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

Chapter 1

Introduction

Statistical models are generally described algebraically, but often they can also be described geometrically. There are two geometric points of view taken with respect to statistical models in the literature: Euclidean geometry and its extension, differential geometry. Euclidean geometry provides an elegant and unified framework for description of linear models (Saville and Wood, 1991) and for multivariate analysis (Dempster, 1969). Differential geometry has initiated crucial advances in a variety of fields of statistics, including the development of new geometries for statistical models (Barndorff-Nielsen, 1987), higher order asymptotic theory (Amari, 1982), invariant asymptotic expansions and inference in nonlinear regression and curved exponential families (Kass and Vos, 1997).

R.A. Fisher's 1915 paper on the distribution of the correlation coefficient was the initial paper placing a statistical model in a Euclidean framework. Following Fisher's ideas, Bartlett (1933 – 1934) discussed the geometry of a Latin square design, Durbin and Kendall (1951) studied the geometry of finding the estimator for one-way ANOVA, and Kruskal (1961) Zyskind (1967) and Watson (1967) described least squares estimation of linear models geometrically. Box, Hunter and Hunter (1978)

presented geometry for specific types of experimental designs. A treatment of the geometry of linear models was given in Christensen (2002) and separately in Saville and Wood (1991).

According to Box (1978), Fisher (1925) had seen the data as a point in finite-dimensional Euclidean space, the linear model as a subspace and least squares fitting as projection of the observation vector onto the model space. Unfortunately, Fisher found his geometric approach was not generally easily understood, so the geometric results were expressed in algebraic form. Saville and Wood (1991), amongst many others, retrieve Fisher's lost insight using simple linear algebra. For example, an ANOVA table is an account of an orthogonal breakup of a vector, with degrees of freedom the dimension of a subspace and sum of squares the squared length of a projected vector.

On the other hand, the history of setting statistical models in a differential geometric framework can be traced back to research by C.R. Rao (1945) and Harold Jeffreys (1948). They used the Fisher information matrix to define a Riemannian metric on a statistical manifold. It was Efron (1975) who defined the statistical curvature of a statistical model so drawing substantial attention to the role of differential geometry in statistics. Furthermore, Efron (1978) discussed the geometry of exponential families, the distributions that generalized linear models follow, using the concept of statistical curvature. Under the strong influence of Efron's paper Amari (1990) constructed a very elegant representation and elaboration of Fisher's theory of information loss. Recently, Kass and Vos (1997) summarized the Fisher-Efron-Amari theory and the Jeffrey-Rao Riemannian geometry using Fisher information to construct the geometry of curved exponential families. In this context they discuss the

geometry of generalized linear models.

The geometry of generalized linear models described by Kass and Vos (1977) uses differential geometry based on the log-likelihood function. This geometric framework leads to the scoring method. In this thesis, we develop a simple, general geometric framework for generalized linear models using only Euclidean geometry. This new geometric framework, which depends on sufficiency, leads to a new algorithm for fitting generalized linear models with canonical link. A linearization of the new algorithm yields the scoring method. Our work closely relates to the thinking evident in the development of a geometric framework for loglinear models, a special case of generalized linear models, by Fienberg (1968) and Haberman (1974).

In 1968 Fienberg represented a two-way table with n cells and entries in the form of probabilities, as a point within an $n-1$ dimensional simplex S_{n-1} in $\mathbf{R}^n$. There are several types of two-way tables characterized as subsets of the simplex S_{n-1}, including tables whose rows and columns are independent, tables with a given interaction structure, and tables with a fixed set of margins. On the other hand, Haberman (1974) viewed a log-transformed table (not necessarily two-way) with n cells and entries in the form of counts as a vector in Euclidean space $\mathbf{R}^n$ and the model space of a loglinear model as a subset in $\mathbf{R}^n$. Fitting a loglinear model maps the observation vector to a q-dimensional ($q \leq n$) model space contained in $\mathbf{R}^n$ (where q is the number of parameters of the loglinear model). Here, we combine these two geometric views of loglinear models, which we term "Fienberg geometry" and "Haberman geometry", and link them in a commutative diagram. As with linear models, the space $\mathbf{R}^n$ is partitioned into two orthogonal subspaces, $\mathcal{S}$ (called the sufficiency space) and $\mathcal{A}$ (called the auxiliary space). Two important geometric properties, then, are revealed

on the sufficiency and auxiliary spaces. Further it is shown that all results for loglinear models can be extended to a geometry of generalized linear models, where the new algorithm is constructed.

Recently a subset of loglinear models, graphical loglinear models, have been extensively studied. Those models are the intersection of a finite set of conditional independence statements (Darroch et al., 1980). In this thesis, we describes a conditional independence model space through the notions of "corresponding point convex hull" and "set convex hull". The workings of iterative proportional fitting, and also a direct fitting method for finding the maximum likelihood estimate of a conditional independence model, are described in this geometric framework.

In order to lay a foundation for the geometric framework constructed in this thesis, we will now review the geometry of linear models from Fisher's point of view, including the traditional fitting methods for linear models, namely the least squares method and the maximum likelihood method. Two geometric properties related to the geometry to be described later are emphasized and demonstrated by an example. Finally, the structure of the whole thesis is outlined.

Consider a linear model with dependent variable Y and the design matrix X. The linear model has matrix form

$$Y = X\beta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I_n) \tag{1.1}$$

where β is a parameter vector $[\beta_1, \beta_2, \ldots, \beta_q]^T$ (where "T" denotes transpose) to be

estimated, ε is an error vector $[\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T$, and the design matrix X has form

$$\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1q} \\ x_{21} & x_{22} & \ldots & x_{2q} \\ \vdots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nq} \end{bmatrix} = [x_1\ x_2\ \ldots\ x_q]$$

where $x_j = [x_{1j}, x_{2j}, \ldots, x_{nj}]^T$ for $j = 1, 2, \ldots, q$. Note that in ANOVA models the column vectors of the design matrix are contrasts of interest. If realized values of Y are $y_1, y_2, \ldots, y_n$ $(n > q)$, then these realized values can be represented by a vector y in the Euclidean vector space $\mathbf{R}^n$. The vector y has the coordinates $(y_1, y_2, \ldots, y_n)$ with respect to the standard basis $\{e_1, e_2, \ldots, e_n\}$ in $\mathbf{R}^n$.

We assume that there is no collinearity here, so the column vectors of X are linearly independent. After applying a variation of the Gram-Schmidt process we can construct a new basis $\{x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_n\}$ instead of the standard basis in $\mathbf{R}^n$ such that $x_i \cdot x_j = 0$ for $i = 1, 2, \ldots, q$ and $j = q+1, q+2, \ldots, n$. Now the whole space $\mathbf{R}^n$ can be partitioned into two orthogonal spaces, specifically

$$\mathbf{R}^n = \mathbb{M} \oplus \mathbb{E}$$

where $\mathbb{M} = \text{span}\{x_1, x_2, \ldots, x_q\}$, called the model space, and $\mathbb{E} = \text{span}\{x_{q+1}, x_{q+2}, \ldots, x_n\}$, called the error space, with $\mathbb{E} = \mathbb{M}^{\perp}$. Later (in Chapter 3), the model space $\mathbb{M}$ becomes the sufficiency space $\mathcal{S}$, and the error space $\mathbb{E}$ becomes the auxiliary space $\mathcal{A}$.

For linear models, an estimate of the parameter β can be found by the least squares method or maximum likelihood method. The least squares method was discovered independently by Adrien Marie Legendre and Carl Friedrich Gauss (Draper and Smith

1998, p.45). The estimate $\hat{\beta}$ of β is found by minimizing the residual sum of squares

$$Q = \|y - X\beta\|^2$$

Following Herr (1980), Q is the squared distance of y from the model space $\mathbb{M}$. Minimizing Q corresponds, then, to finding the point in $\mathbb{M}$ closest to y. The answer is readily visualized as the "point in $\mathbb{M}$ directly below y," the orthogonal projection of y on $\mathbb{M}$, so $X\hat{\beta}$ is unique and satisfies

$$y = X\hat{\beta} + z, \quad z \text{ where is perpendicular to } \mathbb{M}$$

Multiplying both sides by X^T we have

$$X^T y = X^T X\hat{\beta}$$

since $X^T z = 0$. Thus

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

since the column vectors of X are linear independent and hence $X^T X$ is invertible. It can be shown that a least squares estimator is an unbiased and minimum variance estimator (George and Roger 1990, p.564).

