子空间降维算法研究与应用

姜 伟 著

科学出版社

子空间降维算法研究与应用

姜 伟 著

科学出版社

北 京

内 容 简 介

本书结合作者近几年的相关研究工作，全面系统地介绍子空间降维的概念、主要原理、经典方法和国内外有关研究的最新成果。第 1～2 章介绍子空间降维的基本内容包括发展概述与基本方法；第 3～6 章介绍作者关于子空间降维的最新研究成果；第 7 章引入一些秩极小化方法。本书选取一些经典方法进行介绍，并结合作者的研究成果加以论述，较好地反映了该研究领域的全貌，并具有一定的关联性与层次性，便于初学者学习和使用。

本书可供大数据、信号处理、模式识别、机器学习、计算机视觉、数据挖掘等领域的科研人员参考。

图书在版编目(CIP)数据

子空间降维算法研究与应用/姜伟著. —北京：科学出版社，2015

ISBN 978-7-03-043657-3

Ⅰ. ①子… Ⅱ. ①姜… Ⅲ. ①子空间-计算机技术-研究

Ⅳ. ①O186.14 ②TP3

中国版本图书馆CIP数据核字(2015)第046247号

责任编辑：董素芹/ 责任校对：郭瑞芝

责任印制：赵 博/ 封面设计：迷底书装

科 学 出 版 社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销

*

2015 年 3 月第 一 版 开本：720 ×1 000 1/16

2015 年 3 月第一次印刷 印张：8 1/2

字数：156 000

定价：45.00 元

（如有印装质量问题，我社负责调换）

本书由大连市学术著作出版基金资助出版

The published book is sponsored by the Dalian Evaluation Committee for Publishing Academic Works Financed

前　言

近年来，随着数据获取技术的飞速发展，高维数据不断涌现。对高维数据的分析不仅成为模式识别的重要任务，也给模式识别研究带来了极大的挑战。高维问题不但会显著地增加计算和存储代价，更严重的是导致“维数灾难”。本书对基于子空间的降维技术进行了探索，发展出的一系列新算法在某些标准数据集上获得了良好的性能。

在编写方针上，本书从科研的角度出发，注重理论性、实用性、系统性和前瞻性，既参考了许多有关文献，也结合了作者多年来在相关领域的研究成果，有的成果在国内外重要学术刊物与国际学术会议上发表，体现了内容的创新性，学术思想的新颖性和重要的理论、应用价值。目前，国内缺乏子空间降维方面的学术专著，而广大研究生和科技工作者迫切需要了解本领域前沿的最新进展，以满足科研的工作需要，本书颇具理论与应用价值。

在内容选取上，本书围绕子空间降维技术展开，覆盖了线性子空间降维与非线性子空间降维以及这两方面的整合方法。子空间降维是一个新兴的研究方向，新的技术方法和工具不断涌现，真可谓五花八门、层出不穷。因此，在一本书中不可能包含子空间降维的全部内容。本书主要从理论与应用角度讲述子空间降维的基本原理、概念和技术方法，同时也尽量注意到全面性和系统性。希望本书既能对广大工程人员的实际应用有所帮助，又能为科技人员深入研究该方向奠定良好的基础。

全书共 7 章，第 1 章介绍子空间降维技术在数据挖掘和机器学习中的重要作用，简要介绍降维的概念和含义，回顾目前流行的子空间降维算法。

第 2 章按照局部和全局子空间算法的分类原则，详细介绍其中比较经典的几种子空间学习算法，并分析这些算法的各自特点和优缺点，为后面的章节做铺垫。

第 3 章简单介绍核函数的定义，讨论二维线性判别分析及其实质，回顾多核的定义，提出将多核方法推广到二维判别分析，阐述算法的推导过程，给出相应实验设置和实验结果。

第 4 章讨论全局谱嵌入和局部谱嵌入的数学基础，简单介绍图的基本概念和图的 Laplacian 及其性质。回顾 LPP 算法、LPP 算法与 PCA 算法、LDA 算法之间的关系，扼要说明子空间学习算法可以统一在谱图理论的降维方法框架下，提出自适应半监督边界费希尔分析算法。

