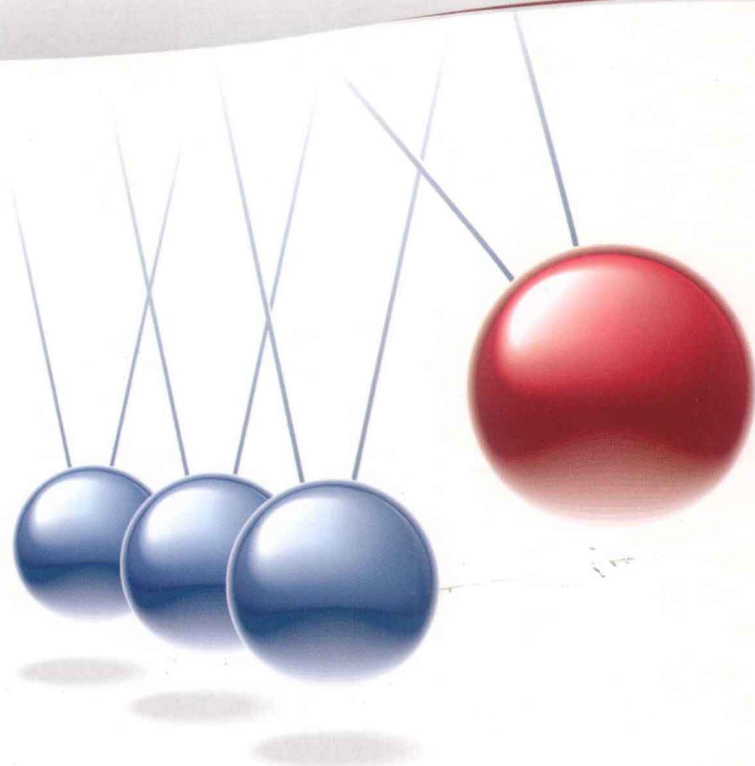


Some Topics in Dimension Reduction and Clustering

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

版权所有。未经许可, 本书的任何部分不准以任何方式在世界任何地区
以任何文字翻印、拷贝、仿制或转载。

中国统计版图书, 如有印装错误, 本社发行部负责调换。

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

1	Introduction	1
1.1	PCA and Latent Variable Models	4
1.1.1	PCA	4
1.1.2	Latent Variable Models	5
1.1.3	FA and PPCA	7
1.2	Motivations and Contributions	8
1.3	Organization of the Book	12
2	ML Estimation for Factor Analysis: EM or non-EM	13
2.1	Introduction	13
2.2	FA Model and Three Estimation Algorithms	16
2.2.1	FA model	16
2.2.2	Lawley (1940)'s simple iteration algorithm	17
2.2.3	EM type algorithms	19
2.3	The ECME2 algorithm	22

2.3.1	The maximization in the first CM-step	22
2.3.2	The maximization in the second CM-step	25
2.3.3	Practical consideration	26
2.3.4	ECME2 vs. simple iteration algorithm	27
2.4	The CM Algorithm	28
2.4.1	The maximization in the second CM-step	29
2.4.2	When will <i>condition I</i> be satisfied	31
2.4.3	Recursive computation of the matrix B_i^{-1}	33
2.4.4	On the nature of stationary points	34
2.5	Simulations	35
2.5.1	Simulation Data	35
2.5.2	Performance Analysis	40
2.5.3	On different starting values	47
2.6	Conclusion and Future Work	48
2.7	Appendix	50
2.7.1	Proofs	50
2.7.2	Some Notes	52
3	Fast ML estimation for the Mixture of Factor Analyzers via an ECM Algorithm	57
3.1	Introduction	57
3.2	MFA model and an ECM algorithm	60

3.2.1	The MFA model	60
3.2.2	The EM algorithm	60
3.2.3	The AECM algorithm	61
3.2.4	The ECM algorithm	63
3.2.5	Computational complexity	64
3.2.6	On speed of convergence	66
3.3	Experiments	67
3.3.1	Artificial data	68
3.3.2	Real data	68
3.4	Concluding remarks	73
4	Mixture Model Selection: BIC or Hierarchical BIC	74
4.1	Introduction	74
4.2	BIC and H-BIC	76
4.2.1	BIC	76
4.2.2	H-BIC	77
4.3	H-BIC: a large sample limit of variational Bayesian lower bound	78
4.4	Experiments	82
4.4.1	Artificial data	84
4.4.2	Wisconsin Diagnostic Breast Cancer Data	85
4.5	Conclusion and Discussions	87

4.6	Appendix	90
4.6.1	Proof of Proposition 4.1	90
4.6.2	Updating equations of MAPGMM	91
5	A Note on Variational Bayesian Factor Analysis	92
5.1	Introduction	92
5.2	FA model	94
5.3	VBFA1	96
5.3.1	Problem 1 of VBFA1	100
5.3.2	Problem 2 of VBFA1	101
5.4	VBFA2	102
5.4.1	The large sample limit of VBFA2	106
5.4.2	Backward learning of VBFA2	106
5.5	Simulations	108
5.5.1	VBFA2 vs. VBFA1	110
5.5.2	Model selection: VB vs. BIC	115
5.6	Concluding remarks and future work	116
5.7	Appendix	120
5.7.1	The lower bound of VBFA2	120
5.7.2	Justification of using constraint (5.24) on \mathbf{A}	121
5.7.3	On the rotation problem of VBFA-ARD	122
5.7.4	The large sample limit for VBFA1	126

