# CCA 与 SVD 分析方法比较研究*

严华生

(云南大学地球科学系, 昆明, 650091)

孟　捷　　李　艳

(云南大学统计系, 昆明, 650091)　　(湖南省气象台, 长沙, 410083)

摘　要

文中从理论分析、方法比较以及实例计算几方面, 对目前气象资料处理分析中常用的 CCA 与 SVD 分析方法进行了比较, 结果表明: (1) 对同样的变量组 $\boldsymbol{X}$, $\boldsymbol{Y}$ 分别用 CCA 和 SVD 方法进行相关分析, 得到了完全不同的分析结果, CCA 所得到的相关就是原始变量组 $\boldsymbol{X}$, $\boldsymbol{Y}$ 之间的相关; SVD 所得到的相关是组合变量 $\boldsymbol{L}$, $\boldsymbol{M}$ 间的相关而不是原始变量组 $\boldsymbol{X}$, $\boldsymbol{Y}$ 之间的相关。理论分析和实例计算都表明, 两种方法分析得到的最大相关有非常显著的差别, CCA 明显比 SVD 要大得多, 且 CCA 收敛快而 SVD 收敛慢。所以 SVD 不能有效地提取两组变量或两个变量场之间相关关系的主要特征, 只有 CCA 才能最大限度地提取它们之间相关关系的主要特征。(2) CCA 所得出的两变量组的变量是独立正交变量, 所以通过分析 CCA 组合变量间的相关来表示原变量组间的相关关系是有意义的。而 SVD 所得到的两变量组的变量不具有独立性和正交性, 信息提供重复, 存在共线性, 所以通过分析 SVD 组合变量来表示两变量组的相关关系没有 CCA 方法有意义。(3) CCA 是在考虑了各个变量场自身变化的情况下来分解两个变量场间的相关关系; 而 SVD 是在没有考虑各个变量场自身变化的情况下来分解两个变量场间的协方差关系。很显然, CCA 比 SVD 更全面、更完整和更准确。(4) 凡用 SVD 方法分析得到的结论, 由于总可以重新应用 CCA 方法找到相关更好的不同结果, 所以有值得进一步探讨的必要。

关键词: CCA, SVD, 特征向量, 组合变量, 相关。

## 1　引　言

在气象资料处理分析中, 一个重要的研究课题就是揭露两个气象场变化之间的相关关系。在这一领域内, 目前存在两种不同的分析方法, 一种是典型相关分析(canonical correlation analysis), 简称 CCA; 另一种是奇异值分解(singular value decomposition), 简称 SVD。而当我们揭露两个气象场变化之间相关关系[1~3]时, 必然考虑 “CCA 与 SVD 分析方法有什么不同? 为什么要用 CCA 分析方法而不用 SVD 分析方法?” 为此, 我们查阅了近 10 a 的气象学术刊物, 发现用 SVD 方法揭露两个气象场变化之间相关关系的论文[4~24]显著增多, 再查阅多元统计论著[25,26], 发现 CCA 来源于经典多元统计理论, 具有严密的数理基础; 而在经典多元统计理论中, 却找不到 SVD 的出处。很显然, SVD 是近 10 a 来新出现的一种资料分析方法, 它并非来源于经典多元统计理论, 也非数理统计学家所创。

对同一问题, 有上述两种不同的分析方法, 一般得出两种不同的结论, 那么究竟哪个更符合实际情况呢? 它们各自有什么特点? 它们之间的差别何在? 用哪一种方法更好? 这些问题确实很值得研究。本文针对这些问题, 通过理论分析、方法比较和实例计算来对 CCA 与 SVD 分析方法进行了比较研究。

## 2　CCA 与 SVD 分析方法对比分析

### 2.1　CCA 与 SVD 分析方法数学理论概述

设有 $n$ 个样本, $m$ 个变量组成的变量场, 记为 $\underset{n\times m}{\boldsymbol{X}}$; $n$ 个样本, $p$ 个变量组成的变量场, 记为 $\underset{n\times p}{\boldsymbol{Y}}$; (为便于讨论和公式写法的简洁, 设资料 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 都

* 初稿时间: 2002 年 10 月 5 日; 修改稿时间: 2002 年 12 月 20 日。
资助课题: 云南省重点基金项目(2003D00142), 国家科委基金项目(40065001) 共同资助。