On the other hand, we can estimate β by maximizing likelihood. For a given observation vector $y = [y_1, y_2, \ldots, y_n]^T$, the likelihood function for the linear model (1.1) is

$$l(\beta \mid y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2}\frac{\|y - X\beta\|^2}{\sigma^2}}$$

This likelihood function depends on β solely through the distance $Q = \|y - X\beta\|^2$, so the maximum likelihood is achieved by minimizing the distance Q. Thus for a linear model the estimate of β is the same for both maximum likelihood and least squares

methods. Note that from the form of the likelihood function the contours, defined by $l(\beta \mid y) = c$ where c is a constant, are spherically symmetrical about $X\beta$, a vector in the model space $\mathbb{M}$. We illustrate these ideas now with a simple example.

The very simple model $Y = \mu + \varepsilon$ with two observations has model form

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \qquad \varepsilon \sim N(0, \sigma^2 I_2)$$

Here we have $Y = [Y_1, Y_2]^T$, $x_1 = [1, 1]^T$ and parameter μ, so the model space is $\mathbb{E} = \text{span}\{x_1\}$, the equiangular line in $\mathbf{R}^2$. The least squares estimate $\hat{\mu}$ of μ is formed by projecting $y = [y_1, y_2]^T$ onto $[1, 1]^T$ where

$$\begin{aligned} \hat{\mu} \begin{bmatrix} 1 \\ 1 \end{bmatrix} &= \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Big/ \sqrt{2} \right) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Big/ \sqrt{2} \\ &= \begin{bmatrix} \frac{y_1 + y_2}{2} \\ \frac{y_1 + y_2}{2} \end{bmatrix} \\ &= \bar{y} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

where $\bar{y} = (y_1 + y_2)/2$, so $\hat{\mu} = \bar{y}$.

Now we consider estimation of μ using maximum likelihood. The estimate $\hat{\mu}$ of μ is determined by searching along the equiangular direction (the model space $\mathbb{M}$). For a given data vector y, the maximum likelihood estimate is obtained when $\hat{\mu} = \bar{y}$, since the length of $\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \end{bmatrix}$ is less than the length $\begin{bmatrix} y_1 - \mu \\ y_2 - \mu \end{bmatrix}$ for any $\mu \neq \bar{y}$ and $[\mu, \mu]^T \in \mathbb{M}$ (see Figure 1.1). Thus the maximum likelihood estimate of μ is the same as the least squares estimate of μ.

There are two geometric properties for linear models we want to emphasize here.

Property 1. The observation vector y and its fitted vector $\hat{y}$ have the same projection onto the model space.

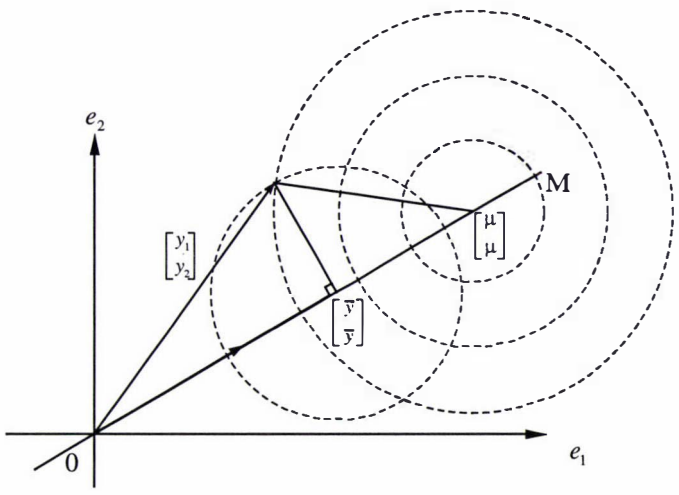


Figure 1.1: The maximum likelihood estimate is the same as the least squares estimate for μ in the model $Y = \mu + \varepsilon$.

Proof. Since $\hat{y} = X\hat{\beta}$ and $\hat{\beta} = (X^T X)^{-1} X^T y$, then

$$\begin{aligned} X^T \hat{y} &= X^T X \hat{\beta} \\ &= X^T X (X^T X)^{-1} X^T y \\ &= X^T y \end{aligned}$$

Thus $\hat{y}.\, x_j = y.\, x_j$ for $j = 1, 2, \ldots, q$. □

Property 2. The fitted vector $\hat{y}$ has zero projection onto the error space.

Proof. Since $\hat{y} = X\hat{\beta} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1 + \ldots + \hat{\beta}_q x_q$ where $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q]^T$, we have $\hat{y} \in \mathbb{M}$. Now $\mathbb{E} = \mathbb{M}^{\perp}$ so we obtain $\hat{y}.\, x_j = 0$ for $j = q+1, q+2, \ldots, n$. □

Therefore with respect to the new basis $\{x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_n\}$ the first property shows that the first q coordinates of the observation y will be

$$\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q$$

The second property indicates that the last $(n-q)$ coordinates of the fitted value $\hat{y}$ are zeros.

Now in order to estimate the parameters β geometrically, we only need change the coordinates system in $\mathbf{R}^n$ from the standard basis to the new basis; the estimate $\hat{\beta}$ is just the coordinates of the observation y with respect to the new basis of the model space. Later (in Chapter 3), we will see that these coordinates are sufficient statistics for the parameter β; the sufficient statistics play a key role in the geometry developed in this thesis. Coordinates with respect to the standard basis are mapped to coordinates with respect to the new basis by the transformation

$$A = [x_1\ x_2\ \ldots\ x_q\ x_{q+1}\ \ldots\ x_n]^{-1}$$

where A is called the change of basis matrix.

These geometric properties of linear models can be illustrated by the following example.

Example: Suppose we have an observation vector $y = [1,2,3]^T$ and independent variables have values $x_1 = [1,1,1]^T$ and $x_2 = [2.5,1,3]^T$. We now calculate the estimates of parameters β_0 and β_1 for the model

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & 2.5 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I_3)$$

Step 1 Construct a new basis for $\mathbf{R}^3$ by extending $\{x_1, x_2\}$ to $\{x_1, x_2, x_3\}$ where $x_3 = [-0.7845, 0.1962, 0.5883]^T$ (using the variation of the Gram-Schmidt process). Note that $x_1.x_3 = x_2.x_3 = 0$.

Step 2 Partition the whole space $\mathbf{R}^3$ into two orthogonal spaces as

$$\mathbf{R}^3 = \mathbb{M} \oplus \mathbb{E}$$

where the model space $\mathbb{M} = \text{span}\{x_1, x_2\}$ and the error space $\mathbb{E} = \text{span}\{x_3\}$.

Step 3 Find the coordinates of the observation vector y on the new basis as

$$\begin{bmatrix} 1 & 2.5 & -0.7845 \\ 1 & 1 & 0.1962 \\ 1 & 3 & 0.5883 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1.4998 \\ 0.2309 \\ 1.3728 \end{bmatrix}$$

Step 4 Estimate the parameters β_0 and β_1 as the coordinates of the observation vector y on the model space basis x_1 and x_2, so $\hat{\beta}_0 = 1.4998$ and $\hat{\beta}_1 = 0.2309$. Thus the fitted value of the observation vector y is

$$\hat{y} = \begin{bmatrix} 1 & 2.5 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1.4998 \\ 0.2309 \end{bmatrix} = \begin{bmatrix} 2.0770 \\ 1.7307 \\ 2.1925 \end{bmatrix}$$

The fitting process for the above example is illustrated in Figure 1.2.

In the next chapter, the geometry of categorical data models (representable by loglinear models) will be discussed and the two geometric properties revealed here will be extended. Two geometric approaches to categorical data models, one by Fienberg and the other by Haberman, are embedded in a unified geometric framework in which the whole space is split into a sufficiency space and an auxiliary space. We find that, in the fitting process, the coordinates of the basis of the sufficiency space are preserved in Fienberg geometry, while the coordinates of the basis of the auxiliary space are zeroed in Haberman geometry. The relationship between the two geometries is highlighted by a commutative diagram.

