

Information Science and Statistics

Series Editors:

M. Jordan

J. Kleinberg

B. Schölkopf

Information Science and Statistics

Akaike and Kitagawa: The Practice of Time Series Analysis.

Bishop: Pattern Recognition and Machine Learning.

Cowell, Dawid, Lauritzen, and Spiegelhalter: Probabilistic Networks and Expert Systems.

Doucet, de Freitas, and Gordon: Sequential Monte Carlo Methods in Practice.

Fine: Feedforward Neural Network Methodology.

Hawkins and Olwell: Cumulative Sum Charts and Charting for Quality Improvement.

Jensen and Nielsen: Bayesian Networks and Decision Graphs, Second Edition.

Lee and Verleysen: Nonlinear Dimensionality Reduction.

Marchette: Computer Intrusion Detection and Network Monitoring: A Statistical Viewpoint.

Rissanen: Information and Complexity in Statistical Modeling.

Rubinstein and Kroese: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation, and Machine Learning.

Studený: Probabilistic Conditional Independence Structures.

Vapnik: The Nature of Statistical Learning Theory, Second Edition.

Wallace: Statistical and Inductive Inference by Minimum Message Length.

John A. Lee Michel Verleysen

Nonlinear Dimensionality Reduction

 Springer

John Lee
Molecular Imaging and Experimental
Radiotherapy
Université catholique de Louvain
Avenue Hippocrate 54/69
B-1200 Bruxelles
Belgium
john.lee@uclouvain.be

Michel Verleysen
Machine Learning Group – DICE
Université catholique de Louvain
Place du Levant 3
B-1348 Louvain-la-Neuve
Belgium
michel.verleysen@uclouvain.be

Series Editors:

Michael Jordan
Division of Computer
Science and
Department of Statistics
University of California,
Berkeley
Berkeley, CA 94720
USA

Jon Kleinberg
Department of Computer
Science
Cornell University
Ithaca, NY 14853
USA

Bernhard Schölkopf
Max Planck Institute for
Biological Cybernetics
Spemannstrasse 38
72076 Tübingen
Germany

Library of Congress Control Number: 2006939149

ISBN-13: 978-0-387-39350-6

e-ISBN-13: 978-0-387-39351-3

Printed on acid-free paper.

© 2007 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science + Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

To our families

Preface

Methods of dimensionality reduction are innovative and important tools in the fields of data analysis, data mining, and machine learning. They provide a way to understand and visualize the structure of complex data sets. Traditional methods like principal component analysis and classical metric multidimensional scaling suffer from being based on linear models. Until recently, very few methods were able to reduce the data dimensionality in a nonlinear way. However, since the late 1990s, many new methods have been developed and nonlinear dimensionality reduction, also called manifold learning, has become a hot topic. New advances that account for this rapid growth are, for example, the use of graphs to represent the manifold topology, and the use of new metrics like the geodesic distance. In addition, new optimization schemes, based on kernel techniques and spectral decomposition, have led to spectral embedding, which encompasses many of the recently developed methods.

This book describes existing and advanced methods to reduce the dimensionality of numerical databases. For each method, the description starts from intuitive ideas, develops the necessary mathematical details, and ends by outlining the algorithmic implementation. Methods are compared with each other with the help of different illustrative examples.

The purpose of the book is to summarize clear facts and ideas about well-known methods as well as recent developments in the topic of nonlinear dimensionality reduction. With this goal in mind, methods are all described from a unifying point of view, in order to highlight their respective strengths and shortcomings.

The book is primarily intended for statisticians, computer scientists, and data analysts. It is also accessible to other practitioners having a basic background in statistics and/or computational learning, such as psychologists (in psychometry) and economists.

