EOF在分解什么?

EOF和SVD在分解什么？原数据，还是协方差?

简单的回答是，EOF在分解时空数据时，是在通过*特征分解*分解该数据的协方差矩阵，或者通过*奇异值分解*分解数据本身。

标题中EOF方法全称为：Empirical Orthogonal Functions (EOFs)，机器学习领域常称作Principal Component Analysis（PCA)； 后面会提到Singular Value Decomposition（SVD），中文分别翻译为经验正交函数，主成分分析和奇异值分解。EOF分解常用于提取空间模态及其对应的时间序列。关于该方法相关的讲解材料很多，本文推荐 A manual for EOF and SVD analyses of Climate Data。这篇文章不会进行系统的讲解（因为没有人家讲得好）。需要明确的是，EOF分解多用于研究单个场（或变量）的分析，如SLP, SST, SAT etc. 尽管该方法可以很容易得扩展应用到包含多个变量的数据（比如同时包含SST和SLP的数据），但后一问题多通过SVD的方法解决。后面的讨论会看到，奇异值分解可以帮助我们在数学上进一步理解EOF分解，为了区分，当涉及两个场的数据的分解时，我们用缩写SVD，当讨论EOF分解和奇异值分解的关系时，我们用全称。

为了阅读容易，规定以下数学符号：

> X:大写字母，表示一个变量，代表单个时间序列；
> x: 小写字母，表示一个数字；
> $\mathbf{X}$:加粗大写字母，表示一个矩阵；
> $\mathbf{x}$:加粗小写字母，表示一个向量（只有一行（行向量）或一列（列向量）的矩阵）。

简单的数学概念

我们先回顾几个简单的统计学概念，一个**随机变量**是对一类事件出现结果的数学表示，比如多次抛硬币这一事件的结果可以表示为一个随机变量，它的取值为$[1,0,1,1,0...]$。线性代数中，我们用大写的字母X表示该随机变量，小写字母 $x_1, x_2, \dots x_n$表示其所有取值，即$X=[x_1, x_2, \cdots x_n]$，此处X是一个行向量（即只有一行的矩阵），代表单个序列。**期望**反映该随机变量平均值大小，数学表示为$\overline{x}=\frac{1}{n}\sum_i^n x_i$，线性代数表示为$E[X]$。**方差**描述该序列的离散程度，我们最早接触到的定义是该序列中各个值与其平均值的距离的平均值，数学表示为$\sigma=\frac{1}{n}\sum_i^n(x_i-\bar{x})^2$，线性代数表示为$Var(X)=E[X-\mu]$， 其中，$\mu=E[X]$，简单一推导，可得，$Var(X)=E[X^2]-(E[X])^2$，抛去物理意义不谈，方差表示为该变量所有取值平方之后的期望与期望之后的平方的差。

以上术语只涉及单个变量，现在引入第二个变量$Y=[y_1, y_2 \dots y_n]$。两个变量之间的**协方差**表征两个序列协变的程度，如果X的较大值与Y的较大值相对应（在序列中基本相同的位置出现），X的较小值与Y的较小值相对应，则可以说X与Y有正的协方差，数学定义为$Cov(x,y)=\frac{1}{n}\sum_i^n(x_i-\bar{x})\cdot(y_i-\bar{y})$，（对比方差的公式其实很好理解），线性代数表示为$Cov(X,Y)=E[(X-E[X])(Y-E[Y])]=E[XY]-E[X]E[Y]$[1]。与方差一样，协方差不具有尺度独立性（not scale invariant），比如，同样是抛硬币（事件），将正反两面表示为0和1的方差，与将正反两面表示为-1和1的方差是不同的。为了克服协方差的这一问题，利用两个变量的方差对协方差进行标准化，结果表示为**相关系数**，即$Corr(X,Y)=Cov(X,Y)/\sqrt{Var(X)Var(Y)}$。

更进一步地，我们讨论多个变量。涉及的概念是**协方差矩阵**，上文仅提到的两个变量，变量X或Y都是单个序列，代表行向量，x_i 和y_i都代表数字。更常见的情形是用一个加粗的大写字母如$\mathbf{X}$（矩阵）一次性表示所有的m个变量，即$\mathbf{X}=[X_1, X_2 \dots X_m]^T$，其中，每个$X_i$长度为$n$，都代表一个变量（行向量），该变量本身可以求期望和方差，不同的两个变量之间可以求协方差（比如X_1与X_2之间），这样，一个变量可以依次与n个变量（X_i与X_i的协方差即为其方差）求协方差，这样，就可以得到一个$m\times m$的矩阵$\mathbf{K}_{m\times m}$，该矩阵即为协方差矩阵，该过程表示如下：

$$\mathbf{K}_{n\times n}=\begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Var(X_2) & \cdots & Cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \cdots & Var(X_n) \end{bmatrix} \tag{1.1}$$

