

SPRINGER BRIEFS IN STATISTICS
JSS RESEARCH SERIES IN STATISTICS

Yuichi Mori
Masahiro Kuroda
Naomichi Makino

Nonlinear Principal Component Analysis and Its Applications



Springer

SpringerBriefs in Statistics

JSS Research Series in Statistics

Editor-in-Chief

Naoto Kunitomo
Akimichi Takemura

Series editors

Genshiro Kitagawa
Tomoyuki Higuchi
Yutaka Kano
Toshimitsu Hamasaki
Shigeyuki Matsui
Manabu Iwasaki
Yasuhiro Omori
Yoshihiro Yajima

The current research of statistics in Japan has expanded in several directions in line with recent trends in academic activities in the area of statistics and statistical sciences over the globe. The core of these research activities in statistics in Japan has been the Japan Statistical Society (JSS). This society, the oldest and largest academic organization for statistics in Japan, was founded in 1931 by a handful of pioneer statisticians and economists and now has a history of about 80 years. Many distinguished scholars have been members, including the influential statistician Hirotugu Akaike, who was a past president of JSS, and the notable mathematician Kiyosi Itô, who was an earlier member of the Institute of Statistical Mathematics (ISM), which has been a closely related organization since the establishment of ISM. The society has two academic journals: the Journal of the Japan Statistical Society (English Series) and the Journal of the Japan Statistical Society (Japanese Series). The membership of JSS consists of researchers, teachers, and professional statisticians in many different fields including mathematics, statistics, engineering, medical sciences, government statistics, economics, business, psychology, education, and many other natural, biological, and social sciences.

The JSS Series of Statistics aims to publish recent results of current research activities in the areas of statistics and statistical sciences in Japan that otherwise would not be available in English; they are complementary to the two JSS academic journals, both English and Japanese. Because the scope of a research paper in academic journals inevitably has become narrowly focused and condensed in recent years, this series is intended to fill the gap between academic research activities and the form of a single academic paper.

The series will be of great interest to a wide audience of researchers, teachers, professional statisticians, and graduate students in many countries who are interested in statistics and statistical sciences, in statistical theory, and in various areas of statistical applications.

More information about this series at <http://www.springer.com/series/13497>

Yuichi Mori · Masahiro Kuroda
Naomichi Makino

Nonlinear Principal Component Analysis and Its Applications

Yuichi Mori
Okayama University of Science
Okayama
Japan

Naomichi Makino
Osaka University
Osaka
Japan

Masahiro Kuroda
Okayama University of Science
Okayama
Japan

ISSN 2191-544X
SpringerBriefs in Statistics

ISSN 2364-0057
JSS Research Series in Statistics
ISBN 978-981-10-0157-4
DOI 10.1007/978-981-10-0159-8

ISSN 2191-5458 (electronic)

ISSN 2364-0065 (electronic)

ISBN 978-981-10-0159-8 (eBook)

Library of Congress Control Number: 2016957863

© The Author(s) 2016

This book was advertised with a copyright holder in the name of the publisher in error, whereas the author holds the copyright.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer Nature Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #22-06/08 Gateway East, Singapore 189721, Singapore

Preface

As the data size increases and the data structure becomes increasingly complex, considering the nonlinearity, mixed measurement level data, simultaneous analysis and computational efficiency, for example, become increasingly necessary.

Accordingly, we have been investigating data analysis and developing methods related to the above problems: how to jointly analyze data that include not only categorical variables but also numerical variables; matrix operations approaching categorical data analyses; how to select a reasonable subset of variables from mixed measurement level data; analyses based on the sparseness of data and simultaneous estimations, such as a joint method with dimension reduction and clustering, which have recently been investigated; and how to accelerate the computations involving the alternating least squares (ALS) algorithm.

Although we have heretofore been investigating these topics separately, we decided to include the above-mentioned methods/techniques in this series when we started to write this Springer Brief Statistics series.

We present our research interests in two parts as follows:

The first part consists of two chapters that introduce the principles of nonlinear principal component analysis (PCA). After a brief introduction of the ordinary PCA, a PCA for categorical (nominal and ordinal) data is introduced as a nonlinear PCA, in which the optimal scaling technique is used to quantify the categorical variables. The ALS algorithm is the main algorithm used in this method. Next, multiple correspondence analysis (MCA), which is a special case of nonlinear PCA, is introduced, and ALS is also used in the computation. All formulations in these methods are integrated in the same manner as the matrix operations. Since any measurement level data can be treated consistently as numerical data and ALS is a very powerful tool for estimations, these methods can be used in a variety of fields, including biometrics, econometrics, psychometrics, and sociology.

The second part of the book consists of four chapters, which describe applications of the nonlinear PCA: variable selection for mixed measurement level data, sparse MCA, reduced k-means clustering, and an acceleration of the ALS algorithm. The variable selection methods used in PCA, which were originally developed for numerical data, can be applied to any type of measurement level data

using nonlinear PCA. Sparseness and k-means clustering for nonlinear data, which were proposed in recent studies, are extensions obtained using the same matrix operations used in nonlinear PCA and MCA. Finally, an acceleration algorithm is proposed to reduce the computational cost of ALS iteration in nonlinear multivariate methods.

This book demonstrates the usefulness of nonlinear PCA and MCA, which can be applied to different measurement level data in a variety of fields and covers the latest topics, including the extension of traditional statistical methods, newly proposed nonlinear methods, and the computational efficiency of these methods. The ALS algorithm is a key concept in all chapters, and optimal quantifications and matrix operations are also key concepts in this book.

We sincerely hope that the concepts and applications presented herein will be of use.

Okayama, Japan
June 2016

Yuichi Mori
Masahiro Kuroda
Naomichi Makino

*The original version of the book was revised:
Typographical errors have been corrected.
The erratum to the book is available at
DOI [10.1007/978-981-10-0159-8_8](https://doi.org/10.1007/978-981-10-0159-8_8).*

Contents

1	Introduction	1
Part I Nonlinear Principal Component Analysis		
2	Nonlinear Principal Component Analysis	7
2.1	Principal Component Analysis	7
2.1.1	Eigen-Decomposition of $\mathbf{Y}^\top \mathbf{Y}/n$	8
2.1.2	Singular Value Decomposition of $\mathbf{Y}$	8
2.2	Quantification of Qualitative Data	9
2.3	Nonlinear PCA	10
2.3.1	Low-Rank Matrix Approximation	11
2.3.2	Homogeneity Analysis	11
2.3.3	Alternating Least Squares Algorithm for Nonlinear PCA	13
2.4	Example: Sleeping Bags	19
	References	20
3	Multiple Correspondence Analysis	21
3.1	Introduction	21
3.2	Formulation	22
3.3	Rank-Restricted MCA	23
3.4	Algorithm of Rank-Restricted MCA	24
3.5	Numerical Example	25
	References	28
Part II Applications and Related Topics		
4	Variable Selection in Nonlinear Principal Component Analysis	31
4.1	Introduction	31