Chapter 3 reveals a geometric framework for generalized linear models. Here an existing geometry of generalized linear models, discussed by Kass and Vos, is reviewed with an example. As in Chapter 2, the whole space is split into a sufficiency space

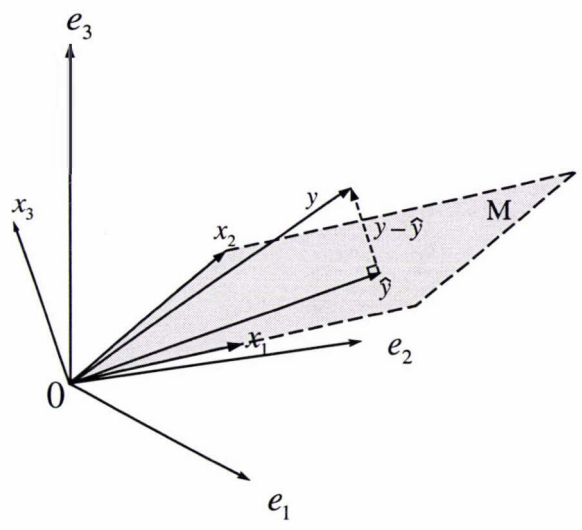


Figure 1.2: This graphic shows how to estimate parameter β for the linear model $Y = X\beta+\varepsilon$ with given data $y = [1, 2, 3]^T$ and independent variables $x_1 = [1, 1, 1]^T$ and $x_2 = [2.5, 1, 3]^T$ geometrically. Here we simply change the standard basis $\{e_1, e_2, e_3\}$ to the new basis $\{x_1, x_2, x_3\}$. Then y has coordinates $[1.4998, 0.2309, 1.3728]^T$ with respect to the new basis, so the estimates are $\hat{\beta}_0 = 1.4998$ and $\hat{\beta}_1 = 0.2309$, the coordinates of y with respect to x_1 and x_2 respectively.

and an auxiliary space. Again we find that in the Fienberg geometry the coefficients of the basis of the sufficiency space are preserved during model fitting, while in the Haberman geometry the coefficients of the basis of the auxiliary space are zeroed.

A new algorithm is constructed for fitting generalized linear models with canonical link in Chapter 4. This algorithm depends on sufficient statistics, and uses two projections alternately, orthogonal projection onto the sufficiency affine plane and non-orthogonal projection onto the transformed model space. In the process, we match sufficient statistics and the model space iteratively until convergence. A linearization of the new algorithm yields the scoring method. The geometry of the scoring method is given using Kass and Vos' approach.

The geometry of graphical loglinear models, a subset of loglinear models, is considered in Chapter 5. Graphical loglinear models occur as intersections of finite sets of conditional independence models. The model space of a conditional independence model with categorical variables is a highly structured subset within a simplex. Here we describe a conditional independence model space using the concepts of "corresponding point convex hull" and "set convex hull". In this geometric framework, two methods (the iterative proportional fitting method and the direct fitting method) for finding the maximum likelihood estimate of a conditional independence model are described.

Chapter 6 summarizes the main results of the whole thesis and highlights directions for further research.

Chapter 2

The geometry of categorical data models

2.1 Introduction

Probabilistic models for measurement variables are commonly based on the normal distribution and modelling interest centres on an additive decomposition of the observation mean μ. Linear models, developed by Gosset and Fisher early last century, have become the workhorses of statistical analysis in handling such situations. Probabilistic models for categorical variables, however, focus on a multiplicative decomposition of a probability π. Such models capture the conditional independence structure of the variables under study and so-called "loglinear" models have become a standard tool in this area. The basis for these models first appeared in Roy and Kastenbaum (1956), with prominent later exposition and development by Bishop, Fienberg and Holland (1975) and Agresti (1990).

From Chapter 1, we see that geometry made a significant contribution to the development of linear models. For loglinear models, cell probabilities π are transformed by a logarithm. Thus, two apparently quite distinct approaches to the underlying

geometry exist in the literature, a description of the untransformed π, as described by Fienberg (1968, 1970) and a description of the transformed $\log \pi$, as described by Haberman (1974). Ideas, which motivate the general work in Chapter 3, will be established for loglinear models in this chapter.

In the next section we set the scene for categorical data models by working through the simplest case, a 2×2 contingency table, using an Australia survey data set. In Section 2.3 we turn to loglinear models and sufficient statistics, the foundations of the geometry of loglinear models. Section 2.4 builds a geometry for categorical data models by linking the Fienberg and Haberman geometries. Here we describe a new basis for an associated Euclidean space, motivated by the sufficient statistics of the saturated categorical data model, then illustrate the partitioning of Euclidean space associated with any unsaturated model. We draw the two geometries together and describe the way in which they are linked. The chapter concludes with a summary.

2.2 Categorical data models

In order to illustrate the essential ideas behind categorical data models we consider the very simple case of a 2×2 contingency table. An example of such a table, together with the underlying model parameters, is given in Figure 2.1, using data from a recent Australian survey of attitudes to genetic engineering (Norton et al., 1998). The total number of respondents was 894 which is distributed among four categories defined by income level and attitude. The question of interest is whether the attitude to genetic engineering is influenced by the income level.

Three common distributional assumptions are made depending on the sample

Income	Attitude: For	Attitude: Against
Low	258	222
High	263	151

(a)

X_1 \ X_2		
	π_{11}	π_{12}
	π_{21}	π_{22}

(b)

Figure 2.1: In (a) is shown a cross-tabulation of income level against acceptance of genetic engineering, with data drawn from a recent Australia-wide survey. In (b) notation for the underlying cell probabilities is presented.

scheme: cell counts are either Poisson, multinomial, or product multinomial. We illustrate each of these sampling schemes by reference to the Australia survey data.

(i) Nothing fixed by design.

In reality, 2500 survey letters were posted out and a return date specified. After this date, information was recorded on income level and attitude towards genetic engineering from the returned surveys. The number of people who would reply was unknown at the start of the survey, so the observed counts in each cell are considered as independent Poisson random variables with means $N\pi_{ij}$ for all i and j, where N is the total sample size, here the number of respondents.

(ii) Total sample size fixed.

If it were possible to fix the total response of the 894 in advance, then 894 observations would be distributed among the four categories with probabilities π_{ij} for all i and j. Then the appropriate distributional model is a multinomial distribution.

(iii) One or more margins fixed.

Such a scheme would arise if we decided to investigate attitude towards genetic engineering within two groups of people (Low income and High income) and stop the mail checking when we have received say 480 letters from the low income group and 414 letters from the high income group. Thus, the margin of Income is fixed in advance. The cells are now divided into two sets, each set having an independent multinomial distribution for a given margin of Income. Jointly the whole table follows a product multinomial distribution.

Fortunately, the multinomial and Poisson sampling schemes have the same maximum likelihood estimates of the cell probabilities for a given categorical data model. These will, however, equal the maximum likelihood estimates for the product multinomial only when the terms associated with the fixed margins are included in the model. Otherwise, there is a contradiction with the product multinomial sampling scheme. For example, in the Australian survey data if the margin of Attitude was assumed fixed, the term representing the main effect of Attitude must be included the model.

The variables in categorical data models can be classified as response and explanatory variables as with the traditional linear models, but they can also be all jointly regarded as responses, modelling the cell probabilities to reveal the relationship among the variables. We will denote an observed relative frequency table as $\{p_{ij}\}$, and the underlying true probability table as $\{\pi_{ij}\}$ for $i = 1, 2$ and $j = 1, 2$ (as shown in Figure 2.1(b)). In the Australian survey data we have two response variables, Income (denoted X_1) and Attitude (denoted X_2). Our interest is in whether X_1 and X_2 are independent, written $X_1 \perp\!\!\!\perp X_2$, i.e. in the dependence structure of the joint distribution $\pi = [\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}]^T$. Four multiplicative decompositions of

π now are discussed.

1. Constant model

$$\pi_{ij} = \mu \quad \text{for all } i, j$$

This model indicates that all cells in the table have the same probability μ, so for any 2×2 table we have an ML estimate $\hat{\mu} = 0.25$ as shown in Figure 2.2 (Note that an ML estimator of a parameter is denoted by adding a '$\wedge$' sign over the associated parameter.) It is clear that the variables X_1 and X_2 have no effects here. Since the model includes the constant term only, it is written symbolically as (1).

		X_2
	0.2500	0.2500
X_1		
	0.2500	0.2500

Figure 2.2: This table shows the constant model for a 2×2 table.

2. One-way model

$$\pi_{ij} = \mu \theta_i^{X_1} \qquad \text{for all } i, j$$

where $\theta_i^{X_1}$ represents the X_1 effect at level i with constraint $\prod_i \theta_i^{X_1} = 1$ to achieve identifiability.

This model indicates that cell probabilities in the table are the same within each row but may vary between rows. A fitted table for the Australia survey data is shown in Figure 2.3 (1) with estimates $\hat{\mu} = 0.2493$, $\hat{\theta}_1^{X_1} = 1.0768$, and $\hat{\theta}_2^{X_1} = 0.9287$. Since the model includes the constant term and the X_1 effect term, it is represented by the

model symbol (1, X_1).

Alternatively, we could have

$$\pi_{ij} = \mu \theta_j^{X_2} \qquad \text{for all } i, j$$

where $\theta_j^{X_2}$ represents the X_2 effect at level j with constraint $\prod_j \theta_j^{X_2} = 1$.

This model indicates that cell probabilities in the table are the same within each column but may vary between columns. A fitted table for the Australia survey data is shown in Figure 2.3 (2) with estimates $\hat{\mu} = 0.2466$, $\hat{\theta}_1^{X_2} = 1.1819$, and $\hat{\theta}_2^{X_2} = 0.8461$, where $\theta_j^{X_2}$ represents the X_2 effect at level j. Since the model includes the constant term and X_2 effect term, it is represented by the model symbol (1, X_2).