接下来，我们可以根据上文中协方差的公式，确定协方差矩阵中的每个元素的值。由脚注[1]可知$Cov(\mathbf{X})=E[(\mathbf{X}-E[\mathbf{X}])(\mathbf{X}-E[\mathbf{X}])^T]=E[\mathbf{X}\mathbf{X}^T]-E[\mathbf{X}]E[\mathbf{X}^T]$。特别地，当$\mathbf{X}$中每个行向量均值为零（比如气候研究中常用的“距平” anomaly的概念）时，上式最后一项$E[\mathbf{X}]E[\mathbf{X}^T]=0$，可得$Cov(\mathbf{X})=E[\mathbf{X}\mathbf{X}^T]$。由于$E[\mathbf{X}\mathbf{X}^T]$与$\mathbf{X}\mathbf{X}^T$仅差一个常数因子$\frac{1}{n}$（总体均值已知的情形，$E[\mathbf{X}\mathbf{X}^T]=\frac{1}{n}\mathbf{X}\mathbf{X}^T$）或$\frac{1}{n-1}$（适用于通过样本标准差推测总体标准差，即$E[\mathbf{X}\mathbf{X}^T]=\frac{1}{n-1}\mathbf{X}\mathbf{X}^T$），实际应用中，也将协方差矩阵简化表示为$Cov(\mathbf{X})=\mathbf{X}\mathbf{X}^T$.

气候场数据

回到气候数据，我们常见的单变量气候数据（比如SST或者SLP）一般都是三维的，分别代表{*time*,*lon*,*lat*}。但上述统计学概念都是定义在二维空间里的（行和列），如矩阵$\mathbf{X}$。为了将气候场数据和随机变量对应起来，通常的做法是，将空间维（也就是*lat*和*lon*）合并成一维，比如利用*lat*的弧度的余弦(cos)的平方根对数据加权，$weights=sqrt(cos(deg2rad(lat)))$[2]，然后一行一行首尾相接。这样，原来$time\times lon\times lat$的三维气候数据就变成了$time\times location$的二维数据，即在时间$t_1,t_2\dots t_n$和位置$l_1,l_2\dots l_p$处测量得到的某变量的值，将结果表示为$n$行（n个map，图1中的任一行，某一时间点的所有测量值组成一个map），p列（每个地图包含p个location，图1中的任一列）的矩阵$\mathbf{F}=[F_1,F_2,\cdots F_p]$（图1），其中，$F_i$为一列向量。对应机器学习中的叫法，map对应于sample，location对应于feature。所以我们可以说，矩阵$\mathbf{F}$包含n个map，每个map具有p个location，也可以说矩阵F包含n个sample，每个sample具有p个feature。这样，就回到了上段中的统计学问题，我们可以基于数据矩阵$\mathbf{F}$求其协方差矩阵$\mathbf{R}$。同样的，当移除$\mathbf{F}$中所有时间序列的均值后，每一个location所代表的时间序列为零均值的，因此，我们有$\mathbf{R}_{p\times p}=\mathbf{F}^T\mathbf{F}$。细心的同学会发现，此处转置符号$T$是在左边，也就是，一个$p\times n$的矩阵与一个$n\times p$的矩阵相乘，结果为$p\times p$的矩阵。这体现出与上文数学概念的另一处不同，此处$F_i$为一个列向量（不同于上文的行向量$X_i$），形象地表示一个location i处测得的时间序列，是在计算两个采样点（或像素点）处两个不同*时间*序列的协方差，而不是两个map的（空间）协方差，因此我们有时候也会看到时域协方差矩阵(temporal covariance matrix)的叫法[3]。