已经过标准化处理，于是在公式推导中，计算协方差时略去了常数项 $1/n$)。研究 $\boldsymbol{X}$ 场与 $\boldsymbol{Y}$ 场变化之间的相关关系的基本思路为：以最大限度地提取 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 场变化之间相关关系的主要特征为准则，从 $\boldsymbol{X}$ 中提取组合变量 $\boldsymbol{L}$，从 $\boldsymbol{Y}$ 中提取组合变量 $\boldsymbol{M}$，如下式所示

$$\begin{cases} \underset{n\times 1}{\boldsymbol{L}} = \underset{n\times m}{\boldsymbol{X}}\ \underset{n\times m}{\boldsymbol{A}} \\ \underset{n\times 1}{\boldsymbol{M}} = \underset{n\times p}{\boldsymbol{Y}}\ \underset{p\times 1}{\boldsymbol{B}} \end{cases} \tag{1}$$

式中，$\boldsymbol{A}$，$\boldsymbol{B}$ 为线性变换，又称为空间特征向量。按式(1)把具有较多个变量的变量场 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 之间的相关化为较少组合变量 $\boldsymbol{L}$ 与 $\boldsymbol{M}$ 间的相关(注意 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相关与 $\boldsymbol{L}$ 和 $\boldsymbol{M}$ 之间相关的区别和联系)；通过 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的数值分布来确定 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的空间相关分布形式，而数值的大小则表示了所对应变量的重要程度。于是问题归结为如何求解 $\boldsymbol{L}$，$\boldsymbol{M}$，$\boldsymbol{A}$，$\boldsymbol{B}$。由于定解条件的不同，就形成了 CCA 和 SVD 方法。

2.1.1 CCA 方法

CCA 方法是在满足

$$\begin{cases} v(\boldsymbol{L}) = \boldsymbol{L}^{\mathrm{T}} L = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{A} = 1 \\ v(\boldsymbol{M}) = \boldsymbol{M}^{\mathrm{T}} \boldsymbol{M} = \boldsymbol{B}^{\mathrm{T}} \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \boldsymbol{B} = 1 \end{cases} \tag{2}$$

的约束条件下，使相关系数

$$r(\boldsymbol{L}, \boldsymbol{M}) = \frac{\mathrm{cov}(\boldsymbol{L}, \boldsymbol{M})}{\sqrt{v(\boldsymbol{L})\, v(\boldsymbol{M})}} = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{V}_{xy} \boldsymbol{B} \tag{3}$$

达到最大。应用拉氏乘子法，构造 $G$ 函数

$$G = r(\boldsymbol{L}, \boldsymbol{M}) - \lambda_1(\boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} - 1) - \lambda_2(\boldsymbol{M}^{\mathrm{T}} \boldsymbol{M} - 1) \tag{4}$$

求 $\frac{\partial G}{\partial A} = 0$，$\frac{\partial G}{\partial B} = 0$，得到方程组

$$\begin{cases} \boldsymbol{V}_{xy} \boldsymbol{B} - \lambda_1 \boldsymbol{V}_{xx} A = 0 \\ \boldsymbol{V}_{yx} \boldsymbol{A} - \lambda_2 \boldsymbol{V}_{yy} \boldsymbol{B} = 0 \end{cases} \tag{5}$$

根据相关系数 $r(\boldsymbol{L}, \boldsymbol{M}) = r(\boldsymbol{M}, \boldsymbol{L})$，可以证明 $\lambda_1 = \lambda_2$，于是由式(5)可得

$$\boldsymbol{V}_{yy}^{-1} \boldsymbol{V}_{yx} \boldsymbol{V}_{xx}^{-1} \boldsymbol{V}_{xy} \boldsymbol{B} - \lambda^2 \boldsymbol{B} = 0 \tag{6}$$

式中矩阵 $\boldsymbol{V}_{yy}^{-1} \boldsymbol{V}_{yx} \boldsymbol{V}_{xx}^{-1} \boldsymbol{V}_{xy}$ 称为 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的线性关联阵，于是就把 CCA 方法成为求 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的线性关联阵的特征值和特征向量问题，对此可再化为如下广义特征值问题求解

$$\boldsymbol{V}_{yx} \boldsymbol{V}_{xx}^{-1} \boldsymbol{V}_{xy} \boldsymbol{B} = \lambda^2 \boldsymbol{V}_{yy} \boldsymbol{B} \tag{7}$$