第 5 章简述非负矩阵分解的含义，回顾非负矩阵分解及其比较流行的诸多改

进算法，提出半监督凸非负矩阵分解算法，算法的目的是寻求一个嵌入映射，既考虑矩阵分解的非负性又考虑由带标签数据与不带标签数据所推出数据的内在几何结构，构造了一个有效的乘积更新算法并且在理论上证明算法的收敛性。

第 6 章介绍格拉斯曼核的定义，回顾半监督判别分析，提出一个新的格拉斯曼流形上的半监督判别分析方法，将其应用于图像集合的识别问题。

第 7 章回顾压缩感知与矩阵低秩恢复的定义，重点描述矩阵低秩恢复最新研究成果。

本书编写前征求了辽宁师范大学数学学院张永清副教授对写作大纲的意见，初稿完成后又征求了辽宁师范大学数学学院张新立副教授和北京科技大学杨炳儒教授等对内容编排的意见，承蒙他们给予真诚的鼓励并且提出了许多宝贵的建议。此外，本书由大连市学术著作出版基金资助出版。在此表示衷心的感谢！

由于作者水平有限，书中难免存在疏漏与不足之处，希望读者批评指正。

作　者

2015 年 3 月 1 日

目　录

第1章 绪 论

随着数据存取技术的快速发展，数据库中积累了大量数据，面临着如何从海量数据中提取有用知识的问题，数据挖掘[1,2]的出现，为人们提供了一条解决“数据丰富而知识贫乏”的有效途径。数据挖掘的定义有含义相同、描述不同的多个版本，比较多的一种定义是“在海量数据中识别出有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程”。从定义中可以看出，数据挖掘是数据库技术和机器学习的交叉，利用数据库技术对数据进行存储和管理，利用机器学习技术对数据进行分析。数据挖掘技术已经在多学科中得到广泛的应用，特别是在计算机视觉和生物信息等学科已经成为最流行的技术。

数据挖掘最初被认为是数据库中知识发现(Knowledge Discovery in Databases, KDD)的一个阶段[3]。后来，数据挖掘被看成 KDD 的同义词。数据挖掘是一门多学科交叉的学科，用到了人工智能、机器学习、数据库、统计学等领域的技术和方法，吸引了许多计算机工作者[4-23]投入这方面的研究中。

数据挖掘中面临的一个重要的问题是如何处理海量、高维、非线性的数据。高维数据的大量涌现，对数据挖掘和机器学习提出了挑战，高维数据不但会提高存储和计算代价，而且会导致维数灾难[24]。也就是说，随着维数的增加，为了保持分类器的性能，样本的数量需呈指数增加。维数灾难现象的几何现象是空空间现象，即高维空间本质上是稀疏空间。当样本的数量远小于样本的维数时，将导致小样本问题。高维数据的另一个困难是难以被人理解，行之有效的方法就是空间降维。数据降维不但是解决高维数据分析的有效手段，也是高维数据可视化的重要途径。因此，对数据进行降维成为机器学习和数据挖掘的重要研究课题。

降维包括特征选择和特征提取两种途径。特征选择是指在所有特征中选择最具代表性一些特征，得到原有特征的一个真子集。特征提取是指通过对原始特征进行线性组合，得到更有意义的低维投影。在本书中所讨论的算法都基于特征提取方法，通过变换将数据从高维空间变换到低维特征空间。

根据不同的分类标准，子空间降维算法可有不同分类。

(1)根据所处理数据的分布，降维算法可分为线性和非线性。典型的线性算法有主成分分析(Principal Component Analysis, PCA)[25]、线性判别分析(Linear Discriminant Analysis，LDA)[26]等。非线性算法有等距映射(Isometric Feature Mapping, ISOMAP)[27]和拉普拉斯特征映射 (Laplacian Eigenmaps，LE)[28]等。

(2)根据是否利用标签信息，降维算法可分为有监督算法和无监督算法两类。

PCA[25]、局部保持映射(Local Preserving Projections，LPP)[29]和局部线性嵌入(Locally Linear Embedding，LLE)[30]等是无监督降维学习算法。LDA[26]和最大间隔准则(Maximum Margin Criterion，MMC)[31]等是有监督降维学习算法。无监督降维学习算法的目标是使数据在降维后信息损失最小，而有监督降维学习算法的目标是最大化各类别之间的鉴别信息。