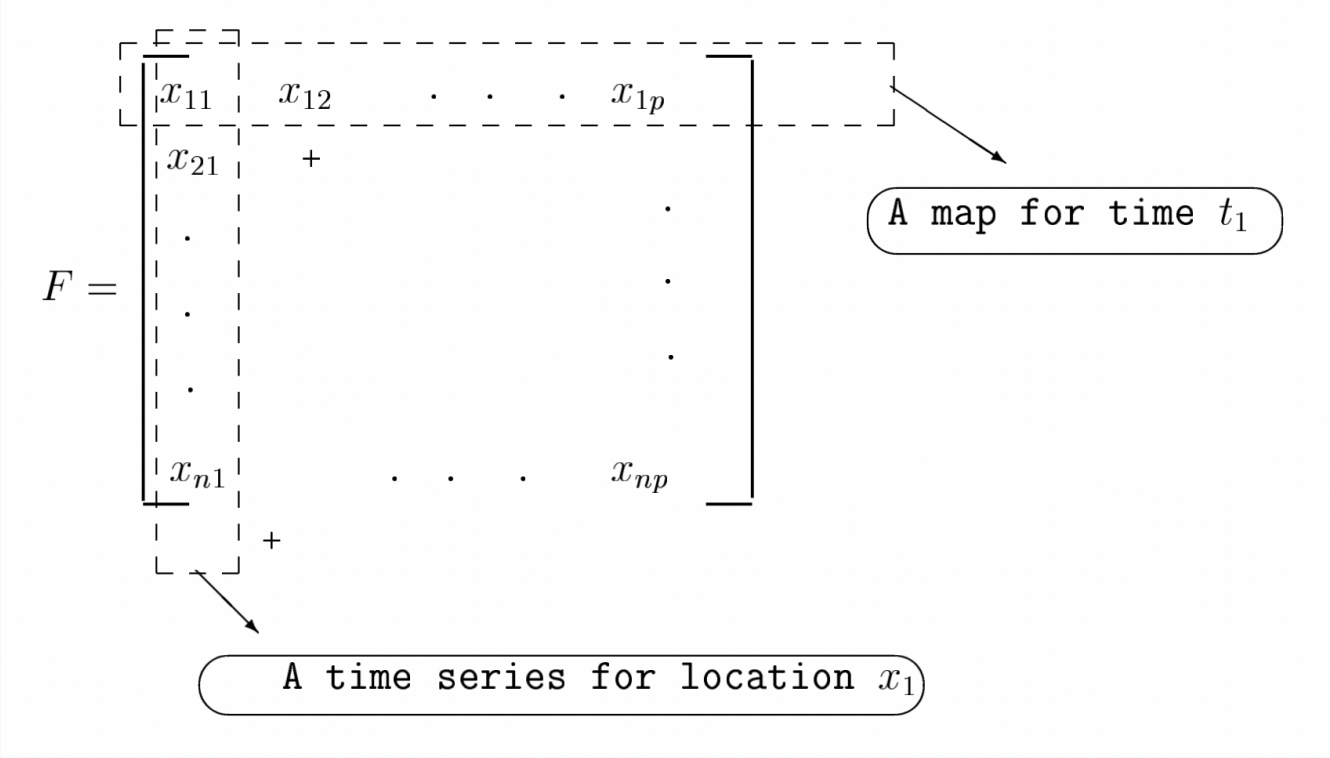

Fig 1 $time \times location$气候数据,每一行为一个地图，每一列为某一位置的时间序列，源自H.Bjoernsson, et, al

特征分解

可以说，对时空数据进行EOF分解，实际操作是对其对应的协方差矩阵进行**特征分解**（Eigendecomposition)。我对特征分解还是不能从感性上进行认识，但顾名思义，特征分解，是从被分解的数据中提取最具代表性（或者最包含最多“特征”）的成分，这里*特征*用两个概念表示，一个是**特征向量**(eigenvector)，表示为$\mathbf{v}$，一个是**特征值**(eigenvalue)，表示为λ，对于一个矩阵$\mathbf{A}$（比如上文数学基础部分的$\mathbf{X}$或者单变量气候场$\mathbf{F}$，尤其是协方差矩阵$\mathbf{K}$或者$\mathbf{R}$），其计算公式为

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{2.1}$$

其中，$\mathbf{v}$是一个向量，代表一个“方向”，λ是一个标量，可以表示该方向上的离散情况。我们这里不做推导，只要理解，对于一个矩阵，可以找到它的特征向量和特征值，使得它本身和特征向量的乘积，与特征向量平行。一个矩阵可以有多个特征向量，对应多个特征值。但公式(2.1)并没有给我“分解”的感觉，只是说一个矩阵可以满足这样的一个方程。那怎么理解此处的“分解”呢?
我们已经提到一个矩阵$\mathbf{A}$可以有多个特征值和特征向量，接下来我们构建一个矩阵$\mathbf{S}$，它的列由矩阵$\mathbf{A}$的所有特征向量（列向量）组成，称作**特征矩阵**，即：

$$\mathbf{S} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{v}_1 & \mathbf{v}_1 & \cdots & \mathbf{v}_l \\ | & | & \cdots & | \end{bmatrix} \tag{2.2}$$