Louvain-la-Neuve, Belgium
October 2006

John A. Lee
Michel Verleysen

Contents

Notations	XV
Acronyms	XVII
1 High-Dimensional Data	1
1.1 Practical motivations	1
1.1.1 Fields of application	2
1.1.2 The goals to be reached	3
1.2 Theoretical motivations	3
1.2.1 How can we visualize high-dimensional spaces?	4
1.2.2 Curse of dimensionality and empty space phenomenon .	6
1.3 Some directions to be explored	9
1.3.1 Relevance of the variables	10
1.3.2 Dependencies between the variables	10
1.4 About topology, spaces, and manifolds	11
1.5 Two benchmark manifolds	14
1.6 Overview of the next chapters	16
2 Characteristics of an Analysis Method	17
2.1 Purpose	17
2.2 Expected functionalities	18
2.2.1 Estimation of the number of latent variables	18
2.2.2 Embedding for dimensionality reduction	19
2.2.3 Embedding for latent variable separation	20
2.3 Internal characteristics	22
2.3.1 Underlying model	22
2.3.2 Algorithm	23
2.3.3 Criterion	23
2.4 Example: Principal component analysis	24
2.4.1 Data model of PCA	24
2.4.2 Criteria leading to PCA	26

2.4.3	Functionalities of PCA	29
2.4.4	Algorithms	31
2.4.5	Examples and limitations of PCA	33
2.5	Toward a categorization of DR methods	37
2.5.1	Hard vs. soft dimensionality reduction	38
2.5.2	Traditional vs. generative model	39
2.5.3	Linear vs. nonlinear model	40
2.5.4	Continuous vs. discrete model	40
2.5.5	Implicit vs. explicit mapping	41
2.5.6	Integrated vs. external estimation of the dimensionality	41
2.5.7	Layered vs. standalone embeddings	42
2.5.8	Single vs. multiple coordinate systems	42
2.5.9	Optional vs. mandatory vector quantization	43
2.5.10	Batch vs. online algorithm	43
2.5.11	Exact vs. approximate optimization	44
2.5.12	The type of criterion to be optimized	44
3	Estimation of the Intrinsic Dimension	47
3.1	Definition of the intrinsic dimension	47
3.2	Fractal dimensions	48
3.2.1	The q -dimension	49
3.2.2	Capacity dimension	51
3.2.3	Information dimension	52
3.2.4	Correlation dimension	53
3.2.5	Some inequalities	54
3.2.6	Practical estimation	55
3.3	Other dimension estimators	59
3.3.1	Local methods	59
3.3.2	Trial and error	60
3.4	Comparisons	62
3.4.1	Data Sets	63
3.4.2	PCA estimator	63
3.4.3	Correlation dimension	63
3.4.4	Local PCA estimator	65
3.4.5	Trial and error	66
3.4.6	Concluding remarks	67
4	Distance Preservation	69
4.1	State-of-the-art	69
4.2	Spatial distances	70
4.2.1	Metric space, distances, norms and scalar product	70
4.2.2	Multidimensional scaling	73
4.2.3	Sammon's nonlinear mapping	82
4.2.4	Curvilinear component analysis	88
4.3	Graph distances	97
4.3.1	Geodesic distance and graph distance	97

4.3.2	Isomap	102
4.3.3	Geodesic NLM	111
4.3.4	Curvilinear distance analysis	114
4.4	Other distances	119
4.4.1	Kernel PCA	120
4.4.2	Semidefinite embedding	125
5	Topology Preservation	133
5.1	State of the art	133
5.2	Predefined lattice	135
5.2.1	Self-Organizing Maps	135
5.2.2	Generative Topographic Mapping	143
5.3	Data-driven lattice	152
5.3.1	Locally linear embedding	152
5.3.2	Laplacian eigenmaps	159
5.3.3	Isotop	165
6	Method comparisons	173
6.1	Toy examples	173
6.1.1	The Swiss roll	173
6.1.2	Manifolds having essential loops or spheres	193
6.2	Cortex unfolding	199
6.3	Image processing	203
6.3.1	Artificial faces	206
6.3.2	Real faces	214
7	Conclusions	225
7.1	Summary of the book	225
7.1.1	The problem	225
7.1.2	A basic solution	226
7.1.3	Dimensionality reduction	226
7.1.4	Latent variable separation	228
7.1.5	Intrinsic dimensionality estimation	229
7.2	Data flow	230
7.2.1	Variable Selection	230
7.2.2	Calibration	230
7.2.3	Linear dimensionality reduction	231
7.2.4	Nonlinear dimensionality reduction	231
7.2.5	Latent variable separation	232
7.2.6	Further processing	232
7.3	Model complexity	232
7.4	Taxonomy	233
7.4.1	Distance preservation	236
7.4.2	Topology preservation	237
7.5	Spectral methods	238
7.6	Nonspectral methods	241