(3)根据算法是否计算每个数据点与所有其他数据点的关系，降维算法可分为局部和全局降维算法。近邻保持嵌入(Neighborhood Preserving Embedding，NPE)[32]和LPP[29]等均为局部降维算法。PCA[25]和LDA[26]等为全局降维算法。

尽管将子空间学习算法按照上述方式进行了分类，但是，它们又是相互融合、互相交叉的，如LDA是一种有监督全局线性算法。

子空间方法（subspace method）的基本思想是把高维空间数据，通过线性或非线性变换到低维的子空间中，使数据在低维空间中更利于分类，降低计算复杂度。在线性降维算法中，最著名的算法有PCA[25]、LDA[26]、典型相关分析(Canonical Correlation Analysis, CCA)[33]、多维尺度分析(Multidimensional Scaling, MDS)[34]、非负矩阵分解(Non-negative Matrix Factorization, NMF)[35]等。PCA是无监督学习算法，算法的原理是寻找一组最佳的正交基向量，使重构误差达到最小。LDA是有监督学习算法，算法的原理是寻找一个投影方向使得沿着该方向同类样本之间的离散度最小，而异类样本之间的离散度最大。在分类问题上，LDA比PCA更具优势，但是LDA要求样本为高斯分布。CCA是一种无监督方法，算法的目标是求得一对基向量，使得两数据集之间的相关最大，它只关注成对样本之间的相关性，并将相关作为不同空间中样本之间的相关性度量。MDS的基本思想是寻找高维数据的低维表示，忠实地保持输入模式内积的低维描述，也就是降维后低维空间中任意两点之间的距离应该与原高维空间中的距离尽量接近。NMF的基本思想是在非负性约束下，对非负矩阵进行非负分解，采用简单有效的乘性迭代算法，通过学习得到基向量中含有关于物体的局部特征信息。NMF反映局部和整体之间的关系，整体是局部的非负线性组合，局部特征在构成整体时不会出现正负抵消的情况。

线性降维算法只能发现数据中的全局线性结构，无法揭示数据内在的非线性结构。为了解决这一问题，研究人员提出了许多非线性降维算法。具有代表性的有两类，分别是核方法和流形学习方法。核方法的基本思想是首先将数据从原始的非线性空间映射到一个更高维甚至是无限维的特征空间，然后再利用传统的线性方法在该特征空间中对数据进行处理。利用核技巧，大多数传统线性降维方法都可以推广到非线性的情况，典型的有核主成分分析(Kernel Principle Component Analysis, KPCA)[36]、核Fisher判别分析(Kernel Fisher Discriminant Analysis, KFDA)[37]和核典型相关分析(Kernel Canonical Correlation Analysis, KCCA)[38]等。核方法的

一个缺点就是如何选择合适的核函数和核函数中的参数，核函数和参数的变化会隐式地改变从输入空间到特征空间的映射，进而影响分类效果，影响核方法的性能；另一个缺点是核方法往往依赖于某种隐式映射，不易直观地理解其降维机理[39]。

另一类非线性的降维方法是基于流形学习的方法，流形是描述许多自然现象的一种空间形式，是欧氏空间在大尺度分析情况下的推广，而欧氏空间是它的特例。流形在局部上与欧氏空间存在着同胚映射，因此，从局部上看，流形与欧氏空间几乎一样。2000 年 *Science* 上发表的 ISOMAP[27]、LLE[30]和感知的流形假说[30]，被认为是流形学习的研究热潮开始的标志，以后又相继发现许多流形学习算法，包括 LE[28]、局部切空间排列(Local Tangent Space Alignment, LTSA)[40]、最大方差展开（Maximum Variance Unfolding, MVU）[41]、黑塞局部线性嵌入（Hessian Locally Linear Embedding, HLLE）[42]和局部坐标排列（Local Coordinates Alignment, LCA）[43]等。流形学习的一个缺点导致“外样本”问题[44]，对于新的样本，算法无法直接得到它在低维空间中对应的坐标。针对这个问题，多个线性化版本被提出，如基于 LLE 的 ONPP(Orthogonal Neighborhood Preserving Projection)[45]、NPE[32]，基于 LE 的 LPP[29]和基于 LTSA 的线性局部切空间排列（Linear Local Tangent Space Alignment, LLTSA）等。流形学习的另一个缺点是由于算法是基于样本的局部结构的，都涉及邻域选择问题。