公式(2.2)中竖线只是为了形象地表示矩阵的样子，l表示分解得到的特征向量的总数。

构建另一个对角矩阵 $\mathbf{\Lambda}$,它主对角线上的值为矩阵$\mathbf{A}$的特征值，即：

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{2.3}$$

很容易的推导[4]就可以得到，数据矩阵$\mathbf{A}$,特征矩阵$\mathbf{S}$和对角矩阵$\mathbf{\Lambda}$，满足以下关系：

$$\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^T \tag{2.4}$$

同时，有：

$$\mathbf{\Lambda} = \mathbf{S}^T\mathbf{A}\mathbf{S} \tag{2.5}$$

公式（2.4）实际上给我们一种"分解"的感觉。数学中习惯将公式（2.5）称作是矩阵$\mathbf{A}$的对角化。

EOF 分解

为了规范，我们将EOF分解得到的空间模态称为**EOF**s，对应的时间模态称为**PC**s。

我们已经知道了如何获得三维时空数据$\mathbf{F}$的的协方差矩阵$\mathbf{R}$，也知道了如何对矩阵$\mathbf{R}$进行特征分解。EOF分解实际上就是对协方差矩阵$\mathbf{R}$做特征分解，即将公式(2.1)中的$\mathbf{A}$换成$\mathbf{R}_{p\times p}$，即$\mathbf{R}\mathbf{v} = \lambda\mathbf{v}$，对于分解得到的每一个特征值$\lambda_i$（标量），都可以得到其对应的特征向量$\mathbf{v}_i$，其长度为$1 \times p$，恰好就是一个*map*的大小，将该map 去权重（de- weight）之后恢复成$lat \times lon$大小，便是要找的一个EOF。将所有的特征值λ_i 按从大到小的顺序排列，对应的EOF称为是$EOF1$，$EOF2\cdots$每一个特征值λ_i除以所有特征值的和的商（$SCF = \frac{\lambda_i}{\sum_{i=1}^{l}\lambda_i}$，分母也等于$\mathbf{\Lambda}$的迹），对应其对应模态（特征向量，EOF）解释的方差的百分比。将原数据与各个EOF做回归，便可得到对应的PC_1，$PC_2\cdots$，用列向量$\mathbf{a}_i$表示，$\mathbf{a}_i = \mathbf{F}\mathbf{v}_i$。

跟上文对应，数据矩阵表示为$\mathbf{F}_{n\times p}$，协方差矩阵为$\mathbf{R}_{p\times p}$，特征矩阵表示为$\mathbf{S}_{p\times l}$，对角矩阵为$\mathbf{\Lambda}_{l\times l}$，分解得到的时间模态矩阵为$\mathbf{A}_{n\times l}$，它的列由所有的$PC$（即所有的$\mathbf{a}_i$）组成。其中各个小写字母代表的数据意义如下：

> n代表时间维的长度；
> p代表空间维的长度（$lat \times lon$）;
> l代表分解得到的特征值（特征向量）的个数。

他们满足以下关系：

$$\mathbf{R} = \mathbf{F}^T F \tag{3.1}$$

$$\mathbf{R} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^T \tag{3.2}$$

因为$\mathbf{R}$是一个对称方阵（即$\mathbf{R}^T = \mathbf{R}$），有以下公式成立：

$$\mathbf{S}^T\mathbf{S} = \mathbf{I} \tag{3.3}$$

公式（3.3）体现各个特征向量之间的正交性。

$$\mathbf{A} = \mathbf{F}\mathbf{S} \tag{3.4}$$

以下的推导将带我们意识到EOF分解与奇异值分解的关系。

我们对时间模态矩阵$\mathbf{A}$做一些转换，定义另外一个矩阵$\mathbf{\Phi}$，它的的每一列定义如下：

$$\phi_i = \frac{\mathbf{a}_i}{\sqrt{\lambda_i}} \tag{3.5}$$

公式（3.5)可以看作是对时间模态的一个标准化。矩阵$\mathbf{\Phi}$满足以下关系[5]：

$$\mathbf{\Phi}^T\mathbf{\Phi} = \mathbf{I} \tag{3.6}$$

再定义另一对角矩阵$\mathbf{D}$，它主对角线元素为$\sqrt{\lambda_i}$，因此，$\mathbf{A} = \mathbf{\Phi D}$，带入公式（3.4），可得：

$$\mathbf{F} = \mathbf{A}\mathbf{S}^T = \mathbf{\Phi D}\mathbf{S}^T \tag{3.7}$$

公式（3.7）恰好就是对矩阵$\mathbf{F}$做奇异值分解分解的结果，所以我们最开始的说，EOF分解，相当于特征分解数据的协方差矩阵，或者奇异值分解分解数据本身。