(1)

X_1 \ X_2		
	0.2685	0.2685
	0.2315	0.2315

(2)

X_1 \ X_2		
	0.2914	0.2086
	0.2914	0.2086

Figure 2.3: In the case of 2 × 2 table, (1) shows a fitted table with X_1 effects and (2) a fitted table with X_2 effects for the Australia survey data employing a one-way model.

3. Two-way model

$$\pi_{ij} = \mu \theta_i^{X_1} \theta_j^{X_2} \qquad \text{for all } i, j$$

where $\theta_i^{X_1}$ and $\theta_j^{X_2}$ represent the ith level of X_1 and jth level of X_2 effects respectively with constraints $\prod_i \theta_i^{X_1} = \prod_j \theta_j^{X_2} = 1$.

This model indicates that the cross-product ratio of the cell probabilities (the odds ratio) equals one in the table, the model where $X_1 \perp\!\!\!\perp X_2$. A fitted table for the Australia survey data is shown in Figure 2.4 with estimates $\hat{\mu} = 0.2459$, $\hat{\theta}_1^{X_1} = 1.0768$, $\hat{\theta}_2^{X_1} = 0.9287$, $\hat{\theta}_1^{X_2} = 1.1818$, and $\hat{\theta}_2^{X_2} = 0.8461$. Since the model includes the constant, X_1 and X_2 effect terms, it is represented by the model symbol $(1,\ X_1,\ X_2)$.

	X_2	
X_1	0.3129	0.2240
	0.2699	0.1932

Figure 2.4: This figure shows a fitted table for the Australia survey data using a two-way model.

4. Saturated model

$$\pi_{ij} = \mu\, \theta_i^{X_1}\, \theta_j^{X_2}\, \theta_{ij}^{X_1X_2} \qquad \text{for all } i, j$$

where $\theta_i^{X_1}$ and $\theta_j^{X_2}$ represent the X_1 and X_2 effects respectively, while $\theta_{ij}^{X_1X_2}$ models the dependence between X_1 and X_2. The model has constraints $\prod_i \theta_i^{X_1} = \prod_j \theta_j^{X_2} = 1$, $\prod_i \theta_{ij}^{X_1X_2} = 1$ for j fixed, and $\prod_j \theta_{ij}^{X_1X_2} = 1$ for i fixed.

This model is sufficiently rich that each cell probability in the table is the same as the observed relative frequency. For the Australia survey data, the associated estimates are $\hat{\mu} = 0.2443$, $\hat{\theta}_1^{X_1} = 1.0959$, $\hat{\theta}_2^{X_1} = 0.9125$, $\hat{\theta}_1^{X_2} = 1.1928$, $\hat{\theta}_2^{X_2} = 0.8384$, $\hat{\theta}_{11}^{X_1X_2} = \hat{\theta}_{22}^{X_1X_2} = 0.9038$, and $\hat{\theta}_{12}^{X_1X_2} = \hat{\theta}_{21}^{X_1X_2} = 1.1064$. Since the model

model	G^2	df	p-value
Constant	38.26	3	0.000
$(1, X_1)$	33.38	2	0.000
$(1, X_2)$	13.65	2	0.001
$(1, X_1, X_2)$	8.77	1	0.003
Saturated	0	0	–

Table 2.1: Goodness-of-fit tests for categorical data models relating to the Australia survey data.

includes the constant term, the X_1 and X_2 effect terms and interaction effect between X_1 and X_2, it is represented by the model symbol $(1,\ X_1,\ X_2,\ X_1X_2)$.

To test goodness of fit of the above models, for the Australia survey data, the likelihood-ratio statistic G^2 and p-value are shown in Table 2.1. For two-way tables, the likelihood-ratio statistic is calculated by

$$G^2 = 2\sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{\hat{\pi}_{ij}}\right)$$

where $\{p_{ij}\}$ is the observed table and $\{\hat{\pi}_{ij}\}$ the associated fitted table for a given model. When the model is suitable, G^2 has large-sample chi-squared distribution with degrees of freedom equalling the difference between the number of cells and the number of parameters in the model. For given degree of freedom, larger G^2 values indicate smaller right-tail probabilities (p-values), and represent poorer fits. Table 2.1 indicates that none of the unsaturated models fit the data well, hence the attitude towards genetic engineering is not independent of the level of income.

We can summarize the dependence in the table using an odds ratio. For the low income group, the odds of attitude "For" are 1.16 which means there were 116 "For" responses for every 100 "Against" response. For the high income group, the odds of attitude "For" are 1.74 which means there were 174 "For" responses for every 100

"Against" response. Hence the table's odds ratio is 0.667 which indicates that the odds for the attitude "For" towards genetic engineering in the low income group is 0.667 times the odds in the high group.

In general, suppose an m-way observed relative frequency table $\{p_i\}$ where $i = (i_1, i_2, \ldots, i_m)$ has true probability table $\{\pi_i\}$ with categorical variables $X_1, X_2, \ldots, X_m$, which have $n_1, n_2, \ldots, n_m$ levels indexed respectively by $i_1, i_2, \ldots, i_m$. Then in the saturated model the joint probability distribution of the table has a multiplicative decomposition as

$$\pi_i = \prod_{A \subseteq T} \theta^A_{i_A}$$

Here the product is over all possible subsets A of $T = \{X_1, X_2, \ldots, X_m\}$, $\theta^A_{i_A}$ represents the interaction effect among variables in A and depends on i only through i_A where i_A is the corresponding sub m-tuple of i for A. Note that $\theta^A_{i_A} = \mu$ when $A = \varnothing$. To achieve identifiability the model is constrained by requiring that the product of the parameter $\theta^A_{i_A}$ for any index in i_A equals one. Thus, for categorical data models, we have a multiplicative decomposition of a probability π_i. Traditional linear models, however, are based on an additive decomposition of the observation mean. By using a log-transformation a familiar additive decomposition is constructed for $\log \pi_i$. This leads to the loglinear model, a standard tool for dealing with categorical data.

2.3 Loglinear models and sufficient statistics

In this section, we will discuss the form of loglinear models, and sufficient statistics for parameters in a loglinear model. These provide the foundations on which we construct the geometry of loglinear models.

After log-transformation, the saturated model for categorical variables $X_1, X_2, \ldots, X_m$ with cell index $i = (i_1, i_2, \ldots, i_m)$ has form

$$\log \pi_i = \sum_{A \subseteq S} \lambda^A_{i_A} \tag{2.1}$$

where $\lambda^A_{i_A} = \log \theta^A_{i_A}$ is the interaction effect among variables in A and depends on i only through i_A, the sub-tuple of i corresponding to A. Conventionally we write $\lambda^A_{i_A} = \mu$ when $A = \varnothing$. To achieve identifiability the model has the constraints that the sum of the parameter $\lambda^A_{i_A}$ for any index in i_A equals zero. Here we call form (2.1) the symbolic form of a loglinear model.

For example, the saturated loglinear model for a 2×2 table has symbolic form

$$\log \pi_{ij} = \mu + \lambda^{X_1}_i + \lambda^{X_2}_j + \lambda^{X_1X_2}_{ij} \quad \text{for} \quad i, j = 1, 2$$

constrained by $\sum_i \lambda^{X_1}_i = \sum_j \lambda^{X_2}_j = 0$, $\sum_i \lambda^{X_1X_2}_{ij} = 0$ for j fixed, and $\sum_j \lambda^{X_1X_2}_{ij} = 0$ for i fixed. We still denote this model using the model symbol $(1,\ X_1,\ X_2,\ X_1X_2)$.

However, considering those constraints the loglinear model has an alternative expression

$$\begin{bmatrix} \log \pi_{11} \\ \log \pi_{12} \\ \log \pi_{21} \\ \log \pi_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \lambda^{X_1}_1 \\ \lambda^{X_2}_1 \\ \lambda^{X_1X_2}_{11} \end{bmatrix} \tag{2.2}$$

Note that columns of the matrix in the model are contrasts in a 2×2 factorial design. These contrasts correspond to the effects in the model symbol $(1,\ X_1,\ X_2,\ X_1X_2)$ respectively.

In general, if we let the table $\{\pi_i\}$ where $i = (i_1, i_2, \ldots, i_m)$ have vector form $\pi = [\pi_1, \pi_2, \ldots, \pi_n]^T$ where $n = n_1 n_2 \ldots n_m$, then under the constraints that the sum

of the parameters for any index associated with the variables in i_A equals zero, the saturated loglinear model (2.1) has form

$$\log \pi = X\beta \tag{2.3}$$

where X is the design matrix with size $n \times n$ containing the constraints constructed using the full factorial design involving variables $X_1, X_2, \ldots, X_m$, and β is a column vector of parameters of size n. Hence, loglinear models have an analogous form to linear models, but parameters in β are not totally free as they are in linear models, since the parameters are constrained by $\sum \pi_i = 1$. The form (2.3) is called the matrix form of a loglinear model.

