

## DOCTORAL THESIS

### Computational methods for bioinformatics and image restoration

Liao, Haiyong

*Date of Award:*  
2010

[Link to publication](#)

#### **General rights**

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

# Computational Methods for Bioinformatics and Image Restoration

**LIAO Haiyong**

A thesis submitted in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy

Principal Supervisor: Prof. Michael K. Ng

Hong Kong Baptist University

January 2010

# Abstract

The main goal of Part I in this thesis is to study and analysis categorical data and Single Nucleotide Polymorphism (SNP) data in Human genome. To better understand the intrinsic structure in human genome, we propose and develop a unidimensional nonnegative scaling model to construct linkage disequilibrium (LD) maps in Chapter 2. The proposed constrained scaling model can be efficiently solved by transforming it to an unconstrained model. The method is implemented in PC Clusters at Hong Kong Baptist University. The LD maps are constructed for four populations from Hapmap data sets with chromosomes of several ten thousand single nucleotide polymorphisms (SNPs). The similarities and dissimilarities of the LD maps are studied and analyzed. Computational results are also reported to show the effectiveness of the method using parallel computation.

It is known that there are many categorical data existing in practice, for example in human genome analysis, market and customer analysis. However, many well-developed numerical methods cannot be applied to this kind of data directly. In Chapter 3, we investigate the problem of determining the number of clusters in the  $k$ -modes based categorical data clustering. We propose a new categorical data clustering algorithm with automatic selection of  $k$ . The new algorithm extends the  $k$ -modes clustering algorithm by introducing a penalty term to the objective function to make more clusters compete for objects. In the new objective function, we employ a regularization parameter to control the number of clusters in a clustering result. Instead of finding  $k$  directly, we choose a suitable value of regularization parameter such that the optimal clustering result is the most stable one among all the generated clustering results. Experimental results on synthetic data sets and real data sets are used to demonstrated the effectiveness of the proposed algorithm.

In Part II, some algorithms for image restoration are investigated. In the last decade total variation (TV) based methods has been being popular for image restoration because of edge-preserving property. There are many excellent TV-based algorithms proposed for image restoration. However, there is little research work on the choice of regularization parameter in TV based model. In Chapter 4, we try

different strategies to make the algorithms find the appropriate regularization parameter automatically. Numerical results are given to demonstrate the performance of the proposed algorithms. We also extend the proposed algorithm to blind image deconvolution (BID) problem in Chapter 5. Blind image deconvolution is a more challenging problem since the point spread function (PSF) is unknown. With Laplacian regularization for PSF and suitable imposed constraints on PSF, the proposed algorithm can recover both image and PSF successfully.

# Contents

<b>Declaration</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Table of Contents</b>	<b>v</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>xi</b>
<b>List of Algorithms</b>	<b>xiv</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 DNA, SNP and Human Genome . . . . .	1
1.1.1 Terms, Definitions and Abbreviations . . . . .	1
1.1.2 Single Nucleotide Polymorphisms . . . . .	3
1.1.3 Haplotype . . . . .	5
1.1.4 The International HapMap Project . . . . .	6
1.1.5 Complex Disease and the CD-CV Hypothesis . . . . .	8
1.1.6 The Goals of Human Genomic Data Analysis . . . . .	9
1.2 Categorical Clustering . . . . .	10
1.2.1 Types of Data . . . . .	10
1.2.2 Cluster Analysis . . . . .	11
1.2.3 Types of Cluster Analysis . . . . .	11
1.2.4 K-means and Its Derivatives . . . . .	12

1.3	Numerical Methods for Image Restoration . . . . .	16
1.3.1	Background and Problem Formulation . . . . .	16
1.3.2	Boundary Conditions and Structured Matrices . . . . .	17
1.3.3	DFT and DCT . . . . .	18
1.3.4	Types of PSFs . . . . .	20
1.4	Structure of This Thesis . . . . .	22

## **I Human Genome Data Analysis and Categorical Clustering**

### **24**

<b>Chapter 2</b>	<b>Unidimensional Scaling for Genome-wide LD Maps</b>	<b>25</b>
2.1	Related Work . . . . .	25
2.2	The Unidimensional Nonnegative Scaling Model . . . . .	27
2.3	Numerical Results . . . . .	31
2.3.1	HapMap Data Analysis . . . . .	31
2.3.2	Parallel Computing . . . . .	43
2.4	Summary . . . . .	45
<b>Chapter 3</b>	<b>Categorical Clustering with Cluster Number Selection</b>	<b>47</b>
3.1	Background . . . . .	47
3.2	Related Work . . . . .	48
3.2.1	Cluster Validation . . . . .	48
3.2.2	Categorical Data Clustering . . . . .	49
3.3	Regularized Fuzzy $k$ -modes . . . . .	50
3.3.1	Optimization Procedure . . . . .	51
3.3.2	The Cluster Structure . . . . .	53
3.4	Experimental Results . . . . .	55
3.4.1	Generation of Synthetic Data Sets . . . . .	55
3.4.2	An Example . . . . .	56
3.4.3	Synthetic Data Sets Study . . . . .	59
3.4.4	Real Data Sets Study . . . . .	62
3.5	Summary . . . . .	70

<b>II</b>	<b>Numerical Methods for Image Restoration</b>	<b>71</b>
<b>Chapter 4</b>	<b>Parameter Selection for TV Image Restoration</b>	<b>72</b>
4.1	Background . . . . .	72
4.2	Previous Work on Total Variation Image Restoration . . . . .	73
4.3	Alternating Minimization Algorithm for TV Image Restoration. . . . .	75
4.4	Selection of Regularization Parameter for TV Image Restoration . . . . .	78
4.4.1	L-curve . . . . .	79
4.4.2	Discrepancy Principle . . . . .	80
4.4.3	Generalized Cross Validation . . . . .	81
4.5	Experimental Results . . . . .	83
4.5.1	Comparison on Regularization Parameter Selection Methods . . . . .	84
4.5.2	Test on Algorithm 5: L2-TV . . . . .	85
4.5.3	Test on Algorithm 6: L1-TV . . . . .	93
4.5.4	Computational Issues on the Two Algorithms . . . . .	96
4.6	Summary . . . . .	98
<b>Chapter 5</b>	<b>Blind Deconvolution with GCV for Parameters Estimation</b>	<b>99</b>
5.1	Previous Work . . . . .	99
5.2	GCV-based Blind Deconvolution Algorithm . . . . .	102
5.2.1	Alternating Minimization Scheme . . . . .	102
5.2.2	Restoration of Image . . . . .	103
5.2.3	Regularization Parameter Estimation in Image Restoration . . . . .	104
5.2.4	Restoration of PSF . . . . .	104
5.2.5	Regularization Parameter Estimation in PSF Restoration . . . . .	105
5.2.6	Remarks on the Restoration of Image and PSF . . . . .	106
5.3	Experimental Results . . . . .	106
5.3.1	Test on "Satellite" Image . . . . .	107
5.3.2	Comparisons with Non-blind Deconvolution Method . . . . .	109
5.3.3	Comparisons with Variational Bayesian BD Methods . . . . .	113
5.3.4	Test on Sensitivity of Parameter Setting . . . . .	117
5.4	Summary . . . . .	118

<b>Chapter 6 Conclusion and Future Work</b>	<b>119</b>
<b>Bibliography</b>	<b>121</b>
<b>Curriculum Vitae</b>	<b>133</b>