Virginia Gray
Editor

MATHEMATICS
RESEARCH
DEVELOPMENTS

Principal Component Analysis

Methods, Applications and Technology

Novinka

Principal Component Analysis

Methods, Applications and Technology

Contributors

Hiroyuki Yamamoto
Tzu-Yi Pai
Yi-Ti Tung
Ming-Ray Lin
Su-Hwa Lin
Lung-Yi Chan
Chia-Fu Lin
Robert Spurr
Vijay Natraj
Pushkar Kopparla
Matt Christi

Virginia Gray
Editor

nova publishers
www.novapublishers.com

ISBN 978-1-53610-889-7
9 781536 108897

Principal Component Analysis: Methods, Applications and Technology ■ Gray

Novinka

MATHEMATICS RESEARCH DEVELOPMENTS

PRINCIPAL COMPONENT ANALYSIS

METHODS, APPLICATIONS AND TECHNOLOGY

VIRGINIA GRAY
EDITOR

NOVINKA
New York

Copyright © 2017 by Nova Science Publishers, Inc.

All rights reserved. No part of this book may be reproduced, stored in a retrieval system or transmitted in any form or by any means: electronic, electrostatic, magnetic, tape, mechanical photocopying, recording or otherwise without the written permission of the Publisher.

We have partnered with Copyright Clearance Center to make it easy for you to obtain permissions to reuse content from this publication. Simply navigate to this publication's page on Nova's website and locate the "Get Permission" button below the title description. This button is linked directly to the title's permission page on copyright.com. Alternatively, you can visit copyright.com and search by title, ISBN, or ISSN.

For further questions about using the service on copyright.com, please contact:
Copyright Clearance Center
Phone: +1-(978) 750-8400 Fax: +1-(978) 750-4470 E-mail: info@copyright.com.

NOTICE TO THE READER

The Publisher has taken reasonable care in the preparation of this book, but makes no expressed or implied warranty of any kind and assumes no responsibility for any errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of information contained in this book. The Publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or in part, from the readers' use of, or reliance upon, this material. Any parts of this book based on government reports are so indicated and copyright is claimed for those parts to the extent applicable to compilations of such works.

Independent verification should be sought for any data, advice or recommendations contained in this book. In addition, no responsibility is assumed by the publisher for any injury and/or damage to persons or property arising from any methods, products, instructions, ideas or otherwise contained in this publication.

This publication is designed to provide accurate and authoritative information with regard to the subject matter covered herein. It is sold with the clear understanding that the Publisher is not engaged in rendering legal or any other professional services. If legal or any other expert assistance is required, the services of a competent person should be sought. FROM A DECLARATION OF PARTICIPANTS JOINTLY ADOPTED BY A COMMITTEE OF THE AMERICAN BAR ASSOCIATION AND A COMMITTEE OF PUBLISHERS.

Additional color graphics may be available in the e-book version of this book.

Library of Congress Cataloging-in-Publication Data

ISBN: 978-1-53610-889-7

Published by Nova Science Publishers, Inc. † New York

此为试读，需要完整PDF请访问：www.ertongbook.com

MATHEMATICS RESEARCH DEVELOPMENTS

PRINCIPAL COMPONENT ANALYSIS

METHODS, APPLICATIONS AND TECHNOLOGY

MATHEMATICS RESEARCH DEVELOPMENTS

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the e-books tab.

PREFACE

This book provides new research on principal component analysis (PCA). Chapter One introduces typical PCA applications of transcriptomic, proteomic and metabolomic data. Chapter Two studies the factor analysis of an outcome measurement survey for science, technology and society. Chapter Three examines the application of PCA to performance enhancement of hyperspectral radiative transfer computations.

Chapter 1 – With recent advances in high-throughput technologies, large amounts of transcriptomic, proteomic, metabolomic and other omics data have become available. Biological knowledge can be extracted from these comprehensive datasets by data-driven approaches such as multivariate analysis. Omics data are high-dimensional; that is, the number of variables is much higher than the number of samples. Principal component analysis (PCA) has been widely applied to visualize omics data in low-dimensional subspaces. In this chapter, the authors review the applications of PCA in different types of omics data. Another important task in PCA is to select variables (e.g., genes, proteins, metabolites) associated with phenotypes such as groups or time series information. After choosing relevant variables by statistical approaches, biologists make biological inferences from the statistically selected variables. In the literature, factor loading is known by a variety of other terms (PC loading, PC coefficient, weights, and eigenvectors), which are frequently interchanged and confused in the literature. Here, the authors theoretically explain the relationships among these technical terms in the presence and absence of scaling. Finally, they introduce the way to select variables by using statistical hypothesis testing of the factor loading in PCA. The authors urge that future PCA of omics data be applied in an appropriate manner.