The discussion above is about saturated loglinear models, but results are easy to apply to unsaturated loglinear models by eliminating some terms (or columns of the design matrix) in the saturated model. The design matrix X then has size $n \times q$ $(q \leq n)$ where q is the size of the parameter vector β. For instance, the independence model of a 2×2 table with variables X_1 and X_2 requires the absence of the interaction between X_1 and X_2, so the model has symbolic form

$$\log \pi_{ij} = \mu + \lambda_i^{X_1} + \lambda_j^{X_2} \quad \text{for} \quad i, j = 1, 2$$

with constraints $\sum_i \lambda_i^{X_1} = \sum_j \lambda_j^{X_2} = 0$, denoted by a model symbol $(1,\ X_1,\ X_2)$. Correspondingly, the model also has matrix form

$$\begin{bmatrix} \log \pi_{11} \\ \log \pi_{12} \\ \log \pi_{21} \\ \log \pi_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \lambda_1^{X_1} \\ \lambda_1^{X_2} \end{bmatrix} \tag{2.4}$$

Sufficient statistics

A sufficient statistic for a parameter θ is a statistic that contains all the information about θ in the sample. Thus any inference about θ depends on the sample only through the sufficient statistic. The sufficient statistic provides a form of data reduction or data summary for the parameter θ. A sufficient statistic is formally defined as follows.

Definition 2.3.1. A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

For example, if a population follows a normal distribution with known variance, then the sample mean is a sufficient statistic for the population mean. Note that a sufficient statistic for a parameter may not be unique; any one-to-one function of a sufficient statistic is also a sufficient statistic.

A loglinear model can be represented in symbolic form or matrix form. Correspondingly, we have two ways to find sufficient statistics for parameters in the model. When a loglinear model is represented in the symbolic form, Bishop, Fienberg and Holland (1995) showed that for the Poisson or multinomial sampling scheme, sufficient statistics for the parameter λ are simply the marginal tables corresponding to the terms in the model symbol.

Recall that for a 2×2 observed relative frequency table $\{p_{ij}\}$, the saturated model has model symbol $(1, X_1, X_2, X_1X_2)$, so we have sufficient statistics $\{p_{i+}\}$ $\{p_{+j}\}$ and $\{p_{ij}\}$ (where '$+$' denotes the summation over the associated index) for parameters $\lambda_i^{X_1}$, $\lambda_j^{X_2}$ and $\lambda_{ij}^{X_1X_2}$ respectively. Thus there is no reduction of the data for the saturated model.

For the independence model with the model symbol $(1,\ X_1,\ X_2)$, the marginal tables $\{p_{i+}\}$ and $\{p_{+j}\}$ are sufficient statistics for parameters $\lambda_i^{X_1}$, and $\lambda_j^{X_2}$ respectively.

When a loglinear model is represented in the matrix form (2.3), Haberman (1973) showed that for an observed table $\{p_i\}$ where $i = (i_1, i_2, \ldots, i_m)$ with vector form $p = [p_1, p_2, \ldots, p_n]^T$ where $n = n_1 n_2 \ldots n_m$, we have a sufficient statistic vector $X^T p$ for the parameter vector β. Since column vectors in X are factorial contrasts, the sufficient statistics in $X^T p$ are some marginal tables. Note that for each parameter in β, the sufficient statistic is the associated component in $X^T p$.

From the matrix form (2.2), the saturated model of a 2×2 table has sufficient statistic vector

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}^T \begin{bmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ p_{1+} - p_{2+} \\ p_{+1} - p_{+2} \\ p_{11} - p_{12} - p_{21} + p_{22} \end{bmatrix}$$

for parameter vector $[\mu, \lambda_1^{X_1}, \lambda_1^{X_2}, \lambda_{11}^{X_1 X_2}]^T$. Then sufficient statistics for the parameters $\lambda_1^{X_1}$, $\lambda_1^{X_2}$, and $\lambda_{11}^{X_1 X_2}$ are $p_{1+} - p_{2+}$, $p_{+1} - p_{+2}$, and $p_{11} - p_{12} - p_{21} + p_{22}$ respectively. We know that the parameters in the model, however, have sufficient statistics $\{p_{i+}\}$ $\{p_{+j}\}$ and $\{p_{ij}\}$ from the symbolic form of the model. There is no contradiction here, because $p_{1+} - p_{2+}$, $p_{+1} - p_{+2}$, and $p_{11} - p_{12} - p_{21} + p_{22}$ are one-to-one functions of $\{p_{i+}\}$ $\{p_{+j}\}$ and $\{p_{ij}\}$ respectively.

The independence model of a 2×2 table with matrix form (2.4) has sufficient

statistic vector

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}^T \begin{bmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ p_{1+} - p_{2+} \\ p_{+1} - p_{+2} \end{bmatrix}$$

for the parameter vector $[\mu, \lambda_1^{X_1}, \lambda_1^{X_2}]^T$. Again $p_{1+} - p_{2+}$ and $p_{+1} - p_{+2}$ are one-to-one functions of $\{p_{i+}\}$ and $\{p_{+j}\}$ respectively.

Once a set of sufficient statistics is determined, Birch (1963) showed that the likelihood equations for loglinear models match sufficient statistics to their expected values. Specifically, suppose an observed table $\{p_i\}$ with vector form $p = [p_1, p_2, \ldots, p_n]^T$ has a maximum likelihood estimate $\{\hat{\pi}_i\}$ with vector form $\hat{\pi} = [\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n]^T$ for a loglinear model $\log \pi = X\beta$. Then Birch's result determines that

$$X^T p = X^T \hat{\pi}$$

Note that this equation plays an important role in the geometry of loglinear models. It will be interpreted geometrically in the next section.

To summarize, loglinear models provide an additive decomposition of log-transformed cell probabilities. A loglinear model has a symbolic form and a matrix form, and correspondingly sufficient statistics for parameters in the model are determined by the associated model symbol or the design matrix and the observation vector. According to Birch's result, the sufficient statistics for parameters in a loglinear model will be preserved in the fitting process.

2.4 Geometry of loglinear models

Since cell probabilities π are transformed by a logarithm for loglinear models, the geometry of loglinear models has been described in two distinct ways. Fienberg (1968, 1970) described the untransformed π using a simplex, while Haberman (1974) represented the transformed $\log \pi$ using a subset in Euclidean space.

In Fienberg geometry (Fienberg, 1968), all possible $r \times c$ probability tables correspond to the points within an $(rc-1)$-dimensional simplex in $\mathbf{R}^n$ where $n = rc$. Then the loci of three types of two-way table are described by Fienberg in the simplex:

(a) all points corresponding to tables whose rows and columns are independent,

(b) all points corresponding to tables with a given interaction structure,

(c) all points corresponding to a table with a fixed set of margins.

All results are illustrated explicitly by 2×2 tables using a three dimensional simplex (a tetrahedron). For example, the model space of the saturated model of a 2×2 table is the whole tetrahedron in $\mathbf{R}^4$, while the model space of the independence model is a portion of a hyperbolic paraboloid in the tetrahedron (see Figure 2.5).

On the other hand, Haberman (1974) viewed a log-transformed probability table with n cells as a vector in Euclidean space $\mathbf{R}^n$ and the model space of a loglinear model as a subset in $\mathbf{R}^n$. Fitting a loglinear model maps the observation vector to a q-dimensional ($q \leq n$) model space contained in $\mathbf{R}^n$ (where q is the number of parameters in the loglinear model). Thus the whole space $\mathbf{R}^n$ is partitioned into a q-dimensional model space and its orthogonal complement. For $r \times c$ tables, all possible log-transformed probability tables $\{\log \pi_{ij}\}$ form a subset of rc-dimensional Euclidean space. This is the model space of the saturated model. Unsaturated models

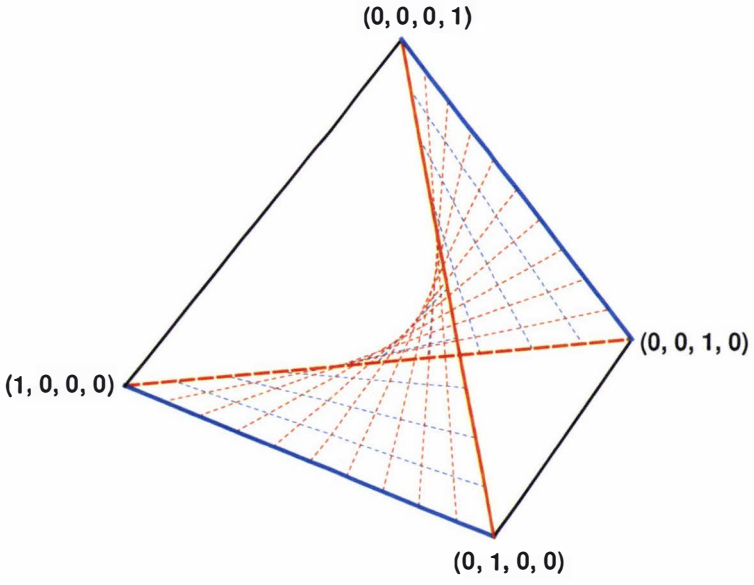


Figure 2.5: For a 2×2 contingency table, the saturated model space is a tetrahedron in $\mathbf{R}^4$, while the model space of the independence model is a portion of a hyperbolic paraboloid in the tetrahedron.

constrain $\{\log \pi_{ij}\}$ to a t-dimensional linear manifold in that subset, with $t < rc$. For the independence model, $t = r + c - 1$.