经典的流形学习算法是无监督的学习算法，不适用于分类问题。在流形学习算法中引入标签信息，许多学者提出了一些针对分类目的的监督流形学习算法。监督流形学习算法也包括非线性算法和线性算法。前者只定义在训练集上，可以做数据分析的工具；后者能够提供显式的线性映射，可用于模式识别中的特征提取。引入监别信息的非线性算法有 Supervised-LLE[46]、LFE(Local Fisher Analysis)[47]、1-DA(local Discriminant Analysis)[48]。线性有监督流形学习算法的典型算法有 LDE(Local Discriminant Embedding)[45]和 MFA(Marginal Fisher Analysis)[49]。MFA 与 LDE 并无本质上的差异。在 LDE 和 MFA 中，目标函数定义的形式不同，但它们可以转化为同样形式的广义特征值问题。以后又相继出现半监督判别分析(Semi-Supervised Discriminant Analysis, SDA)[50]、正交局部保持投影(Orthogonal Locality Preserving Projections, OLPP)[45]、判别局部线性嵌入(Discriminant Locally Linear Embedding，DLLE)[51]等。这些方法同时利用了样本的局部几何结构和数据的判别信息，因而具有很好的判别性能。

上面所提出的大多数算法或者是监督学习算法，或者是无监督学习算法，对于分类问题，通常有监督算法能够取得较好的效果。然而实际问题中通常只能取得少量的有标签的数据和大量未知标签的数据，如果对大量数据进行标记，可能要花费大量的人力和物力。一种可行的办法是同时利用少量的有标签数据和大量

的无标签数据，即半监督学习[52-55]。半监督学习是基于流形假设和聚类假设的：①流形假设，数据分布在高维欧氏空间中的低维流形上；②聚类假设，即距离相近的数据很可能具有同样的标签。基本思想是利用无标签的数据提供数据集的几何信息，同时利用有标签的数据指导分类。

模式是模式识别的基本操作对象，传统子空间学习方法总是将模式转换成对应的向量来处理[56]，但向量化表示并非总是有效的。前面所介绍的 PCA、LDA、LPP 等经典的子空间学习算法都是向量表示方法，将高维向量空间数据降维到低维向量空间。对于人脸识别和其他图像识别问题，数据的内在表示模式是矩阵或张量，模式向量化会产生如下问题：一方面，向量化模式使得模式的结构遭到破坏[57,58]；另一方面，模式在向量化后将成为高维数据，导致高的计算复杂度和存储代价，产生维数灾难[59]和小样本问题。作为向量模式表示的扩展和补充，研究人员提出了各种基于矩阵或张量模式的学习算法，并在模式识别、数据挖掘、机器学习、计算机视觉等领域引起了广泛关注[60]。向量表示线性降维算法大多数已被推广到矩阵或张量模式，典型的有二维主成分分析(Two-Dimensional Principal Component Analysis, 2DPCA)[61]和二维线性判别分析(Two-Dimensional Linear Discriminant Analysis，2DLDA)[62]。以后又相继提出基于 PCA 的多线性主成分分析(Multilinear Principal Component Analysis，MPCA)[63-65]、LDA 的高阶张量推广算法(Discriminant Analysis with Tensor Representation，DATER)[66,67]和广义张量判别分析(General Tensor Discriminant Analysis，GTDA)[68]。最近，Yan 等[49]提出了降维技术的统一框架，将降维算法归结为图构造及其嵌入方式。