Chapter 2 – In this study, a factor analysis of outcome measurement survey for "Science, Technology and Society" (STS) course of general education was studied based on principal component analysis (PCA). STS course is mainly designed to familiarize students with knowledge, attitude, and action associated with science, technology and society including nature evolution, information and network, mathematics, nano-technology, global climate change, and sustainable development etc. STS course also intends to develop students' civil responsibility, information and science accomplishments, and ability for logical thinking and identification. The results showed that the questionnaire for evaluating the outcome of the STS course was edited. The questionnaire was divided into 15 respects. Besideds, the construct was divided into three classifications including knowledge, attitude, and action. The construct validity of the outcome measurement survey was calculated using PCA. There were totally 33 PCs extracted from PCA. According to PCA, the previous 19 and 27 PCs could account for about 70% and 80% of total variance, respectively. Total 33 PCs could account for 86.313% of total variance. The first PC consisted of 37 components, mainly the components of action, i.e., the action dimension explained the most variance. The first PC which was mainly made up of the action dimension accounted for 14.110% of variance alone. The second PCC consisted of 32 components, mainly the components of attitude. The attitude dimension explained as much of the remaining variation. The second PC which was mainly made up of the attitude dimension accounted for 10.250% of variance alone. The third PC consisted of 11 components, mainly the components of attitude. The remaining attitude dimension explained as much of the remaining variation. The third PC which was mainly made up of the attitude dimension accounted for 5.089% of variance alone. The 5th, 6th, 7th, 8th, 9th, 10th, 14th, 15th, 16th, 17th, and 18th PC mainly consisted of the components of knowledge. The knowledge dimension explained as much of the remaining variation.

Chapter 3 – For retrievals and sensitivity studies with passive Earth- (and planetary-) atmosphere remote-sensing instruments having moderate or high spectral resolution, the major time-limiting step is the forward-model multiple-scatter (MS) radiative transfer (RT) computation. In this paper, the authors use a principal component analysis (PCA)-based spectral binning technique to enhance RT performance. For a given spectral window, one may apply PCA to a data set of optical profiles (of trace gas absorption and single-scatter albedo, for example) to derive a much smaller set of optical profiles that captures most of the data variability. Further, 2-stream (2S) RT calculations, although not accurate, are highly correlated with full MS RT computations. The key to

enhanced performance is: (1) to separate the (fast) single-scatter (SS) and (slow) MS computations, and (2) to limit the number of full MS RT calculations to the PCA-reduced data set, and thus develop a series of PC-based correction factors that are applied to the fast hyperspectral calculations made by the 2S and SS RT models. This technique delivers accuracy of 0.1% or better, depending on the number of EOFs selected for the PCA; 2–4 EOFs are usually sufficient. Speed-ups of 1–2 orders of magnitude are possible for a wide range of spectral windows.

For global atmospheric chemical and climate modeling studies, performance considerations currently require that RT calculations must be done using fast and analytic 2S RT models; in this regard, 2S fluxes are frequently in error by 10–15% compared with high-accuracy MS RT results, and this uncertainty is a significant source of error in radiative forcing calculations. The authors have developed a set of hyperspectral PCA-based RT tools that can perform fast radiance and/or flux calculations across the entire solar and thermal regions from 260 nm to 20 microns (including the cross-over region where both solar and thermal sources apply) with an accuracy that is comparable (at the 0.1% level) to that obtained with full MS RT.

The authors show that the fast-PCA method works for RT with polarization, e.g., for the VLIDORT model — multiple-scatter correction factors for the intensity can be derived with the (scalar) 2S model, and the technique can also be applied to fast calculation of Stokes parameters Q and U. The fast-PCA method is also applicable to calculations of analytically-derived atmospheric profile and total column Jacobians, as well as surface weighting functions; they present an example for ozone profile Jacobians in the UV range from 306–335 nm.

