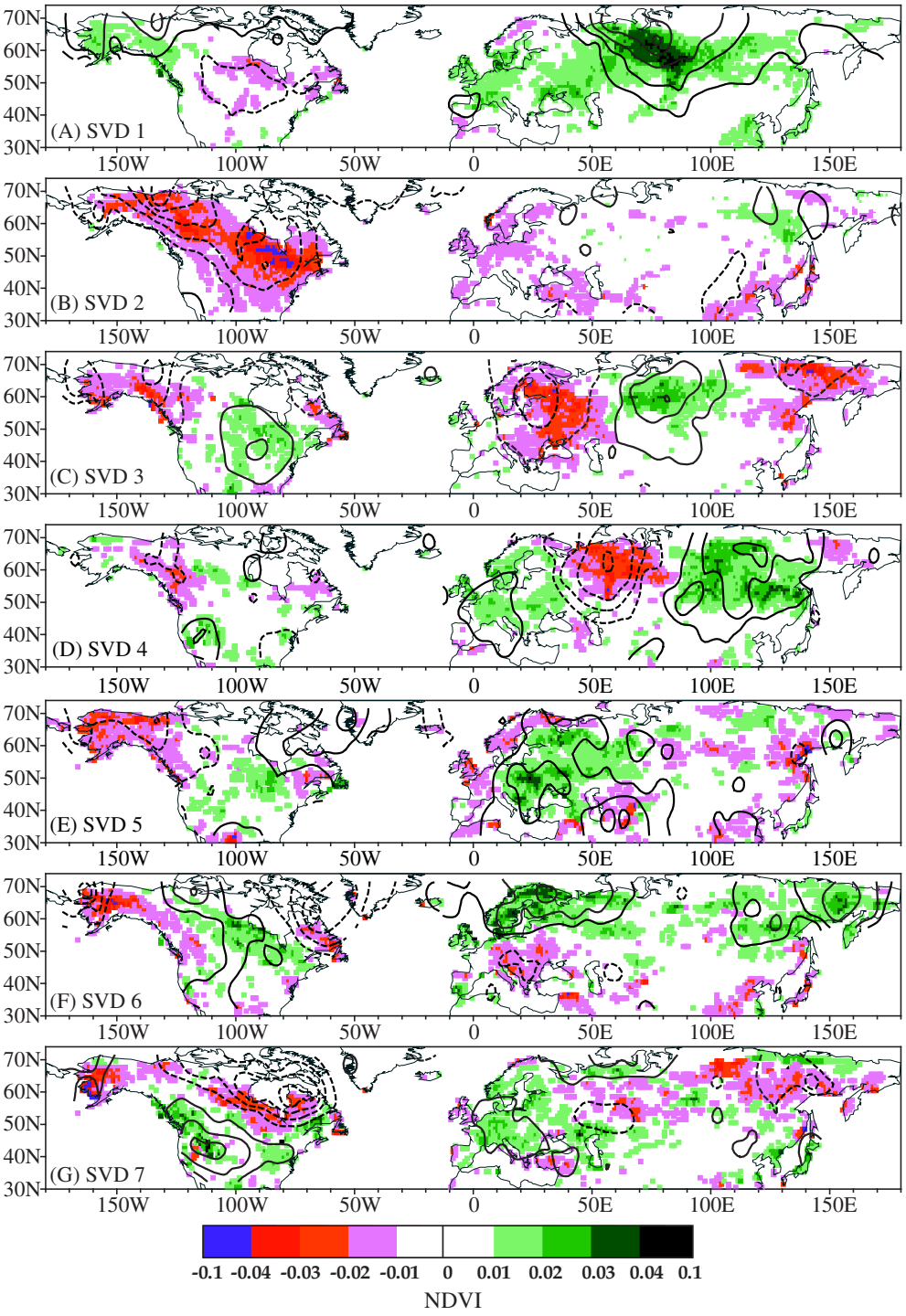

图 A.18: 北半球春季NDVI和气温SVD分析前七对相关模态. 1982 – 2000年资料.