解此广义特征值问题，即可求出 $\boldsymbol{B}$；代入式(5)，即求出 $\boldsymbol{A}$；再代入式(1)，求出 $\boldsymbol{L}$，$\boldsymbol{M}$。其中，最大特征值为满足式(4)的最大相关系数，而对应的最大特征向量为由 $\boldsymbol{X}$，$\boldsymbol{Y}$ 提取的满足式(4)的线性组合。特征值按从大到小排列，将得到的第1特征值对应的($\boldsymbol{L}$，$\boldsymbol{M}$)和($\boldsymbol{A}$，$\boldsymbol{B}$)记为($\boldsymbol{L}_1$，$\boldsymbol{M}_1$)和($\boldsymbol{A}_1$，$\boldsymbol{B}_1$)…，第 $K$ 特征值对应的记为($L_K$，$M_K$)和($\boldsymbol{A}_K$，$\boldsymbol{B}_K$)，定义

$$\begin{cases} \bar{\boldsymbol{L}} = (\boldsymbol{L}_1 \cdots \boldsymbol{L}_K),\ \tilde{\boldsymbol{M}} = (\boldsymbol{M}_1 \cdots \boldsymbol{M}_K) \\ \bar{\boldsymbol{A}} = (\boldsymbol{A}_1 \cdots \boldsymbol{A}_K),\ \bar{\boldsymbol{B}} = (\boldsymbol{B}_1 \cdots \boldsymbol{B}_K) \end{cases} \tag{8}$$

则

$$\begin{cases} \bar{\boldsymbol{L}}^{\mathrm{T}} \bar{\boldsymbol{L}} = I \\ \tilde{\boldsymbol{M}}^{\mathrm{T}} \tilde{\boldsymbol{M}} = I \end{cases} \tag{9}$$

即由此得到的各组($\boldsymbol{L}_i$，$\boldsymbol{M}_i$)($i = 1, 2 \cdots k$)之间是互不相关的。

2.1.2 SVD 方法

传统的 SVD 方法是从求两组变量场协方差矩阵的左右特征向量来叙述的，但并未给出为什么可以对协方差矩阵进行奇异值分解的理论推导以及分解结果的统计学意义证明。文献[11]首先对此进行了分析，这里把 SVD 方法的理论推导和统计分析归纳总结如下。

在 CCA 方法中，改变约束条件为

$$\begin{cases} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} = 1 \\ \boldsymbol{B}^{\mathrm{T}} \boldsymbol{B} = 1 \end{cases} \tag{10}$$

求使协方差 $\mathrm{cov}(\boldsymbol{L}, \boldsymbol{M}) = \boldsymbol{L}^{\mathrm{T}} \boldsymbol{M} = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{V}_{xy} \boldsymbol{B}$ 达到最大的($\boldsymbol{L}$，$\boldsymbol{M}$)以及($\boldsymbol{A}$，$\boldsymbol{B}$)，即为 SVD 方法。应用拉氏乘子法，构造 $G$ 函数

$$G = \mathrm{cov}(\boldsymbol{L}, \boldsymbol{M}) - \lambda_1(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} - 1) - \lambda_2(\boldsymbol{B}^{\mathrm{T}} \boldsymbol{B} - 1) \tag{11}$$

求 $\frac{\partial G}{\partial A} = 0$，$\frac{\partial G}{\partial B} = 0$，得到方程组

$$\begin{cases} \boldsymbol{V}_{xy} \boldsymbol{B} - \lambda_1 \boldsymbol{A} = 0 \\ \boldsymbol{V}_{yx} \boldsymbol{A} - \lambda_2 \boldsymbol{B} = 0 \end{cases} \tag{12}$$

将约束条件式(10)代入式(12)可得：

$$\lambda_1 = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{V}_{xy} \boldsymbol{B} = (\boldsymbol{A}^{\mathrm{T}} \boldsymbol{V}_{xy} \boldsymbol{B})^{\mathrm{T}} = \lambda_2 \tag{13}$$

此式按 SVD 方法的传统叙述，即对两组变量或者两个变量场间的协方差矩阵进行奇异值分解为左右特征向量，由奇异值分解计算方法，结合式(12)和(13)成为：

$$\begin{cases} \boldsymbol{V}_{yx} \boldsymbol{V}_{xy} \boldsymbol{B} = \lambda^2 \boldsymbol{B} \\ \boldsymbol{V}_{xy} \boldsymbol{V}_{yx} \boldsymbol{A} = \lambda^2 \boldsymbol{A} \end{cases} \tag{14}$$

于是，对 $\boldsymbol{V}_{yx} \boldsymbol{V}_{xy}$ 或 $\boldsymbol{V}_{xy} \boldsymbol{V}_{yx}$ 进行特征值分解，最大特征值对应的特征向量 $\boldsymbol{A}$，$\boldsymbol{B}$ 以及相应的 $\boldsymbol{L}$，$\boldsymbol{M}$ 即为使式(11)达到最大的值。与 CCA 类似，对不同的特征值，可以得到一系列的($\boldsymbol{A}$，$\boldsymbol{B}$)及相应的($\boldsymbol{L}$，$\boldsymbol{M}$)可用于计算 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 同性和异性相关。特征值按从