7.7	Tentative methodology	242
7.8	Perspectives	245
A	Matrix Calculus	247
A.1	Singular value decomposition	247
A.2	Eigenvalue decomposition	248
A.3	Square root of a square matrix	248
B	Gaussian Variables	251
B.1	One-dimensional Gaussian distribution	251
B.2	Multidimensional Gaussian distribution	253
B.2.1	Uncorrelated Gaussian variables	254
B.2.2	Isotropic multivariate Gaussian distribution	254
B.2.3	Linearly mixed Gaussian variables	256
C	Optimization	259
C.1	Newton's method	259
C.1.1	Finding extrema	260
C.1.2	Multivariate version	260
C.2	Gradient ascent/descent	261
C.2.1	Stochastic gradient descent	261
D	Vector quantization	263
D.1	Classical techniques	265
D.2	Competitive learning	266
D.3	Taxonomy	266
D.4	Initialization and "dead units"	267
E	Graph Building	269
E.1	Without vector quantization	270
E.1.1	K -rule	270
E.1.2	ϵ -rule	271
E.1.3	τ -rule	271
E.2	With vector quantization	272
E.2.1	Data rule	272
E.2.2	Histogram rule	274
F	Implementation Issues	277
F.1	Dimension estimation	277
F.1.1	Capacity dimension	277
F.1.2	Correlation dimension	277
F.2	Computation of the closest point(s)	279
F.3	Graph distances	280
	References	283

Index 297

Notations

\mathbb{N}	The set of positive natural numbers: $\{0, 1, 2, 3, \dots\}$
\mathbb{R}	The set of real numbers
y, x	Known or unknown random variables taking their values in \mathbb{R}
\mathbf{A}	A matrix
$a_{i,j}$	An entry of the matrix \mathbf{A} (located at the crossing of the i th row and the j th column)
N	Number of points in the data set
M	Number of prototypes in the codebook \mathbf{C}
D	Dimensionality of the data space (which is usually \mathbb{R}^D)
P	Dimensionality of the latent space (which is usually \mathbb{R}^P) (or its estimation as the intrinsic dimension of the data)
\mathbf{I}_D	D -dimensional identity matrix
$\mathbf{I}_{P \times D}$	Rectangular matrix containing the first P rows of \mathbf{I}_D
$\mathbf{1}_N$	N -dimensional column vector containing ones everywhere
\mathbf{y}	Random vector in the known data space: $\mathbf{y} = [y_1, \dots, y_d, \dots, y_D]^T$
\mathbf{x}	Random vector in the unknown latent space: $\mathbf{x} = [x_1, \dots, x_p, \dots, x_P]^T$
$\mathbf{y}(i)$	The i th vector of the data set
$\mathbf{x}(i)$	(Unknown) latent vector that generated $\mathbf{y}(i)$
$\hat{\mathbf{x}}(i)$	The estimate of $\mathbf{x}(i)$
\mathcal{Y}	The data set $\mathcal{Y} = \{\dots, \mathbf{y}(i), \dots\}_{1 \leq i \leq N}$
\mathcal{X}	The (unknown) set of latent vectors that generated \mathcal{Y}
$\hat{\mathcal{X}}$	Estimation of \mathcal{X}
\mathbf{Y}	The data set in matrix notation: $\mathcal{Y} = [\dots, \mathbf{y}(i), \dots]_{1 \leq i \leq N}$
\mathbf{X}	The (unknown) ordered set of latent vectors that generated \mathbf{Y}
$\hat{\mathbf{X}}$	Estimation of \mathbf{X}

