新一代信息科学与技术丛书

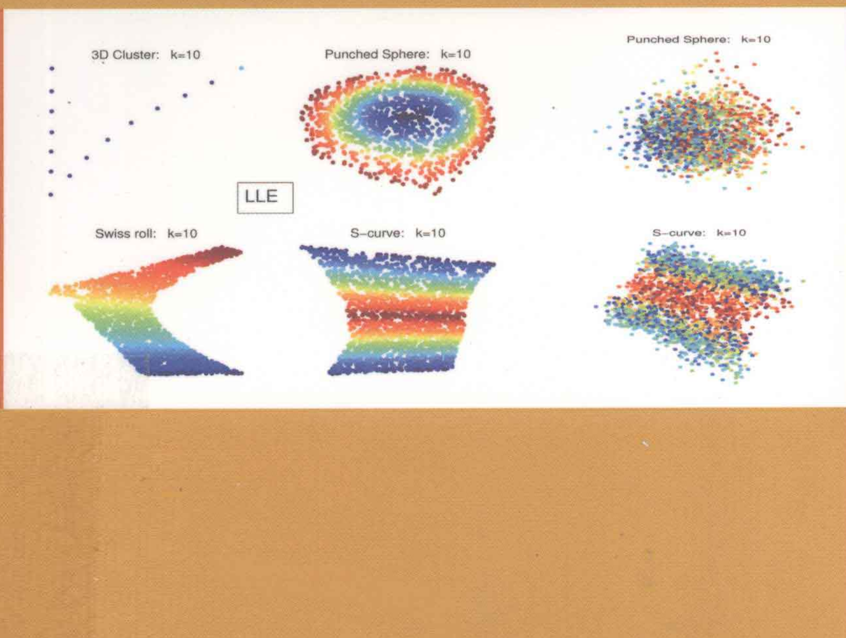
Geometric Structure of High-Dimensional Data and Dimensionality Reduction

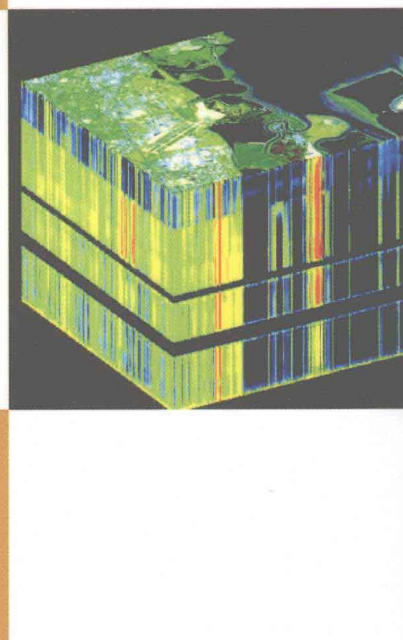
Jianzhong Wang

高维数据几何结构及降维

王建忠

（英文版）





高等教育出版社
HIGHER EDUCATION PRESS

Geometric Structure of High-Dimensional Data and Dimensionality Reduction

新一代信息科学与技术丛书

Geometric Structure of High-Dimensional Data and Dimensionality Reduction

高维数据几何结构及降维（英文版）

王建忠 Jianzhong Wang

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

Author
Prof. Jianzhong Wang
Department of Mathematics and Statistics
Sam Houston State University
1901 Avenue I, Huntsville
TX 77382-2206, USA
E-mail: jzwang@shsu.edu

图书在版编目(CIP)数据

高维数据几何结构及降维=Geometric Structure of High-Dimensional Data and Dimensionality Reduction：英文/王建忠编著．—北京：高等教育出版社，2012.1
ISBN 978-7-04-031704-6
Ⅰ．①高… Ⅱ．①王… Ⅲ．①统计数据-统计分析(数学)-英文 Ⅳ．①O212.1
中国版本图书馆 CIP 数据核字(2011)第184611号

策划编辑 刘 英　责任编辑 刘 英　封面设计 张 楠　版式设计 王 莹
责任校对 胡美萍　责任印制 朱学忠

出版发行	高等教育出版社	咨询电话	400-810-0598
社　址	北京市西城区德外大街4号	网　址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印　刷	涿州市星河印刷有限公司	网上订购	http://www.landraco.com
开　本	787mm×1092mm 1/16		http://www.landraco.com.cn
印　张	23.5	版　次	2012年1月第1版
字　数	540千字	印　次	2012年1月第1次印刷
购书热线	010-58581118	定　价	69.00元

物 料 号 31704-00

To my wife Huilan and my sons Chunhui and Chunze

Preface

Many objects in our world can be electronically represented with high-dimensional data — speech signals, images, videos, electrical text documents. We often need to analyze a large amount of data and process them. However, due to the high dimension of these data, directly processing them using regular systems may be too complicated and unstable to be feasible. In order to process high-dimensional data, dimensionality reduction technique becomes crucial. Dimensionality reduction is a method to represent high-dimensional data by their low-dimensional embeddings so that the low-dimensional data can be effectively used either in processing systems, or for better understanding. This technique has proved an important tool and has been widely used in many fields of data analysis, data mining, data visualization, and machine learning.

Principal component analysis (PCA) and classical multidimensional scaling (CMDS) are two traditional methods based on linear data models. However, they are not effective if data do not well reside on superplanes. Recently, many methods of nonlinear dimensionality reduction, also called manifold learning, have been developed based on nonlinear data models, which assume each observed high-dimensional data resides on a low-dimensional manifold. A nonlinear method *learns* the underlying manifold from the data and then represents the data by its low-dimensional manifold coordinates.