大到小排列, 将 $V_{xy}V_{yx}$ 第 1 特征值对应的 $(\boldsymbol{L}, \boldsymbol{M})$ 和 $(\boldsymbol{A}, \boldsymbol{B})$ 记为 $(L_1, M_1)$ 和 $(A_1, B_1)$ …, 第 $K$ 特征值对应的记为 $(L_K, M_K)$ 和 $(A_K, B_K)$, 定义

$$\begin{cases} \overline{L} = (L_1 \cdots L_K), \widetilde{M} = (M_1 \cdots M_K) \\ \overline{A} = (A_1 \cdots A_K), \overline{B} = (B_1 \cdots B_K) \end{cases} \tag{15}$$

则

$$\begin{cases} \overline{A}^{\mathrm{T}} \overline{A} = I \\ \overline{B}^{\mathrm{T}} \overline{B} = I \end{cases} \tag{16}$$

即由此得到的各组 $(A_i, B_i)$ $(i = 1, 2 \cdots k)$ 之间是正交的, 但各组 $(L_i, M_i)$ $(i = 1, 2 \cdots k)$ 之间是相关的。

**2.2　CCA 与 SVD 对比分析**

2.2.1　CCA 与 SVD 方法的差异

首先, 由 2.1 节的讨论可知, 尽管对原始变量 $(\boldsymbol{X}, \boldsymbol{Y})$ 经过标准化处理后, $V_{xy} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 是相关矩阵, 但 CCA 和 SVD 讨论的是组合变量 $(\boldsymbol{L}, \boldsymbol{M})$ 而不是原始变量 $(\boldsymbol{X}, \boldsymbol{Y})$。从求极值的函数来看, CCA 方法为在组合变量 $(\boldsymbol{L}, \boldsymbol{M})$ 正交归一化的条件下对组合变量 $(\boldsymbol{L}, \boldsymbol{M})$ 间的相关系数求最大; 而 SVD 方法是将约束条件改变为在特征向量 $(\boldsymbol{A}, \boldsymbol{B})$ 正交条件下对组合变量 $\boldsymbol{L}, \boldsymbol{M}$ 间的协方差(注意不是相关系数) 求最大。我们知道由于度量单位和尺度的差异, 协方差不是相关性的一个好的度量, 即 SVD 方法旨在使协方差最大, 而不等价于使相关系数最大, 于是由此提取的线性组合无法很好地体现变量之间的相关。因此, 在一般情况下, 对同一问题, 用两种方法所得结果是完全不同的。只有当 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 均为正交归一时, 式(2) 的约束条件成为:

$$\begin{cases} v(\boldsymbol{L}) = \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A} = \boldsymbol{A}^{T}\boldsymbol{A} = 1 \\ v(\boldsymbol{M}) = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{M} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{B} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = 1 \end{cases} \tag{17}$$

且此时的线性组合 $(\boldsymbol{L}, \boldsymbol{M})$ 的方差为 1, 故极大化相关系数和极大化协方差是等价的, 只有此时两种方法所得结果才会相同。现有一种作法为: 先分别对 $\boldsymbol{X}, \boldsymbol{Y}$ 作 EOF 分解, 提出 $\boldsymbol{X}, \boldsymbol{Y}$ 的主成份, 再作 SVD 分析, 此时虽然主成份正交, 但其与自身转置矩阵相乘不是单位矩阵, 即方差不等 1, 而是为特征值所组成的对角阵。因而这种作法所得结果与 CCA 所得结果仍然不相同。

其次, 文献[25], [26] 从理论上已经证明 CCA 方法求得的特征值就是相关系数, $(\boldsymbol{L}, \boldsymbol{M})$ 间的相关结果就是 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 间的相关结果。但用 SVD 方法对矩阵分解得到的特征值不是相关系数, 进一步计算同性, 异性和模态相关系数表示的是 $L$ 与 $M$ 间的相关结果而不是 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 间的相关结果。

此外, 在统计上, 对 CCA 方法已导出其概率分布及显著性检验, 而对 SVD 方法至今未见这方面的研究, 因而尚未从理论上证明 SVD 分解在统计上是有显著意义的。