子空间学习算法通常根据一定的性能指标来寻找线性或非线性空间变换，把原始数据压缩到一个低维空间。但是这些方法在矩阵进行分解的时候没有对分解的对象和分解结果进行非负限制，分解的结果可能存在负值，但在许多情况下数据取负值没有实际意义，如图像数据不可能有负值的像素，文档统计中，负值也是无法解释的。所以，NMF 具有很重要的现实意义。1999 年 *Nature* 刊登了两位科学家 Lee 和 Seung[35]对数学中 NMF 的研究成果之后，NMF 在各领域得到广泛的使用，奠定了 NMF 的研究基础。

第 2 章　基于局部和全局的子空间降维算法

降维是克服维数灾难和小样本问题的重要途径，子空间降维算法按不同标准分为线性与非线性、监督与无监督、局部与全局降维算法。核方法和半监督算法分别建立起线性与非线性、有监督与无监督降维算法之间的纽带。全局子空间降维方法要求降维时将流形上的近邻点映射到低维空间中的近邻点，同时保证将流形上距离远的点映射到低维空间中远距离的点。局部子空间方法只是将流形上的近邻点映射到低维空间中的近邻点。全局算法简单、高效，在某些情况下能够获得比局部算法更优的性能，但没有考虑数据的内蕴几何结构，而局部算法充分考虑数据的内蕴几何结构，但要获得好的性能往往需要更多的训练样本，算法的性能主要取决于近邻参数的选择。本章将详细介绍几种比较经典的子空间学习算法，并分析这些算法的各自优缺点，为后面的章节做铺垫。

为了简化起见，算法中的数据都需要中心标准化，因此，在描述算法之前，给出数据中心标准化方法。

样本集 $X=\{x_i,i=1,2,\cdots,N\}\in\mathbf{R}^D$，均值向量不为零，则中心化样本定义为 $\bar{X}=XHD^{-1}$，式中，$H=I-\frac{1}{N}1_N1_N^{\mathrm{T}}$，$D=\mathrm{diag}\{\|Hx_{(1)}\|,\cdots,\|Hx_{(N)}\|\}$，$1_N$ 是 $N\times1$ 全 1 矢量，因此

$$\bar{X}=XH=X\left(I-\frac{1}{N}1_N1_N^{\mathrm{T}}\right)=X-\frac{1}{N}\left(\sum_{i=1}^{N}x_i\right)1_N^{\mathrm{T}}=\left[x_1-\frac{1}{N}\left(\sum_{i=1}^{N}x_i\right),\cdots,x_N-\frac{1}{N}\left(\sum_{i=1}^{N}x_i\right)\right]$$

由此可见，XH 是 X 中心化后的结果，XHD^{-1} 是 X 中心标准化的结果。

2.1　基于全局的子空间算法

2.1.1　主成分分析及其核推广

1）PCA

PCA[25]又称为主分量分析、KL 变换或霍特林变换，是著名的数据降维方法之一。PCA 的目的是通过线性变换寻求数据集的低维表示 $Y=\{y_i,i=1,2,\cdots,N\}\in\mathbf{R}^d(d\ll D)$，即 $Y=W^{\mathrm{T}}X$。W 的求解有最小化重构误差和最大方差两种方式，这两种方式从不同的角度刻画 PCA，但最终得到的结果却一致。第一种方式的基本思想是在最小二乘的意义下，寻找一组最佳的正交基向量，使重构误差达到最小。

第二种方式的基本思想是找到高维数据集彼此正交且数据方差变化最大的几个方向。

两种描述方式是等价的，表示为

$$\begin{aligned}&\arg\max_{W} W^{\mathrm{T}}SW = \arg\min_{W} \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - WW^{\mathrm{T}}x_i\right\|^2 \\ &\text{s.t.}\quad W^{\mathrm{T}}W = I\end{aligned} \tag{2.1}$$

证明

$$\begin{aligned}\text{左边} &= \arg\min_{W} \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - WW^{\mathrm{T}}x_i\right\|^2 \\ &= \arg\min_{W} \frac{1}{N}\sum_{i=1}^{N}(x_i - WW^{\mathrm{T}}x_i)^{\mathrm{T}}(x_i - WW^{\mathrm{T}}x_i) \\ &= \arg\min_{W} \frac{1}{N}\sum_{i=1}^{N}(x_i^{\mathrm{T}}x_i - x_i^{\mathrm{T}}WW^{\mathrm{T}}x_i - x_i^{\mathrm{T}}WW^{\mathrm{T}}x_i + x_i^{\mathrm{T}}WW^{\mathrm{T}}WW^{\mathrm{T}}x_i)\end{aligned}$$