奇异值分解

奇异值分解我们不做深入展示。对于矩阵$\mathbf{F}$，奇异值分解的形式为：

$$\mathbf{F} = \mathbf{U\Gamma V}^T \tag{4.1}$$

其中，$\mathbf{U}$ 为一个$n \times n$的正交矩阵，$\mathbf{V}$ 是一个$m \times m$的正交矩阵，$\mathbf{\Gamma}$是一个$n \times m$的对角矩阵，其元素（ρ个）为 $\Gamma_i = \delta_{i,j}\gamma_{i,j}$，$\gamma_{i,i}$称为是奇异值。

我们对比公式（2.4）和公式（4.1）可以看出二者之间的关系，如果$\mathbf{F}$是可逆方阵，则$\mathbf{U} = \mathbf{V}$，且他们的列中包含$\mathbf{F}$的特征向量，并且$\mathbf{\Gamma}$ 中包含特征值。为了让结果更明显，我们可以观察以下推导：

由公式（3.1）可知矩阵$\mathbf{A}$的协方差矩阵为：

$$\mathbf{R} = \mathbf{S\Lambda S}^T \tag{4.2}$$

将式（4.1）带入（3.1）,可得：

$$\mathbf{R} = (\mathbf{U\Gamma V}^T)^T\mathbf{U\Gamma V}^T = \mathbf{V\Gamma}^T\mathbf{U}^T\mathbf{U\Gamma V}^T = \mathbf{V\Gamma}^T\mathbf{\Gamma V}^T \tag{4.3}$$

对比式（4.2）和（4.3）可见，$\mathbf{S} = \mathbf{V}$，$\mathbf{\Lambda} = \mathbf{\Gamma}^T\mathbf{\Gamma}$，特征值$\lambda_i$和奇异值$\gamma_i$之间的关系为：

$$\lambda_i = \gamma_{i,i}^2 \tag{4.4}$$

后记

希望公式能增进我们对EOF分解的理解。利用EOF分解三维时空数据，是通过特征分解其时域协方差矩阵，但在实际计算中（如python eof 包）中，是利用奇异值分解分解数据本身。

值得注意的是，“数据本身”是指demean了之后的数据，即我们常见的anomaly数据。我们可以通过数学推导的视角思考以下问题，数据为什么需要求anomaly，从上文的计算来看，anomaly操作对于得到正确的协方差矩阵是必须的。另一个问题是，为什么是协方差矩阵？从前面的计算可以看出来，协方差矩阵保证了特征分解的对象是一个对称的方阵。但实际上，这些操作本身可能也有物理层面的考虑，比如demean的操作，实际上有一种“方差”的感觉在（没有平方，也没有加和的”方差“），所以量化的是一种时间域上的变异（variability）。而时域协方差矩阵，是求每两个空间位置所测得的时间序列之间的协方差，它实际上量化的是一种空间域上的变异。所以最后得到的协方差矩阵$\mathbf{R}$，算是一个很好的可以反映数据变异的矩阵。

---

1. 此处XY表示两个行向量的点乘，即X和Y中对应元素相乘并求和。如果将X和Y看作两个矩阵$\mathbf{X}_{1\times n}$和$\mathbf{Y}_{1\times n}$，则两个行向量的点积可以用矩阵乘积的形式表示如下$X \cdot Y = \mathbf{X}\mathbf{Y}^T$。相应地，如果二者均表示列向量，$X$和$Y$可看作两个矩阵$\mathbf{X}_{n\times 1}$和$\mathbf{Y}_{n\times 1}$，为了得到合理的结果（标量），则转置符号落在前者（即$X \cdot Y = \mathbf{X}^T\mathbf{Y}$）。这对于理解下文中的协方差矩阵非常有帮助。 ↩ ↩

2. 这种方式使得该数据的协方差矩阵以lat的弧度的余弦加权。 ↩

3. 实际上求各个map之间，也就是空间维上的协方差也是可以的，但是我们不展开。 ↩

4. 推导过程可以参考这个网易公开课视频。 ↩

5. 推导可以参考：$\mathbf{A}^T\mathbf{A} = (\mathbf{FS})^T\mathbf{FS} = \mathbf{SF}^T\mathbf{FS} = \mathbf{SRS} = \mathbf{\Lambda}$，带入公式（$3.5$）就可以得到公式（$3.6$）。 ↩