现把 CCA 和 SVD 分析结果列入表 1, 以便更清楚地看出它们之间空间模态、时间模态、相关结果和统计特征等方面的差异。

2.2.2　CCA 在变量压缩和信息提取方面均优于 SVD

CCA 把变量场 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分解为两两互不相关的独立变量 $\boldsymbol{L}$ 和 $\boldsymbol{M}$, 使 $\boldsymbol{L}$ 与 $\boldsymbol{M}$ 的相关达到最大, 这种相关是清楚的和有意义的, 并且在进行相关分析时不存在共线性问题; 然而 SVD 所分解出的变量 $\boldsymbol{L}$ 和 $\boldsymbol{M}$ 相互不独立, 即 $\boldsymbol{L}_i$ 与 $\boldsymbol{L}_j$ 之间和 $\boldsymbol{M}_i$ 与 $\boldsymbol{M}_j$ 之间存在相关性, 在这种情况下计算出的 $\boldsymbol{L}$ 与 $\boldsymbol{M}$ 间的相关系数, 不能排除其他变量间相互关系的影响作用, 所以不能准确地反映 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 间的真实相关。CCA 方法用最少的变量组合在最大限度上揭示 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 之间的相关关系; 相反, SVD 方法提取的线性组合之间是相关的, 信息提供重复, 而且存在共线性问题。而共线性问题是相关分析中十分复杂的问题之一。

表 1　CCA 和 SVD 分析结果比较

Table 1　The results contrast between CCA and SVD

| 分析结果 | 特征场 $\boldsymbol{A}, \boldsymbol{B}$ | 组合变量 $\boldsymbol{L}, \boldsymbol{M}$ | 相关结果 | 统计显著性检验 |
|---|---|---|---|---|
| CCA 方法 | 不正交 | 正交 | 特征值就是相关系数, 组合变量 $(\boldsymbol{L}, \boldsymbol{M})$ 与原始变量 $(\boldsymbol{X}, \boldsymbol{Y})$ 一致 | $\chi^2$ 检验 |
| SVD 方法 | 正交 | 不正交 | 须进一步计算同性和异性相关系数来表示, 组合变量 $(\boldsymbol{L}, \boldsymbol{M})$ 间的相关结果不是原始变量 $(\boldsymbol{X}, \boldsymbol{Y})$ 间的相关结果 | 无 |

#### 2.2.3 CCA 的典型相关系数大于 SVD 的模态相关系数

SVD 的模态相关系数定义为

$$r_{SVD}(\boldsymbol{L}_{SVD}, \boldsymbol{M}_{SVD}) = \frac{\boldsymbol{L}_{SVD}^{T}\boldsymbol{M}_{SVD}}{\sqrt{v(\boldsymbol{L}_{SVD})v(\boldsymbol{M}_{SVD})}} \quad (18)$$

由于 SVD 是使协方差达到最大而 CCA 是使相关达到最大,而相关是由协方差与方差之比构成,所以仅使协方差达到最大不能保证就使相关达到最大,根据前面的讨论,于是进一步有

$$r_{SVD}(\boldsymbol{L}_{SVD}, \boldsymbol{M}_{SVD}) \leqslant \underset{L,M}{\mathrm{Max}} \frac{\boldsymbol{L}^{T}\boldsymbol{M}}{\sqrt{v(\boldsymbol{L})v(\boldsymbol{M})}}$$

$$= \underset{\substack{v(L)=1 \\ v(M)=1}}{\mathrm{Max}} \boldsymbol{L}^{T}\boldsymbol{M} = r_{CCA}(\boldsymbol{L}_{CCA}, \boldsymbol{M}_{CCA}) \quad (19)$$

即:SVD 定义的模态相关系数小于 CCA 得到的典型相关系数。

#### 2.2.4 CCA 的典型相关系数是传统意义上相关系数的拓广和延伸

记 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的方差协差阵为 $\boldsymbol{V}=\begin{pmatrix} \boldsymbol{V}_{xx} & \boldsymbol{V}_{xy} \\ \boldsymbol{V}_{yx} & \boldsymbol{V}_{yy} \end{pmatrix}$,定义它的行列式为广义方差,记为

$$|\boldsymbol{V}_{xx}\boldsymbol{V}_{yy} - \boldsymbol{V}_{xy}\boldsymbol{V}_{yx}| = |\boldsymbol{V}_{xx}||\boldsymbol{V}_{yy} - \boldsymbol{V}_{yx}\boldsymbol{V}_{xx}^{-1}\boldsymbol{V}_{xy}| \quad (20)$$

则 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 之间的广义相关为

$$\frac{\begin{vmatrix} \boldsymbol{V}_{xx} & \boldsymbol{V}_{xy} \\ \boldsymbol{V}_{yx} & \boldsymbol{V}_{yy} \end{vmatrix}}{|\boldsymbol{V}_{xx}||\boldsymbol{V}_{yy}|} = \frac{|\boldsymbol{V}_{xx}||\boldsymbol{V}_{yy} - \boldsymbol{V}_{yx}\boldsymbol{V}_{xx}^{-1}\boldsymbol{V}_{xy}|}{|\boldsymbol{V}_{xx}||\boldsymbol{V}_{yy}|}$$

$$= 1 - |\boldsymbol{V}_{yy}^{-1}\boldsymbol{V}_{yx}\boldsymbol{V}_{xx}^{-1}\boldsymbol{V}_{xy}| \quad (21)$$

