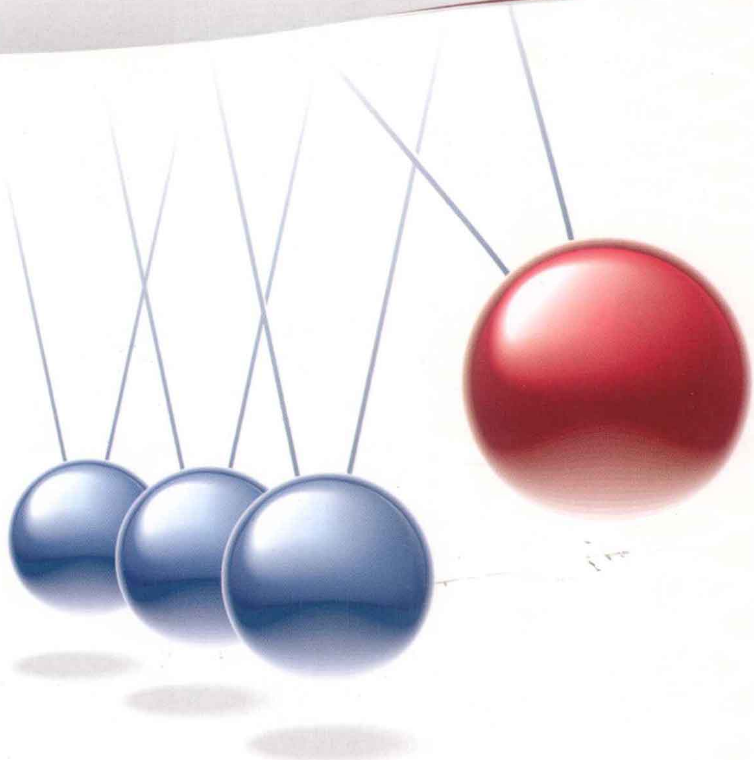
Some Topics in Dimension Reduction and Clustering

赵建华 著
Zhao Jianhua



中国统计出版社
China Statistics Press

（京）新登字 041 号

图书在版编目（CIP）数据

数据降维和聚类中的若干问题研究 = Some Topics in Dimension Reduction and Clustering : 英文 / 赵建华著. -- 北京 : 中国统计出版社, 2011.8
ISBN 978-7-5037-6307-6

Ⅰ.①数… Ⅱ.①赵… Ⅲ.①概率－数学模型－研究－英文 Ⅳ.①O21

中国版本图书馆 CIP 数据核字(2011)第 157193 号

数据降维和聚类中的若干问题研究
Some Topics in Dimension Reduction and Clustering

作　　者/赵建华　著
责任编辑/梁　超
装帧设计/黄　晨
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
电　　话/邮购（010）63376907　书店（010）68783172
印　　刷/北京顺义兴华印刷厂
经　　销/新华书店
开　　本/710×1000mm　1/18
印　　张/12.75
字　　数/230 千字
版　　别/2011 年 8 月第 1 版
版　　次/2011 年 8 月第 1 次印刷
书　　号/ISBN 978-7-5037-6307-6/O · 78
定　　价/26.00 元

Preface

A central research area in data mining and machine learning is probabilistic modeling because it has a number of advantages over non-probabilistic methods. Given a probabilistic model, one could fit the model using maximum likelihood (ML) method or Variational Bayesian (VB) method. In ML method, (1) many algorithms may converge very slowly and thus computationally efficient algorithms are often desirable; and (2) the choice of a suitable model is difficult though many model selection criteria exist and thus criteria with higher accuracy are desired. In VB method, employing different priors may yield different performances and thus studies on how to choose a suitable prior are important. In this book, three sub-topics were studied: *Modeling*, *Estimation* and *Model selection* for dimension reduction and clustering.

Modeling: To overcome the serious problems when probabilistic principal component analysis (PPCA) is applied to 2D data, a bilinear PPCA was proposed, which itself declares a breakthrough from traditional linear latent variable models to the bilinear ones. The result from our extensive empirical studies is encouraging.

Estimation: A new conditional maximization (CM) algorithm was proposed for ML estimation in factor analysis, which, like expectation maximization (EM) algorithm, is easy to implement and converge stably. The novelty is that our CM possesses quadratic convergence. Empirical results show that CM outperforms all existing competing algorithms. The CM algorithm for factor analysis was then extended to mixtures of factor analyzers, resulting in a fast expectation CM (ECM) algorithm. As revealed by experiments, the convergence of our ECM is substantially faster than that of existing algorithms. For VB estimation of factor analysis, existing works were found to suffer two serious problems theoretically and empirically. A novel VB treatment is proposed to resolve the two problems and a simulation study was conducted to testify its improved performance over existing treatments.

Model selection: A novel model selection criterion called hierarchical BIC (H-BIC) was proposed for mixture model selection using ML method. We showed theoretically and empirically that H-BIC is a large sample approximation of VB lower bound and the widely used Bayesian information criterion (BIC) is further an approximation of H-BIC.

Acknowledgements

I am very lucky to have the opportunity to pursue my PhD degree at The University of Hong Kong, where I have learned much academically and personally from my supervisor, Dr. Philip L.H. Yu, Dr. Gary G.L. Tian, Dr. Hu Yue-Qing and a number of people in HKU.

Contents

List of Figures

List of Tables

List of Acronyms

ECME Expectation Conditional Maximization of Either 9

AECM Alternating Expectation Conditional Maximization 10

MFA Mixtures of Factor Analyzers 9

MAP Maximum a Posteriori 10

BIC Bayesian Information Criterion 75

H-BIC Hierarchical Bayesian Information Criterion 75

BPPCA Bilinear Probabilistic Principal Component Analysis 129

2DSVD Two-dimensional Singular Value Decomposition 132

GLRAM Generalized Low Rank Approximations of Matrices 132

Chapter 1

Introduction

Dimension reduction is important in many disciplines such as botany, biology, bioinformatics, social sciences, economics, engineering, etc., because of one of the reasons including: (1) the interesting structure of the high dimensional data generally lies in a low dimensional space and thus more compact and meaningful representation for the data is required for visualization, interpretation, analysis, etc; (2) the dimensionality of the data is too high to be handled for certain algorithm. One example of high dimensional data is face recognition, where if face images are cropped to 40×50 pixels, then the resulting data dimension is two thousands and could be higher if larger size is used.

One way to reduce data dimension is to use *subset selection*, i.e., only select a subset of original features that retain the original information as much as possible according to certain criterion. However, in many applications, instead of original features themselves we are interested in the