ON THE GEOMETRY
OF
GENERALIZED LINEAR MODELS

By

Dongwen Luo

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
MASSEY UNIVERSITY
PALMERSTON NORTH, NEW ZEALAND
FEB 2003

To Jasmine

Acknowledgements

I would like to thank Professor Graham Wood and Dr. Geoff Jones, my supervisors, for their many suggestions and constant support during this research. I have greatly benefited in many ways from their intelligence and wisdom and have been deeply rewarded by their supervision of my PhD program. I will never forget those regular weekly meeting and invaluable discussions. I am certain that everything I have learned throughout my studies will, in one way or another, benefit my whole life.

I would also like to thank Associate Professor Chin Diew Lai, Dr. Siva Ganesh and Dr. Chungui Qiao who provided continuous assistance and help during my study at Massey University.

My special thanks go to the Department of Statistics, Macquarie University, for providing me with a position as an exchange scholar so that I could concentrate upon and complete my PhD, and particularly to Professor H. M. Hudson for his encouragement.

Finally, my deepest thanks go to my wife Caiqin Liu and my daughter Jasmine Luo for their constant support. I also thank my father Changhe Luo and mother Qiaolin He for their support. I really feel that I owe them so much because in order to support my study, they have had to sacrifice some family life. I hope the debt can be paid back in the near future.

Abstract

The perspective afforded by Euclidean geometry led to the rapid development of linear models in the early stages of the twentieth century: Fisher saw the data as a point in finite-dimensional Euclidean space, the model as a subspace and least squares fitting as projection of the observation vector onto the model space. From the late 1960s to early 1970s, Fienberg revealed geometry underlying loglinear models for two-way tables, while Haberman discussed geometry for the log-transformed case. Generalized linear models, however, have largely eluded geometers until recently. In 1997 an extension of Fisher's view to generalized linear models was given by Kass and Vos, using the language of differential geometry.

The aim of this work is to develop a simple, general geometric framework for generalized linear models, closely related to the thinking of Fienberg and Haberman. Whereas Kass and Vos developed a geometric view which leads to the usual scoring method, we develop geometry which leads to a new algorithm. A linearization of this new algorithm yields the scoring method. The geometry discussed by Kass and Vos is based on the log-likelihood function whereas the geometry developed here depends on sufficiency.

In the geometry of generalized linear models, developed through chapters 1 to 3, an observation with n values is viewed as a vector in Euclidian space $\mathbf{R}^n$. This Euclidian space $\mathbf{R}^n$ is partitioned into two orthogonal spaces, the sufficiency space $\mathcal{S}$ and the auxiliary space $\mathcal{A}$, with respect to a new basis. We focus on two mean sets relating to generalized linear models, one for the untransformed model space and another for the link-transformed model space. There are two critical properties of the maximum likelihood estimate of the parameters of a generalized linear model with canonical link. The first property is that the coefficients of the basis of the sufficiency space, the sufficient statistics, are preserved in the untransformed model space in the fitting process. The second property is that the coefficients of the basis of the auxiliary space are zeroed in the link-transformed model space in the fitting process. Linear models and loglinear models serve as special cases of generalized linear models with identity and log link respectively.

Based on the geometric framework discussed in the thesis, a new algorithm is constructed for fitting generalized linear models with canonical link in Chapter 4. This algorithm, which relies on sufficient statistics for the parameters in the model rather than the likelihood function, takes two projections alternately, orthogonal projection onto a sufficiency affine plane and non-orthogonal projection onto the transformed model space. In the process, we match the model space and sufficient statistics iteratively until convergence. Linearization of the new algorithm induces the scoring method.

In Chapter 5 we pay special attention to a subset of loglinear models, graphical loglinear models, those which are the intersection of a finite set of conditional independence statements. The model space of one conditional independence statement is described through the notions of "corresponding point convex hull" and "set convex hull". The fitting of one conditional independence statement is considered geometrically using a direct fitting method and the familiar iterative proportional fitting method.

Table of Contents

Acknowledgements **iii**

Abstract **iv**

Table of Contents **vi**

1 Introduction **1**

2 The geometry of categorical data models **13**

2.1 Introduction 13

2.2 Categorical data models 14

2.3 Loglinear models and sufficient statistics 21

2.4 Geometry of loglinear models 27

2.4.1 Geometry of a 2 × 2 table 28

2.4.2 The general case 41

2.4.3 Some examples 45

2.5 Conclusion 50

3 The geometry of GLMs **52**

3.1 Introduction 52

3.2 GLMs and sufficient statistics 53

3.3 Kass and Vos approach 57

3.4 An alternative geometric approach 63

3.5 Example 70

3.6 Conclusion 75

4 A new algorithm for fitting GLMs **79**

4.1 Introduction 79

4.2 Geometry of the scoring method 80

4.3 The new algorithm 85
4.4 A detailed example 92
4.5 Link between the two algorithms 93
4.6 Numerical comparison of the two algorithms 95
4.7 Conclusions 97

5 The geometry of conditional independence statements 98
5.1 Introduction 98
5.2 Technical preliminaries 103
5.3 Geometric setting for distributions 107
5.4 Geometric setting for CI models 116
5.5 The MLE of a distribution satisfying
$A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ 129
5.6 Conclusion 142

6 Conclusions 144
6.1 Thesis Contribution 144
6.2 Further Research Directions 146

A Matlab functions 148

B Data sources 160

Bibliography 164