CCA 是对矩阵 $\boldsymbol{V}_{yy}^{-1}\boldsymbol{V}_{yx}\boldsymbol{V}_{xx}^{-1}\boldsymbol{V}_{xy}$ 进行特征值和特征向量分解,而 SVD 是对矩阵 $\boldsymbol{V}_{yx}\boldsymbol{V}_{xy}$ 进行特征值和特征向量分解。从所分解矩阵的不同可看出,CCA 是在考虑了各个变量场自身变化的情况下来分解两个变量场间的相关关系;而 SVD 是在没有考虑各个变量场自身变化的情况下来分解两个变量场间的协方差关系。在 CCA 所分解的矩阵 $\boldsymbol{V}_{yy}^{-1}\boldsymbol{V}_{yx}\boldsymbol{V}_{xx}^{-1}\boldsymbol{V}_{xy}$ 中,当 $m=1, p=1$ 时导出两个变量间的单相关系数;当 $m\neq1, p=1$ 时导出一个因变量对多个自变量间的复相关系数;当 $m=1, p\neq1$ 时导出多个因变量对一个自变量间的全相关系数,当 $M\geqslant1, P\geqslant1$ 时导出多因变量对多自变量间的广义相关系数。所以 CCA 是以往统计理论中单、复、全相关分析的进一步推广,而 SVD 就没有这些性质。很显然,CCA 比 SVD 更全面、更完整和更准确。

## 3 CCA 和 SVD 的实例比较

### 3.1 资 料

取云南省 9 个站(昭通、丽江、腾冲、保山、大理、昆明、沾益、玉溪、蒙自) 1951~1999 年 5 月雨量场 $\boldsymbol{X}_{49\times9}$ 和雨季 5~10 月总雨量场 $\boldsymbol{Y}_{49\times9}$ 进行分析,从而研究两场之间的相关关系。

### 3.2 方 法

分别用上述 CCA 和 SVD 方法对 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 进行分析,得到 CCA 和 SVD 的结果,再用下面的多因变量矩阵回归[26,27]相关分析来比较它们的差别。

若 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 之间存在线性关系,则可得它们的矩阵回归方程为

$$\hat{\boldsymbol{Y}} = \boldsymbol{XB} \quad (22)$$

由方差分析,设 $L_{yy}$ 为因变量总方差,$U$ 为回归平方和,$Q$ 为残差平方和,则有

$$L_{yy} = U + Q \quad (23)$$

可用下面的总复相关系数表达式

$$R_{总}^{2} = 1 - \frac{|Q|}{|L_{yy}|} \quad (24)$$

来表示 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 之间线性相关的紧密程度,并用它来对 CCA 和 SVD 方法进行评判。

对两组变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,分别用 CCA 和 SVD 方法得到前 K 个组合变量 $\boldsymbol{L}$ 和 $\boldsymbol{M}$ 及其回归方程 $\boldsymbol{M}=\boldsymbol{LC}$,其中 $\boldsymbol{C}$ 为回归系数矩阵,按式(24)计算前 $K$ 个组合变量 $\boldsymbol{L}$ 和 $\boldsymbol{M}$ 间的总复相关系数(表 2 中第 2,4 列)。再将式(1)代入回归方程 $\boldsymbol{M}=\boldsymbol{LC}$,即可求得返回原始变量 $\boldsymbol{X}$, $\boldsymbol{Y}$ 后的回归方程和总复相关系数(表 2 中第 3,5 列)。表 2 列出了用 CCA 和 SVD 所得的前9个组合变量及返回原始变量后的

表 2 两种方法计算所得的总复相关系数比较

Table 2 The contrast of multiple correlation coefficient between two methods

| 前 $K$ 个组合变量 | CCA 方法 | | SVD 方法 | |
|---|---|---|---|---|
| | $R_{总}$ | $R'_{总}$ | $R_{总}$ | $R'_{总}$ |
| 1 | 0.7095 | 0.7095 | 0.3962 | 0.5099 |
| 2 | 0.8227 | 0.8227 | 0.5344 | 0.6331 |
| 3 | 0.8875 | 0.8875 | 0.7382 | 0.8006 |
| 4 | 0.9237 | 0.9237 | 0.8187 | 0.8588 |
| 5 | 0.9394 | 0.9394 | 0.8546 | 0.8995 |
| 6 | 0.9470 | 0.9470 | 0.9062 | 0.9259 |
| 7 | 0.9523 | 0.9523 | 0.9295 | 0.9381 |
| 8 | 0.9547 | 0.9547 | 0.9418 | 0.9469 |
| 9 | 0.9547 | 0.9547 | 0.9547 | 0.9547 |