The fast-PCA method has been used operationally (in a spectral fitting algorithm requiring total column ozone weighting functions in the forward calculations) to reprocess the long-term global total ozone record from several atmospheric chemistry satellite platforms past and present (GOME, GOME-2, SCIAMACHY). In addition to this application, the authors present two examples of fast-PCA RT performance enhancement for the O_2 A band and the strong and weak CO_2 absorption bands that are central to the Orbiting Carbon Observatory (OCO-2) mission, which is being used to generate global maps of CO_2 concentrations from space.

此为试读，需要完整PDF请访问：www.ertongbook.com

CONTENTS

Preface		**vii**
Chapter 1	The Principal Component Analysis of Omics Data *Hiroyuki Yamamoto*	**1**
Chapter 2	A Factor Analysis of an Outcome Measurement Survey for Science, Technology and Society: A Course of General Education Based on PCA *Tzu-Yi Pai, Yi-Ti Tung, Ming-Ray Lin, Su-Hwa Lin, Lung-Yi Chan and Chia-Fu Lin*	**15**
Chapter 3	The Application of Principal Component Analysis (PCA) to Performance Enhancement of Hyperspectral Radiative Transfer Computations *Robert Spurr, Vijay Natraj, Pushkar Kopparla and Matt Christi*	**33**
Bibliography		**83**
Index		**137**

In: Principal Component Analysis
Editor: Virginia Gray
ISBN: 978-1-53610-889-7
© 2017 Nova Science Publishers, Inc.

Chapter 1

THE PRINCIPAL COMPONENT ANALYSIS OF OMICS DATA

Hiroyuki Yamamoto*
Human Metabolome Technologies, Inc., Mizukami, Kakuganji, Tsuruoka, Yamagata, Japan

ABSTRACT

With recent advances in high-throughput technologies, large amounts of transcriptomic, proteomic, metabolomic and other omics data have become available. Biological knowledge can be extracted from these comprehensive datasets by data-driven approaches such as multivariate analysis. Omics data are high-dimensional; that is, the number of variables is much higher than the number of samples. Principal component analysis (PCA) has been widely applied to visualize omics data in low-dimensional subspaces. In this chapter, we review the applications of PCA in different types of omics data. Another important task in PCA is to select variables (e.g., genes, proteins, metabolites) associated with phenotypes such as groups or time series information. After choosing relevant variables by statistical approaches, biologists make biological inferences from the statistically selected variables. In the literature, factor loading is known by a variety of other terms (PC loading, PC coefficient, weights, and eigenvectors), which are frequently interchanged and confused in the literature. Here, we theoretically explain the relationships among these technical terms in the presence and absence

* Corresponding Author E-mail: h.yama2396@gmail.com.

of scaling. Finally, we introduce the way to select variables by using statistical hypothesis testing of the factor loading in PCA. We urge that future PCA of omics data be applied in an appropriate manner.

Keywords: transcriptomics, proteomics, metabolomics, factor loading, statistical hypothesis testing

1. INTRODUCTION

Recently, gene expressions and protein and metabolite concentrations have been comprehensively measured by high-throughput technologies. Analyses of these large datasets, called *omics analyses*, include transcriptomics [1], proteomics [2] and metabolomics [3]. Transcriptomics measures the expression levels of messenger ribonucleic acid (mRNA) by deoxyribonucleic acid (DNA) microarray [1] or RNA-sequencing [4]. Proteomics assays the protein expression levels or concentrations by combining two-dimensional gel electrophoresis [5] or liquid chromatography [2] with mass spectrometry. Proteins can also be assayed by antibody microarray [6]. Metabolomics typically employs mass spectrometry and hyphenated technologies such as gas chromatography [7], liquid chromatography [8] and capillary electrophoresis mass spectrometry [9]. Nuclear magnetic resonance [10] is also applied in metabolomics.

Each of the omics fields employs different analytical methods, but all omics data are formatted as matrices of samples and variables of genes, proteins or metabolites. Omics data are characterized by their high dimensionality; that is, the number of variables far exceeds the number of samples, and require converting to a simpler format for easy visualization. This is most popularly achieved by principal component analysis (PCA) [11].