Shortly we will relate Fienberg and Haberman geometries as we construct a geometric framework for loglinear models; here probability tables will be used in the discussion. The results about probability tables are easily applied to frequency tables. We begin with the simplest case, a 2×2 table, and then illustrate the general result using examples in a $2 \times 2 \times 2$ table.

2.4.1 Geometry of a 2×2 table

In this section, the geometry of a 2×2 table is discussed for the saturated model and the independence model. In the saturated model, a new basis is constructed for

the associated Euclidean space, motivated by sufficient statistics for parameters in the saturated model. In the independence model, the associated Euclidean space is partitioned into two orthogonal parts, each part spanned by a subset of the new basis.

The saturated model

The saturated model of a 2×2 table will be presented in four stages: a new basis in $\mathbf{R}^n$, Fienberg geometry, Haberman geometry and the link between the two geometries.

A new basis in $\mathbf{R}^n$

A 2×2 probability table $\{\pi_{ij}\}$ with variables X_1 and X_2 corresponds to a point (or vector) in $\mathbf{R}^4$. Fienberg (1970) considered the point (or vector) $\pi = [\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}]^T$ with respect to the standard basis, so the saturated model for the untransformed joint distribution can be expressed as

$$\begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \pi_{11} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \pi_{12} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \pi_{21} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \pi_{22} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

with $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$. This distribution can thus be thought of as a point in a regular tetrahedron (a 3-dimensional simplex) with vertices

$$e_1 = [1, 0, 0, 0]^T, \; e_2 = [0, 1, 0, 0]^T, \; e_3 = [0, 0, 1, 0]^T, \text{ and } e_4 = [0, 0, 0, 1]^T$$

the standard basis in $\mathbf{R}^4$. This saturated model space is shown in Figure 2.6. Thus, a 2×2 observed relative frequency table $\{p_{ij}\}$ can be represented by a vector $p = [p_{11}, p_{12}, p_{21}, p_{22}]^T$ in the tetrahedron.

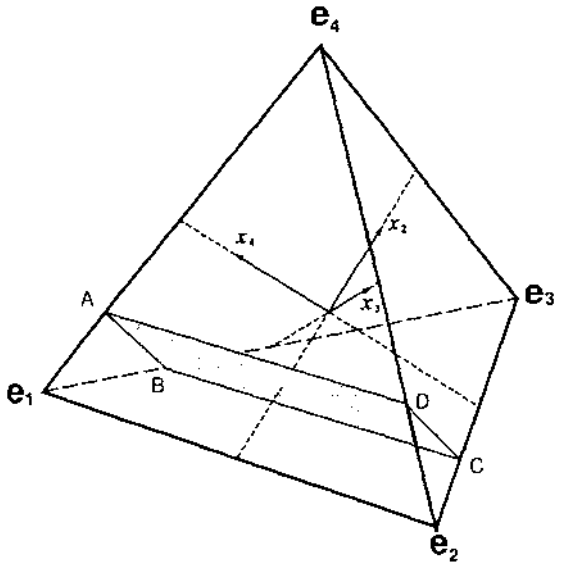


Figure 2.6: The saturated model space for a 2×2 contingency table is a tetrahedron in $\mathbf{R}^4$. Orthogonal vectors x_2, x_3 and x_4 form a new basis for the tetrahedron, while the shaded quadrilateral $ABCD$ represents joint distributions with fixed X_1 margin.

From the last section we know that the sufficient statistics vector for the parameters vector $\beta = [\mu, \lambda_1^{X_1}, \lambda_1^{X_2}, \lambda_{11}^{X_1 X_2}]^T$ in the saturated loglinear model (2.2) is the vector

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}^T \begin{bmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \end{bmatrix} = \begin{bmatrix} p.x_1 \\ p.x_2 \\ p.x_3 \\ p.x_4 \end{bmatrix}$$

where x_1, x_2, x_3 and x_4 are the column vectors in the design matrix (see (2.2)). Specifically

$$x_1 = [1,1,1,1]^T,\ x_2 = [1,1,-1,-1]^T,\ x_3 = [1,-1,1,-1]^T,\ x_4 = [1,-1,-1,1]^T$$

Since $p.x_i$ is a sufficient statistic for the ith element of β for all i, then $p.x_i/\|x_i\|$

also is a sufficient statistic for the ith element of β due to the one-to-one relationship between $p.x_i$ and $p.x_i/\|x_i\|$ for all i. Thus, in Euclidean space $\mathbf{R}^n$, the sufficient statistics for β in the model are the lengths (ignoring the sign) of projection of the observation vector p onto the directions specified by the column vectors of the design matrix. Furthermore, the column vectors x_1, x_2, x_3 and x_4 are linearly independent, so $\{x_1, x_2, x_3, x_4\}$ (the vectors x_2, x_3 and x_4 are illustrated in Figure 2.6) is chosen as a new basis in $\mathbf{R}^4$ motivated by sufficient statistics. Now, projecting the vector p onto the new basis, we obtain a coordinate vector for p with respect to the new basis

$$\left[\frac{1}{4}, \frac{p_{1+} - p_{2+}}{4}, \frac{p_{+1} - p_{+2}}{4}, \frac{p_{11} - p_{12} - p_{21} + p_{22}}{4}\right]^T$$

which are sufficient statistics for parameters in the saturated model.

Finally, note that coordinates with respect to the new basis are the image of coordinates with respect to the standard basis in $\mathbf{R}^4$ under the linear transformation

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}^{-1}$$

which is the inverse design matrix in the saturated model.

Fienberg geometry – the saturated model

Fienberg geometry provides a description of the geometry of an untransformed table. We illustrate this initially using a two-way table $\{\pi_{ij}\}$ with vector form $\pi = [\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}]^T$. The saturated model can be expressed as an additive decomposition of the joint probability π with respect to the new basis

$$\pi = \frac{\pi.x_1}{\|x_1\|^2}x_1 + \frac{\pi.x_2}{\|x_2\|^2}x_2 + \frac{\pi.x_3}{\|x_3\|^2}x_3 + \frac{\pi.x_4}{\|x_4\|^2}x_4$$

Specifically, this is

$$\begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{\pi_{1+} - \pi_{2+}}{4}\begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} + \frac{\pi_{+1} - \pi_{+2}}{4}\begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} + \frac{\pi_{11} - \pi_{12} - \pi_{21} + \pi_{22}}{4}\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \tag{2.5}$$

Parallel to the linear model, we define the "effect" of a model term to be the projection coefficient onto the associated contrast vector in the new basis. Hence the constant term is $1/4$, the projection coefficient for the equiangular vector x_1. The main effect of X_1 is $(\pi_{1+} - \pi_{2+})/4$, the projection coefficient for x_2 and similarly the main effect of X_2 is $(\pi_{+1} - \pi_{+2})/4$, the projection coefficient for x_3. Finally the interaction effect of X_1X_2 is $(\pi_{11} - \pi_{12} - \pi_{21} + \pi_{22})/4$, the projection coefficient for x_4.

With respect to the new basis, the entire tetrahedron lies in the hyperplane whose value along x_1 is $1/4$. Evidently all tables with a given X_1 margin will lie in a hyperplane orthogonal to x_2, while all tables with a given X_2 margin will lie in a hyperplane orthogonal to x_3, since the coordinates of x_2 and x_3 are $(\pi_{1+} - \pi_{2+})/4$ and $(\pi_{+1} - \pi_{+2})/4$ respectively. A slice $ABCD$ of the first type is illustrated in Figure 2.6.