总复相关系数 $R_{总}$ 和 $R'_{总}$。

由表 2 可以看出: CCA 方法得到的两组变量总相关明显高于 SVD 方法, 尤其是最大特征值所对应的第一主分量有特别显著的差异。而且, CCA 方法返回原始变量前后的总复相关系数总是一致, 说明 CCA 所得到的相关就是原始变量组 $\boldsymbol{X}$, $\boldsymbol{Y}$ 之间的相关; SVD 方法返回原始变量前后的总复相关系数不一致, 说明 SVD 所得到的相关是组合变量 $\boldsymbol{L}$, $\boldsymbol{M}$ 间的相关而不是原始变量组 $\boldsymbol{X}$, $\boldsymbol{Y}$ 之间的相关。CCA 方法所得的总复相关系数收敛快, SVD 方法所得的总复相关系数收敛慢。

## 3 结 论

通过以上理论分析、方法比较以及实例计算说明:

(1) 对同样的数据 $\boldsymbol{X}$, $\boldsymbol{Y}$, 分别用 CCA 和 SVD 方法进行相关分析, 得到完全不同的分析结果, CCA 所得到的相关就是原始变量组 $\boldsymbol{X}$, $\boldsymbol{Y}$ 之间的相关; SVD 所得到的相关是组合变量 $\boldsymbol{L}$, $\boldsymbol{M}$ 间的相关而不是原始变量组 $\boldsymbol{X}$, $\boldsymbol{Y}$ 之间的相关。理论分析和实例计算都表明, 两种方法分析得到的最大相关有非常显著的差别, CCA 明显比 SVD 要大得多。且 CCA 收敛快而 SVD 收敛慢。所以 SVD 不能有效提取两组变量或两个变量场之间相关关系的主要特征, 只有 CCA 才能最大限度地提取它们之间相关关系的主要特征。

(2) CCA 所得出的两变量组的组合变量是独立正交变量, 所以通过分析 CCA 组合变量间的相关来表示原变量组间的相关关系是有意义的。而 SVD 所得到的两变量组的组合变量不具有独立性和正交性, 信息提供重复, 存在共线性, 所以通过分析 SVD 组合变量来表示两变量组的相关关系没有 CCA 方法有意义。

(3) CCA 是在考虑了各个变量场自身变化的情况下来分解两个变量场间的相关关系; 而 SVD 是在没有考虑各个变量场自身变化的情况下来分解两个变量场间的协方差关系。很显然, CCA 比 SVD 更全面、更完整和更准确。

(4) 凡用 SVD 方法分析得到的结论, 由于总可以重新应用 CCA 方法找到相关更好的不同结果, 所以有值得进一步探讨的必要。

## 参考文献


1 严华生, 卞福久. 用多因变量多自变量统计方法建立的大气环流——区域水稻产量预报模式. 气象学报, 1988, 46(1): 127~128

2 段旭, 严华生, 董谢琼. 云南 8 月气温与春季气温场的典型相关分析. 高原气象, 1999, 18(2): 192~198

3 晏红明, 严华生, 谢应齐. 中国汛期降水的印度洋SSTA 信号特征分析. 热带气象学报, 2001, 17(2): 109~116

4 Prohaska J. A technique for analyzing the linear relationships between two meteorological fields. Mon Wea Rew, 1976, 104: 1345~1353

5 Lanzante J R. A rotated eigen analysis of the correlation between 700 mb height and sea surface temperature in the Pacific and Atlantic. Mon Wea Rew, 1984, 112: 2270~2280

6 Wallace J M, Smith C, Bretherton C S. Singular value decomposition of wintertime sea surface temperature and 500 hPa height anomalies. J Climate, 1992, 5: 567~576

7 Bretherton C S, Wallace J M. An intercomparison of methods for finding coupled patterns in climate data. J Climate, 1992, 5(6): 541~559

8 徐瑞珍, 张先恭. 经验正交函数在两个气象场相关分析中的应用. 气象学报, 1982, 40(1): 117~122

9 孙照渤, 章基嘉, 华莱士 J M. 冬季北大西洋地区海表温度与 500 hPa 高度奇异值分解. 南京气象学院学报, 1991, 14(3): 287~293