\mathcal{M}	A manifold (noted as a set)
\mathbf{m}	The functional notation of \mathcal{M} : $\mathbf{y} = \mathbf{m}(\mathbf{x})$
$E_x\{x\}$	The expectation of the random variable x
$\mu_x(x)$	The mean value of the random variable x (computed with its known values $x(i)$, $i = 1, \dots, N$)
μ_i	The i th-order centered moment
μ'_i	The i th-order raw moment
$\mathbf{C}_{\mathbf{xy}}$	The covariance matrix between the random vectors \mathbf{x} and \mathbf{y}
$\hat{\mathbf{C}}_{\mathbf{xy}}$	The estimate of the covariance matrix
$f(\mathbf{x}), \mathbf{f}(\mathbf{x})$	Uni- or multivariate function of the random vector \mathbf{x}
$\frac{\partial f(\mathbf{x})}{\partial x_p}$	Partial derivative of f with respect to x_p
$\nabla_{\mathbf{x}}f(\mathbf{x})$	Gradient vector of f with respect to \mathbf{x}
$\mathbf{H}_{\mathbf{x}}f(\mathbf{x})$	Hessian matrix of f with respect to \mathbf{x}
$\mathbf{J}_{\mathbf{x}}\mathbf{f}(\mathbf{x})$	Jacobian matrix of \mathbf{f} with respect to \mathbf{x}
$\langle \mathbf{y}(i) \cdot \mathbf{y}(j) \rangle$	Scalar product between the two vectors $\mathbf{y}(i)$ and $\mathbf{y}(j)$
$d(\mathbf{y}(i), \mathbf{y}(j))$	Distance function between the two vectors $\mathbf{y}(i)$ and $\mathbf{y}(j)$ (often a spatial distance, like the Euclidean one) shortened as $d_{\mathbf{y}}(i, j)$ or $d_{\mathbf{y}}$ when the context is clear
$\delta(\mathbf{y}(i), \mathbf{y}(j))$	Geodesic or graph distance between $\mathbf{y}(i)$ and $\mathbf{y}(j)$
\mathcal{C}, \mathcal{G}	Codebook (noted as a set) in the data and latent spaces
\mathbf{C}, \mathbf{G}	Codebook (noted as a matrix) in the data and latent spaces
$\mathbf{c}(r), \mathbf{g}(r)$	Coordinates of the r th prototypes in the codebook (respectively, in the data and latent spaces)

Acronyms

DR Dimensionality reduction
LDR Linear dimensionality reduction
NLDR Nonlinear dimensionality reduction

ANN Artificial neural networks
EVD Eigenvalue decomposition
SVD Singular value decomposition
SVM Support vector machines
VQ Vector quantization

CCA	Curvilinear component analysis	<i>NLDR method</i>
CDA	Curvilinear distance analysis	<i>NLDR method</i>
EM	Expectation-maximization	<i>optimization technique</i>
GTM	Generative topographic mapping	<i>NLDR method</i>
HLLE	Hessian LLE (see LLE)	<i>NLDR method</i>
KPCA	Kernel PCA (see PCA)	<i>NLDR method</i>
LE	Laplacian eigenmaps	<i>NLDR method</i>
LLE	Locally linear embedding	<i>NLDR method</i>
MDS	Multidimensional scaling	<i>LDR/NLDR method</i>
MLP	Multilayer perceptron	<i>ANN for function approx.</i>
MVU	Maximum variance unfolding (see SDE)	<i>NLDR method</i>
NLM	(Sammon's) nonlinear mapping	<i>NLDR method</i>
PCA	Principal component analysis	<i>LDR method</i>
RBFN	Radial basis function network	<i>ANN for function approx.</i>
SDE	Semidefinite embedding	<i>NLDR method</i>
SDP	Semidefinite programming	<i>optimization technique</i>
SNE	Stochastic neighbor embedding	<i>NLDR method</i>
SOM	(Kohonen's) self-organizing map	<i>NLDR method</i>
TRN	Topology-representing network	<i>ANN</i>