6	Bilinear Probabilistic Principal Component Analysis	128
6.1	Introduction	128
6.1.1	Motivations.	129
6.1.2	Related works	132
6.2	Review of PPCA, GLRAM and 2DSVD	134
6.2.1	PPCA	134
6.2.2	GLRAM	137
6.2.3	2DSVD	138
6.3	BPPCA model	139
6.3.1	Matrix-variate normal distribution	139
6.3.2	BPPCA model	142
6.4	Maximum Likelihood Estimation of BPPCA	146
6.4.1	A CM algorithm	147
6.4.2	An AECM algorithm	149
6.4.3	Compression and reconstruction	151
6.5	Connections with PPCA, GLRAM and one-mode 2DPCA	154
6.5.1	Connection between BPPCA and PPCA	155
6.5.2	Connection between BPPCA and GLRAM	155
6.5.3	Connection between BPPCA and 2DSVD	157
6.5.4	Connection with one-mode 2DPCA	158

6.6	Experiments	158
6.6.1	Artificial data	158
6.6.2	Real data	161
6.7	Conclusion	169
6.8	Appendix	173
6.8.1	Optimal least squares reconstruction under BPPCA	173
6.8.2	Computational Complexity Analysis	175
7	Conclusions and discussions	177
	References	180

List of Figures

1.1	Summary of this book.	8
2.1	A plot of the typical evolvement of log-likelihood for EM, ECME2 and CM, fitted to data sets with different noise types.	38
2.2	The box plot of required time for convergence.	41
2.3	The box plot of required number of iterations for conver- gence.	42
2.4	A plot of the evolution of log-likelihood.	49
3.1	(a) The artificial data; (b), (c) and (d): The typical evolve- ments of log likelihood.	69
3.2	Real data: the typical evolvments of log likelihood.	71
4.1	An example of artificial data sets with different separation δ .	85
4.2	Percentage of success versus sample size n	86
4.3	Two-dimensional views of Wisconsin Diagnostic Breast Cancer Data.	88
4.4	Selected component number versus sample size n	89

5.1	Probabilistic graphical models for VBFA1 and VBFA2. . .	98
5.2	Typical Hinton diagrams of factor loadings by fitting VBFA1 and VBFA2.	111
5.3	(a)-(d), the fitting of VBFA1 and VBFA2 with $q = 9$ to data 2. (e) evolvement of \mathcal{F} for VBFA1 and VBFA2; (f) evolvement of \mathcal{F}_2 by the backward learning algorithm. . .	112
5.4	Learning curves of VBFA1, VBFA2 and MLFA.	114
5.5	Learning curves based on \mathcal{F}_1 , \mathcal{F}_2 , \mathcal{F}_2 -ARD and BIC. . . .	118
5.6	Hinton diagrams of factor loadings matrices.	125
6.1	Probabilistic graphical models of BPPCA.	144
6.2	(a) The averaged reconstruction errors of training data and test data; (b) The averaged angle between the estimated column principal subspace and the true one.	160
6.3	Row 1: the original images with index 2, 4, 6, 7, 9, 11. The reconstructed images by 2DSVD, GLRAM and BPPCA shown in row 2-10.	165
6.4	The matrix PCs $\mathbf{u}_{cj}\mathbf{u}'_{rk}$'s by 2DSVD, GLRAM and BPPCA and (j, k) is in the order shown at the top of figure. . .	166
6.5	The RMSRE versus q by 2DSVD, GLRAM and BPPCA. .	167
6.6	The averaged misclassification rates versus q by BPPCA, 2DSVD and GLRAM for different data sets.	170

List of Tables

2.1	Speedup by CM over EM and ECME2 in CPU time and number of iterations.	43
2.2	Speedup by CM over QN in CPU time and number of iterations.	45
2.3	Speedup by CM over ECME1 in CPU time and number of iterations.	47
2.4	Computation cost of different algorithms in each iteration.	55
3.1	Computational cost of different algorithms for each component.	65
3.2	The averaged CPU time and number of iterations for convergence.	70
3.3	The averaged required CPU time (in seconds) and number of iterations.	72
3.4	The averaged MSE and log likelihood from 10 runs.	72

5.1	Comparisons on latent dimension estimation among \mathcal{F}_1 , \mathcal{F}_2 , \mathcal{F}_2 -ARD and BIC.	117
6.1	The averaged lowest error rates in different models shown as mean(std.).	171
6.2	The optimal latent dimension q_{opt} yielding the lowest error rates by different methods.	172

List of Acronyms

PCA Principal Component Analysis	2
ML Maximum Likelihood	5
VB Variational Bayesian	2
FA Factor Analysis	5
PPCA Probabilistic Principal Component Analysis	3
CM Conditional Maximization	9
EM Expectation Maximization	3

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