10 江志红, 丁裕国. 我国夏半年降水距平与北太平洋海温异常的奇异值分解法分析. 热带气象学报, 1995, 11(2): 133~141

11 江志红, 丁裕国. SVD 方法在气象场诊断分析中的普适性. 气象学报, 1996, 54(3): 365~372

12 魏凤英, 张先恭. 北太平洋海表温度与中国夏季气温的耦合特征. 见: 曹鸿兴等编. 我国短期气候变化与成因研究. 北京: 气象出版社, 1996. 134pp

13 缪锦海, 肖天贵. 复奇异值分解方法的构造及度量特征. 成都气象学院学报, 1996, 11(3): 97~105

14 谢炯光, 秦冰冰, 王静渊. 奇异值分解方法在季降水预测中的应用. 气象学报, 1997, 55(1): 118~123

15 魏凤英, 曹鸿兴. 奇异值分解及其在北美陆地气温与我国降水遥相关中的应用. 高原气象, 1997, 16(2): 174~182

16 王盘兴, 周伟灿, 王欣等. 气象向量场奇异值分解方法及其应用. 南京气象学院学报, 1997, 20(2): 152~157

17 王盘兴, 周伟灿, 薛志华等. 赤道太平洋区域风应力与海表温度年际异常的奇异值分解. 应用气象学报, 1998, 9(3): 265~272

18 吴洪宝. 我国东南部夏季干旱指数研究. 应用气象学报, 2000, 11(2): 137~144

19 朱伟军, 孙照渤. 冬季黑潮区域海温异常对北太平洋风暴轴的影响. 应用气象学报, 2000, 11(2): 145~153

20 朱伟军, 孙照渤. 冬季北太平洋风暴轴的年际变化及其与 500 hPa 高度以及热带和北太平洋海温的联系. 气象学报, 2000, 58(3): 309~319
21 祝从文, 何金梅, 吴国雄. 东亚季风指数及其与大尺度热力环流年际变化关系. 气象学报, 2000, 58(4): 391~401
22 肖天贵, 缪锦海. El Niño 期间移动性海温场与大气环流相互关系的复奇异值分解. 气象学报, 2000, 58(4): 418~427
23 李跃清. 青藏高原上空环流变化与其东侧旱涝异常分析. 大气科学, 2000, 24(4): 470~476
24 李耀辉, 李栋梁, 赵庆云. 中国西北春季降水与太平洋秋季海温的异常特征及其相关分析. 高原气象, 2000, 19(1): 100~110
25 张尧庭, 方开泰. 多元统计分析引论. 北京: 科学出版社, 1983. 580pp
26 严华生, 王学仁. 多因变量及要素场统计预报. 北京: 气象出版社, 1991. 33~41
27 严华生, 张永昆, 曹杰等. 多因变量矩阵回归预报方法. 大气科学, 1997, 21(4): 493~498

# THE STUDY ON CCA AND SVD ANALYTICAL METHODS

Yan Huasheng

*(Department of Atmospheric Science, Yunnan University, Kunming 650091)*

Meng Jie

*(Department of Statistics, Yunnan University, Kunming 650091)*

Li Yan

*(Department of Atmospheric Science, Yunnan University, Kunming 650091)*

## Abstract

An important research subject is to reveal the correlation of two meteorological fields while analyzing in the meteorological data. In this field, two kinds of different analytical methods exist at present: one is canonical correlation analysis, abbreviated as CCA; the other is singular value decomposition, abbreviated as SVD. CCA and SVD is compared using theory analysis, method comparison and example calculation. Four principal results have been achieved as follows. Firstly, to the same data **X** and **Y**, after carrying on relevant analysis with CCA and SVD method separately, the results are total different. The results indicate that SVD cannot distill correlativity of two variable fields effectively, but CCA can distill them as effectively as possible. Correlation of CCA combination variable fields can embody more correlativity of original variable fields than that of SVD, and using CCA eigenvector fields to express their correlation distribution are significant than using that of SVD. The example calculation also shows that maximal correlations obtained from the two methods have prominent difference and the value of correlation coefficient of CCA is obviously bigger than that of SVD. Secondly, it is more significant to study the correlations of source variable fields through studying the correlation of combination variable fields using CCA than that of using SVD. The combination variable fields obtained from CCA are independent and orthogonal but not for SVD. Thirdly, it is obvious that CCA is more fully and truly embody the correlativity of original variable fields than that of SVD for CCA takes into account that the self change of each variable fields when studying original variable fields but not for SVD. Finally, it is deserved to study more about the results obtained from SVD in respect that the correlation between x and y is significant if CCA is adopted.

**Key words:** CCA, SVD, Eigenvector, Combination variables, Correlation.