Haberman geometry – the saturated model

Haberman geometry provides a description of the geometry of a log-transformed table. We illustrate this initially using a two-way table $\{\log \pi_{ij}\}$ with vector form

$\log \pi = [\log \pi_{11}, \log \pi_{12}, \log \pi_{21}, \log \pi_{22}]^T$. Expressed with respect to the new basis $\{x_1, x_2, x_3, x_4\}$, the saturated model for $\log \pi$ is

$$\log \pi = \frac{\log \pi . x_1}{\|x_1\|^2} x_1 + \frac{\log \pi . x_2}{\|x_2\|^2} x_2 + \frac{\log \pi . x_3}{\|x_3\|^2} x_3 + \frac{\log \pi . x_4}{\|x_4\|^2} x_4$$

In vector form this becomes

$$\begin{bmatrix} \log \pi_{11} \\ \log \pi_{12} \\ \log \pi_{21} \\ \log \pi_{22} \end{bmatrix} = \frac{1}{4} \log(\pi_{11}\pi_{12}\pi_{21}\pi_{22}) \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{4} \log \left(\frac{\pi_{11}\pi_{12}}{\pi_{21}\pi_{22}}\right) \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

$$+ \frac{1}{4} \log \left(\frac{\pi_{11}\pi_{21}}{\pi_{12}\pi_{22}}\right) \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} + \frac{1}{4} \log \left(\frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}\right) \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \tag{2.6}$$

Again, a log-transformed 2×2 observed relative frequency table $\{\log p_{ij}\}$ can be represented by a vector $\log p = [\log p_{11}, \log p_{12}, \log p_{21}, \log p_{22}]^T$ in the extended tetrahedron with vertices lying at infinity (discussed in the next section). Then the coordinate vector of $\log p$ with respect to the new basis is

$$\left[\frac{1}{4} \log(p_{11}p_{12}p_{21}p_{22}),\ \frac{1}{4} \log \left(\frac{p_{11}p_{12}}{p_{21}p_{22}}\right),\ \frac{1}{4} \log \left(\frac{p_{11}p_{21}}{p_{12}p_{22}}\right),\ \frac{1}{4} \log \left(\frac{p_{11}p_{22}}{p_{21}p_{12}}\right)\right]^T$$

The link between the two geometries – the saturated model

In order to link the Fienberg and Haberman geometries, it is necessary to study them with respect to the same basis. Here, we consider the two geometries with respect to the standard basis. In Fienberg geometry, the saturated model space for a 2×2 probability table is a tetrahedron with vertices

$$e_1 = [1, 0, 0, 0]^T,\ e_2 = [0, 1, 0, 0]^T,\ e_3 = [0, 0, 1, 0]^T,\ e_4 = [0, 0, 0, 1]^T$$

The componentwise logarithm transformation maps the tetrahedron into an extended tetrahedron with vertices

$$e_1 = [0, -\infty, -\infty, -\infty]^T,\ e_2 = [-\infty, 0, -\infty, -\infty]^T$$

$$e_3 = [-\infty, -\infty, 0, -\infty]^T,\ e_4 = [-\infty, -\infty, -\infty, 0]^T$$

in the negative orthant in extended $\mathbf{R}^4$. The extended tetrahedron is the saturated model space for a 2×2 probability table in Haberman geometry. Hence, with respect to the standard basis, for the saturated model of a 2×2 table, the regular tetrahedron of Fienberg geometry is mapped to the extended tetrahedron of Haberman geometry, with vertices at the limits of diagonals on the planar faces of the negative orthants, as indicated schematically in Figure 2.7.

The independence model

Now we follow the same pattern as used in the discussion of the saturated model to display the geometry of the independence model for a 2×2 probability table $\{\pi_{ij}\}$ with variables X_1 and X_2. Here the whole space $\mathbf{R}^4$ will be partitioned into two parts to reveal geometric properties.

Fienberg geometry – the independence model

In Fienberg geometry, imposition of independence of X_1 and X_2 will restrict π to a subset of the tetrahedron. Independence does not constrain the one-way margins, so on the new basis the coordinates of x_1, x_2 and x_3 will not alter. Since π_{ij} must now equal $\pi_{i+}\pi_{+j}$ for all i and j, in (2.5) the coefficient $(\pi_{11} - \pi_{12} - \pi_{21} + \pi_{22})/4$ of x_4 can be checked to be $(\pi_{1+} - \pi_{2+})(\pi_{+1} - \pi_{+2})/4$, so the independence model space is the

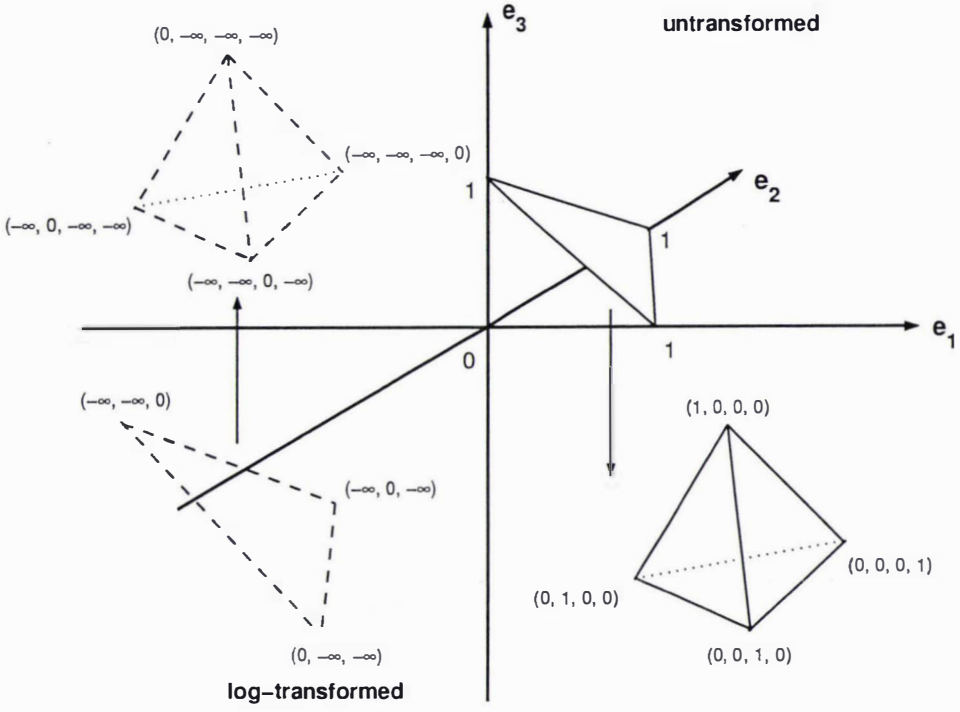


Figure 2.7: The saturated model space for a 2×2 contingency table is a tetrahedron with vertices the standard basis in $\mathbf{R}^4$, while after log-transformation the saturated model becomes an extended tetrahedron in the negative orthant in extended $\mathbf{R}^4$.

subset of the tetrahedron with points having form

$$\begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{\pi_{1+} - \pi_{2+}}{4}\begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} + \frac{\pi_{+1} - \pi_{+2}}{4}\begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$$
$$+ \frac{(\pi_{1+} - \pi_{2+})(\pi_{+1} - \pi_{+2})}{4}\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \tag{2.7}$$

This is shaded in Figure 2.8, a portion of a hyperbolic paraboloid in the tetrahedron, which we informally term the "butterfly".

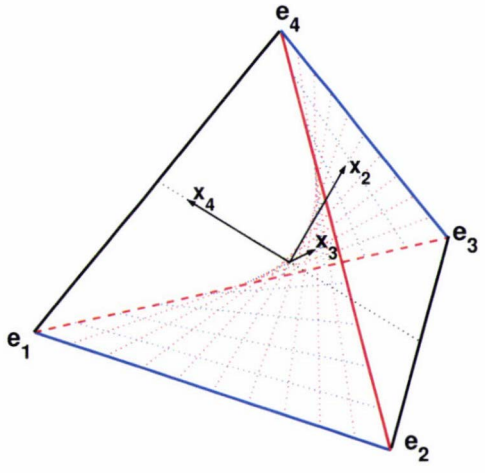


Figure 2.8: The Fienberg independence model space (double-ruled) for a 2×2 contingency table is a two-dimensional surface in the tetrahedral saturated model space.

Hence for an observed table $\{p_{ij}\}$ (corresponding to a vector p on the tetrahedron), the maximum likelihood fitted table $\{\hat{\pi}_{ij}\}$ (corresponding to a vector $\hat{\pi}$ on the butterfly) for the model $X_1 \perp\!\!\!\perp X_2$ has coordinate vector

$$\left[\frac{1}{4}, \frac{\hat{\pi}_{1+} - \hat{\pi}_{2+}}{4}, \frac{\hat{\pi}_{+1} - \hat{\pi}_{+2}}{4}, \frac{(\hat{\pi}_{1+} - \hat{\pi}_{2+})(\hat{\pi}_{+1} - \hat{\pi}_{+2})}{4}\right]^T$$

with respect to the new basis $\{x_1, x_2, x_3, x_4\}$.