因为$W^{\mathrm{T}}W = I$，所以

$$\begin{aligned}&\arg\min_{W} \frac{1}{N}\sum_{i=1}^{N}(x_i^{\mathrm{T}}x_i - x_i^{\mathrm{T}}WW^{\mathrm{T}}x_i - x_i^{\mathrm{T}}WW^{\mathrm{T}}x_i + x_i^{\mathrm{T}}WW^{\mathrm{T}}WW^{\mathrm{T}}x_i) \\ &= \arg\min_{W} \frac{1}{N}\sum_{i=1}^{N}(x_i^{\mathrm{T}}x_i - x_i^{\mathrm{T}}WW^{\mathrm{T}}x_i) \\ &= \arg\min_{W} -\frac{1}{N}\sum_{i=1}^{N}(x_i^{\mathrm{T}}WW^{\mathrm{T}}x_i) \\ &= \arg\max_{W} \frac{1}{N}\sum_{i=1}^{N}(x_i^{\mathrm{T}}WW^{\mathrm{T}}x_i) \\ &= \arg\max_{W} \frac{1}{N}\sum_{i=1}^{N}(W^{\mathrm{T}}x_ix_i^{\mathrm{T}}W) \\ &= \arg\max_{W}(W^{\mathrm{T}}SW)\end{aligned}$$

式中，$S = \frac{1}{N}\sum_{i=1}^{N}x_ix_i^{\mathrm{T}}$。

由上面的推导过程得到，PCA 的两种描述方式是等价的，PCA 在低维空间中保持数据总体方差结构。

于是，待求解问题可表示为

$$\begin{aligned}&\arg\max_{W}(W^{\mathrm{T}}SW)\\&\text{s.t.}\quad W^{\mathrm{T}}W=I\end{aligned}\tag{2.2}$$

构造拉格朗日（Lagrange）辅助函数为

$$f(W)=W^{\mathrm{T}}SW+\lambda(W^{\mathrm{T}}W-I)\tag{2.3}$$

求 $\partial f/\partial W$，并令 $\partial f/\partial W=0$，得到

$$SW=W\Lambda\tag{2.4}$$

PCA 算法的优点是计算简单、解释性强，缺点[69]是当数据是非线性分布时，PCA 算法将失效；PCA 能够找到的方向，对于分类和识别问题未必是最有利的；当特征值变化比较平缓时，难以对主分量进行取舍；在某些问题中，对 PCA 所求得的主分量进行解释是困难的。

数据矩阵 $X=[x_1,x_2,\cdots,x_N]$，$x_i\in\mathbf{R}^D$，有 $D\gg N$，如果对 $D\times D$ 的矩阵进行特征值分解，那么计算量将是非常大的。由奇异值分解定理可知，$\bar{X}\bar{X}^{\mathrm{T}}$ 和 $\bar{X}^{\mathrm{T}}\bar{X}$ 的特征值是相同的。对于特征值 λ_i，对应的特征向量分别为 u_i 和 v_i，则有

$$u_i=\frac{1}{\sqrt{\lambda_i}}\bar{X}v_i$$

因此可以通过对 $N\times N$ 的矩阵 $\bar{X}^{\mathrm{T}}\bar{X}$ 的特征值分解来求解 $\bar{X}\bar{X}^{\mathrm{T}}$ 的特征向量。

2）KPCA

Schölkopf 等[36]将核函数的思想和 PCA 相结合，提出了 KPCA，算法核心部分描述如下。

给定一组样本数据 x_k，$k=1,2,\cdots,N$，$x_k\in\mathbf{R}^D$，且满足均值为 0，即 $\sum_{k=1}^{N}x_k=0$。