This chapter introduces typical PCA applications of transcriptomic, proteomic and metabolomic data. We also clarify what is meant by *factor loading* in PCA, which is often confused in the literature, from a theoretical viewpoint in the presence and absence of scaling. Finally, we introduce variable selection by statistical hypothesis testing of the factor loading in PCA.

2. Typical PCA Applications in Omics Data Analyses

2.1. PCA for Transcriptomics

This subsection discusses PCA applications in transcriptomics. Basse et al. [12] acquired the gene expression data of brown and white adipose tissues in sheep. These tissue types perform opposite biological functions; white adipose tissue accumulates energy, whereas brown adipose tissue consumes energy by metabolizing fatty acids and glucose. Transitions from brown to white adipose tissue were clustered in the first PC of their PCA. Basse et al. then selected genes that differed among the groups by one-way analysis of variance (ANOVA) and Student's t-test.

From the gene expression data of kidney biopsies, Parikh et al. [13] differentiated early-treatment responders to lupus nephritis flare from non-responders and healthy controls. In PC1, healthy controls were separately clustered from responders and non-responders. Based on the PC1 factor loading, the authors identified the genes that differed between healthy controls and the responder and non-responder groups.

Ryan et al. [14] compared the gene expressions in the whole blood of healthy controls and patients with chronic inflammatory response syndrome (CIRS) following exposure to the marine toxin ciguatoxin. Ciguatoxin commonly enters the diet through toxic fish consumption, and leads to CIRS in rare cases. They performed a PCA on the genes selected by a modified t-test and fold-change, focusing on the loading factors of the genes. The control and patient samples were clearly separated on a PC score plot.

To better understand the transcriptional profiles putatively associated with ovarian egg infertility in the marine bivalve mollusc *Venerupis decussata*, Pauletto et al. [15] acquired the gene expression data of oocytes of this organism. They performed a comparative study between released oocytes and ovarian oocytes obtained by stripping, and found that low hatching rate, medium hatching rate and stripped eggs were clearly separated along PC1. They selected genes by their eigenvectors in PC1 and by significance analysis of microarray (SAM) [16].

2.2. PCA for Proteomics

This subsection introduces some representative applications of PCA in proteomics. Meling et al. [17] acquired the serum proteome data of scrapie-infected sheep. Scrapie is caused by an abnormality in a normal protein, and is considered as a type of prion disease. Meling et al. performed a comparative study of control and infected samples using PCA, but the two sample groups were not clearly separated in the score plot. Next, they selected pertinent variables by partial least squares-discriminant analysis. A second PCA confirmed a clear separation along PC1.

Waloszczyk et al. [18] differentiated between normal lung tissue and early-stage lung cancer by proteomic lung tissue data. In the PC score plot, lung cancer samples showed larger variance than the control samples, probably because the dataset included various types of lung cancer. They distinguished proteins in different groups by t-test and the Wilcoxon rank sum test.

Bergquist et al. [19] acquired the proteome data of bronchoalveolar lavage mouse samples, and differentiated between two major inflammatory phenotypes of asthma (aosinophilic and neutrophilic) in a PCA. The control and drug-treated groups formed separate clusters, and the aosinophilic and neutrophilic asthmatic subgroups were separately clustered along PC1. They also selected proteins by their loading in PC1 and PC2 and an ANOVA.

2.3. PCA for Metabolomics

We now introduce some representative applications of PCA in metabolomics. Wang et al. [20] investigated the metabolomic responses of vascular smooth muscle cells to R- and S-carvedilol. Whether the two enantiomers of carvedilol (a therapeutic drug for hypertension and cardiac disease) exert different pharmacological effects is poorly understood at the metabolite level. A PCA revealed that A7r5 cells incubated with R- and S-Carvedilol were separately clustered along PC1. In this study, the differential metabolites were selected from the loading plot in PC1.

Xu et al. [21] investigated the metabolomic changes caused by simvastatin or fenofibrate intervention in diet-induced hyperlipidemia rats. They performed PCA on three datasets. In a comparative study of 6-week drug treatment, the control, simvastatin-treated and fenofibrate-treated groups were

此为试读，需要完整PDF请访问：www.ertongbook.com