This monograph is an introductory treatise on dimensionality deduction with an emphasis on nonlinear methods. It includes a comprehensive introduction to linear methods, and then focuses on the geometric approach to nonlinear dimensionality reduction. The approach adopts manifold embedding technique that embeds a high-dimensional dataset into a low-dimensional manifold. The main idea of this approach can be outlined as follows. The geometry of the underlying manifold is learned from the neighborhood structure of the data, which defines the data similarity. Then a kernel is created to represent the learned manifold geometry and the spectral decomposition of the kernel leads to the manifold embedding. The different nonlinear methods adopt different manifold learning models. I hope that this book will illustrate this idea effectively.

Giving consideration to various readers, I try to keep a balance between intuitive description and strictly mathematical deduction in this book. Each method is first described with the intuitive idea. Then the formulation is derived with necessary details. Finally, the algorithm is presented and interpreted. Graphs are used to illustrate the capability and validity, as well as the advantages and shortcomings, of each method. Several applications, such as face recognition, document ordering and classification, objective visualization, are introduced or outlined. I have no ambition to completely cover these applications. I hope the introduction to them will help readers understand the concepts and contents of dimensionality reduction. The technical mathematical justification is put in the last section of each chapter. The readers who are only interested in the applications of dimensionality reduction can skip, for example, Chapter 2, Section 3.3, and all justification sections.

This monograph first presents a general introduction. It then presents the material in three parts. In Part I, the geometry of data is discussed. In this part, there is a brief review of manifold calculus, followed by a preliminary presentation of spectral graph theory. These are the main tools in the study of the geometry of data. The last chapter of Part I discusses the data models and strategies of dimensionality reduction. In Part II, linear methods are discussed. Besides PCA and CMDS, the random projection methods are also introduced. The main part of this monograph is Part III, in which several popular nonlinear methods are introduced, such as Isomaps, maximum variance unfolding (MVU), locally linear embedding (LLE), local tangent space alignment (LTSA), Laplacian eigenmaps (Leigs), Hessian locally linear embedding (HLLE), and diffusion maps (Dmaps). This part ends with a chapter on fast dimensionality reduction algorithms, which have attracted more interest recently, with the increasing availability of very large data sets.

The book is primarily intended for computer scientists, applied mathematician, statisticians, and data analysts. It is also accessible to other audiences who want to apply dimensionality reduction technique to understanding and treating the complex data in their own fields, such as economists and geophysics.

I am very grateful to my colleague and friend Dr. Barry Friedman who has read the manuscript. To the editorial office, particularly to Dr. Ying Liu, I wish to express my appreciation of their efficient assistance and friendly cooperation.

Jianzhong Wang
Huntsville, Texas
February, 2011

Acronyms

CMDS	classical multidimensional scaling
CSDP	C semidefinite programming
DR	dimensionality reduction
DWT	discrete wavelet transform
EM-PCA	expectation-maximization principle component analysis
FAT	fast anisotropic transformation
FFT	fast Fourier transform
GAT	greedy anisotropic transformation
HEEL	Hessian locally linear embedding
HSI	hyperspectral image
IRAT	interpolative randomizing anisotropic transformation
LLE	locally linear embedding
LMVU	landmark maximum variance unfolding
LTSA	local tangent space alignment
MDS	multidimensional scaling
MVU	maximum variance unfolding
PCA	principle component analysis
PRAT	projective randomizing anisotropic transformation
RAT	randomizing anisotropic transformation
RP	random projection
SDE	semidefinite embedding
SDP	semidefinite programming
SLLE	supervised locally linear embedding
SVD	singular value decomposition
Dmaps	diffusion maps
Isomaps	isometric maps
JL-embedding	Johnson and Lindenstrauss embedding
LB operator	Laplace Beltrami operator
Leigs	Laplacian eigenmaps
i.i.d.	independent and identically distributed
o.n.	orthonormal

o.g.	orthogonal
psd	positive semidefinite
pd	positive definite

Symbols

$\mathfrak{D}_{m,n}$	real $m \times n$ diagonal matrices
$\mathfrak{D}_n$	real $n \times n$ diagonal matrices
$\mathfrak{M}_{m,n}$	real $m \times n$ matrices
$\mathfrak{M}_{m,n}(r)$	real $m \times n$ matrices of rank r
$\mathfrak{M}_n$	real $n \times n$ matrices
$\mathfrak{M}_n(r)$	real $n \times n$ matrices of rank r
$\mathfrak{O}_{m,n}$	real $m \times n$ matrices with o.n. columns if $m \geqslant n$, and with o.n. rows if $m < n$
$\mathfrak{O}_n$	real $n \times n$ o.g. matrices
$\mathfrak{S}_n$	$n \times n$ psd matrices
$S\mathfrak{P}_n(r)$	$n \times n$ psd matrices of rank r
$\mathfrak{P}_n$	$n \times n$ pd matrices
$\mathfrak{R}_{m,n}$	$m \times n$ random matrices
$\mathfrak{R}^1_{m,n}$	$m \times n$ Gaussian random matrices, random matrices of Type 1
$\mathfrak{R}^2_{m,n}$	$m \times n$ random matrices of Type 2
$\mathfrak{R}^3_{m,n}$	$m \times n$ random matrices of Type 3
$\mathbb{R}^n$	space of real n-vectors, n-dimensional Euclidean space
$\mathbb{R}_+$	nonnegative real numbers
$\mathbb{Z}$	integer numbers
$\mathbb{Z}_+$	nonnegative integer numbers
$\mathbb{N}$	natural numbers

Contents

Part I Data Geometry

Part III Nonlinear Dimensionality Reduction