样本的协方差矩阵为

$$C=\frac{1}{N}\sum_{j=1}^{N}x_jx_j^{\mathrm{T}}\tag{2.5}$$

C 的特征值和特征向量为

$$\lambda v=Cv\tag{2.6}$$

将式(2.5)代入式(2.6)得

$$\lambda v=\frac{1}{N}\sum_{j=1}^{N}\langle x_j,v\rangle x_j\tag{2.7}$$

两边分别与 x_k 作内积得到

$$\lambda\langle x_k,v\rangle=\langle x_k,Cv\rangle,\quad i=1,2,\cdots,N\tag{2.8}$$

考虑引入一个非线性映射 ϕ，将样本 x_i 从输入空间 $\mathbf{R}^D$ 映射到更高维的空间 F 中，即

$$\phi:\mathbf{R}^D \mapsto F, x \mapsto \phi(x) \tag{2.9}$$

协方差矩阵为

$$C^{\phi} = \frac{1}{N}\sum_{j=1}^{N}\phi(x_j)\phi(x_j)^{\mathrm{T}} \tag{2.10}$$

对角化协方差矩阵等价于求解特征值问题为

$$\lambda v^{\phi} = C^{\phi} v^{\phi} \tag{2.11}$$

由于 v^{ϕ} 是 $\phi(x_i)$ 的线性组合，即 $v^{\phi} = \sum_{i=1}^{N}\alpha_i\phi(x_i)$，式(2.11)两边分别对 $\phi(x_k)$ 作内积得

$$\lambda\left\langle \phi(x_k), v^{\phi}\right\rangle = \left\langle \phi(x_k), C^{\phi}v^{\phi}\right\rangle, \quad k=1,2,\cdots,N \tag{2.12}$$

将式(2.10)代入式(2.12)得

$$\lambda\left\langle \phi(x_k), \sum_{i=1}^{N}\alpha_i\phi(x_i)\right\rangle = \left\langle \phi(x_k), \frac{1}{N}\sum_{j=1}^{N}\phi(x_j)\phi(x_j)^{\mathrm{T}}\sum_{i=1}^{N}\alpha_i\phi(x_i)\right\rangle, \quad k=1,2,\cdots,N$$

$$\Rightarrow \lambda\sum_{i=1}^{N}\alpha_i\left\langle \phi(x_k),\phi(x_i)\right\rangle = \frac{1}{N}\sum_{i=1}^{N}\alpha_i\left\langle \phi(x_k), \sum_{j=1}^{N}\phi(x_j)\phi(x_j)^{\mathrm{T}}\phi(x_i)\right\rangle$$

$$\Rightarrow \lambda\sum_{i=1}^{N}\alpha_i\left\langle \phi(x_k),\phi(x_i)\right\rangle = \frac{1}{N}\sum_{i=1}^{N}\alpha_i\left\langle \phi(x_k), \sum_{j=1}^{N}\phi(x_j)\left\langle \phi(x_j),\phi(x_i)\right\rangle\right\rangle$$

$$\Rightarrow \lambda\sum_{i=1}^{N}\alpha_i\left\langle \phi(x_k),\phi(x_i)\right\rangle = \frac{1}{N}\sum_{i=1}^{N}\alpha_i\left\langle \sum_{j=1}^{N}\phi(x_j),\phi(x_i)\right\rangle\left\langle \phi(x_k), \sum_{j=1}^{N}\phi(x_j)\right\rangle \tag{2.13}$$

定义 $N\times N$ 的对称核矩阵 K，元素 $K_{ij} = \left\langle \phi(x_i),\phi(x_j)\right\rangle$，式(2.13)可以转化为

$$N\lambda K\alpha = K^2\alpha \text{或} N\lambda\alpha = K\alpha \tag{2.14}$$

核变换不能确保 $\sum_{i=1}^{N}\phi(x_i)=0$，因此有必要对核矩阵 K 进行归一化，令

$$\bar{\phi}(x_i) = \phi(x_i) - \frac{1}{N}\sum_{i=1}^{N}\phi(x_i) \tag{2.15}$$

则训练样本 $x_1,\cdots,x_N$ 在高维特征空间中对应的映射 $\phi(x_1),\cdots,\phi(x_N)$ 的均值就转化为零了，此时核矩阵变为