Recall that maximum likelihood model fitting, with either the Poisson or multinomial distributional assumption, preserves the sufficient statistics for the model (Birch 1963). Thus we have

$$\hat{\pi}_{i+} = p_{i+}, \qquad \hat{\pi}_{+i} = p_{+i} \qquad \text{for } i = 1, 2$$

Now the coordinate vector of $\hat{\pi}$ with respect to the new basis becomes

$$\left[\frac{1}{4}, \frac{p_{1+}-p_{2+}}{4}, \frac{p_{+1}-p_{+2}}{4}, \frac{(p_{1+}-p_{2+})(p_{+1}-p_{+2})}{4}\right]^T$$

Recall that the vector p with respect to the new basis has coordinate vector

$$\left[\frac{1}{4}, \frac{p_{1+}-p_{2+}}{4}, \frac{p_{+1}-p_{+2}}{4}, \frac{p_{11}-p_{12}-p_{21}+p_{22}}{4}\right]^T$$

Hence the coordinates of x_1, x_2 and x_3 are the same for vectors $\hat{\pi}$ and p, and these coordinates are sufficient statistics for parameters in the model $X_1 \perp\!\!\!\perp X_2$. This property will be central in later chapters.

Haberman geometry – the independence model

In Haberman geometry, under the independence assumption $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j, from (2.6), the coordinates of $\log \pi$ with respect to the new basis can be shown to be

$$\begin{aligned}
\frac{\log \pi . x_1}{\|x_1\|^2} &= \frac{1}{4}\log\left(\pi_{1+}\pi_{+1}\pi_{1+}\pi_{+2}\pi_{2+}\pi_{+1}\pi_{2+}\pi_{+2}\right) \\
&= \frac{1}{2}\log\left(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2}\right) \\
\frac{\log \pi . x_2}{\|x_2\|^2} &= \frac{1}{4}\log\left(\frac{\pi_{1+}\pi_{+1}\pi_{1+}\pi_{+2}}{\pi_{2+}\pi_{+1}\pi_{2+}\pi_{+2}}\right) \\
&= \frac{1}{2}\log\left(\frac{\pi_{1+}}{\pi_{2+}}\right) \\
\frac{\log \pi . x_3}{\|x_3\|^2} &= \frac{1}{4}\log\left(\frac{\pi_{1+}\pi_{+1}\pi_{2+}\pi_{+1}}{\pi_{1+}\pi_{+2}\pi_{2+}\pi_{+2}}\right) \\
&= \frac{1}{2}\log\left(\frac{\pi_{+1}}{\pi_{+2}}\right) \\
\frac{\log \pi . x_4}{\|x_4\|^2} &= \frac{1}{4}\log\left(\frac{\pi_{1+}\pi_{+1}\pi_{2+}\pi_{+2}}{\pi_{2+}\pi_{+1}\pi_{1+}\pi_{+2}}\right) \\
&= 0
\end{aligned}$$

Thus after the log-transformation the independence model space becomes

$$\begin{bmatrix} \log \pi_{11} \\ \log \pi_{12} \\ \log \pi_{21} \\ \log \pi_{22} \end{bmatrix} = \frac{1}{2} \log \left(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2} \right) \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{2} \log \left(\frac{\pi_{1+}}{\pi_{2+}} \right) \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} + \frac{1}{2} \log \left(\frac{\pi_{+1}}{\pi_{+2}} \right) \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$$

This independence surface is pictured schematically in Figure 2.9, and is informally termed the "jellyfish". It is a locally two-dimensional surface in the hyperplane orthogonal to x_4, but is necessarily pictured here in $\mathbf{R}^3$.

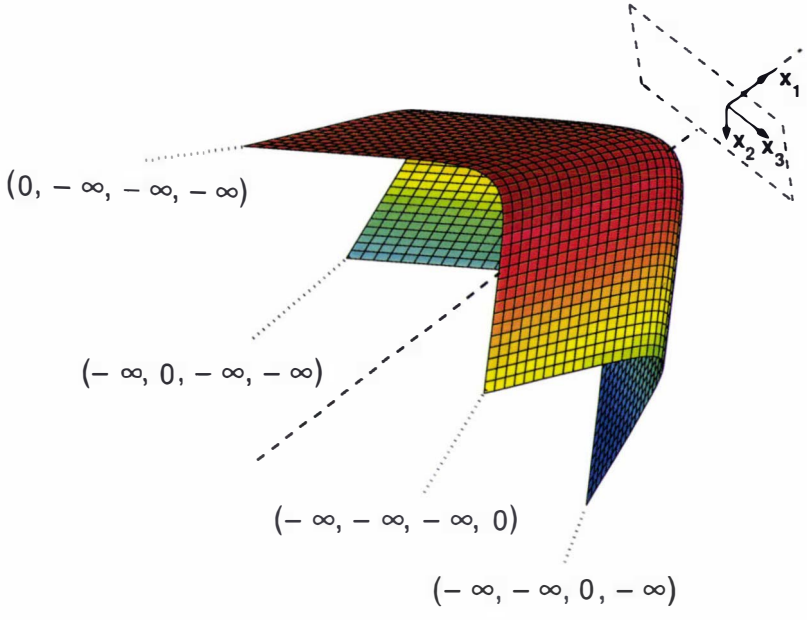


Figure 2.9: A schematic representation of the Haberman view of the independence model space. It lies entirely in a 3-dimensional subspace in the negative orthant of $\mathbf{R}^4$ and has its apex at $(\log \frac{1}{4}, \log \frac{1}{4}, \log \frac{1}{4}, \log \frac{1}{4})$ with respect to the standard basis.

Consider a log-transformed observation table $\{\log p_{ij}\}$ (corresponding to a vector $\log p$ in the extended tetrahedron) and the maximum likelihood fitted table $\{\log \hat{\pi}_{ij}\}$ (corresponding to a vector $\log \hat{\pi}$ on the jellyfish) for the model $X_1 \perp\!\!\!\perp X_2$. After

applying Birch's results these two points have coordinate vectors with respect to the new basis of

$$\log p: \quad \left[\frac{1}{2}\log(p_{11}p_{12}p_{21}p_{22}),\ \frac{1}{2}\log\left(\frac{p_{11}p_{12}}{p_{21}p_{22}}\right),\ \frac{1}{2}\log\left(\frac{p_{11}p_{21}}{p_{12}p_{22}}\right),\ \frac{1}{2}\log\left(\frac{p_{11}p_{22}}{p_{21}p_{12}}\right)\right]^T$$

$$\log\hat{\pi}: \quad \left[\log\left(p_{1+}p_{2+}p_{+1}p_{+2}\right),\quad \log\left(\frac{p_{1+}}{p_{2+}}\right),\quad \log\left(\frac{p_{+1}}{p_{+2}}\right),\quad 0\right]^T$$

We find that the coordinates relative to x_1, x_2 and x_3 are not preserved in the fitting process, but for $\log\hat{\pi}$ the coordinate relative to x_4 is zeroed. This property also will be central in later chapters.

The link between the two geometries – the independence model

To combine the two geometries, for the unsaturated model $X_1 \perp\!\!\!\perp X_2$ with model symbol $(1,\ X_1,\ X_2)$, we split the new basis into two parts, one including elements corresponding to columns of the design matrix (i.e. $\{x_1, x_2, x_3\}$), the other being the remaining element $\{x_4\}$. Then the whole space $\mathbf{R}^4$ can be partitioned into two subspaces as

$$\mathbf{R}^4 = \mathcal{S} \oplus \mathcal{A}$$

where $\mathcal{S} = \text{span}\{x_1, x_2, x_3\}$ and $\mathcal{A} = \mathcal{S}^{\perp} = \text{span}\{x_4\}$.

In Fienberg geometry, the coordinates of the observation p with respect to the basis of subspace $\mathcal{S}$ are sufficient statistics and will be preserved in the fitting process, while in Haberman geometry the coordinate of the fitted vector $\log\hat{\pi}$ with respect to the basis of subspace $\mathcal{A}$ is zeroed in the fitting process. Thus we call $\mathcal{S}$ the "sufficiency space" and $\mathcal{A}$ the "auxiliary space". Furthermore, in Fienberg geometry, the model space of $X_1 \perp\!\!\!\perp X_2$, the "butterfly", straddles the sufficiency space and auxiliary space. In Haberman geometry, the model space of $X_1 \perp\!\!\!\perp X_2$, the "jellyfish", lies entirely in the sufficiency space.

With respect to the standard basis, for the unsaturated model $X_1 \perp\!\!\!\perp X_2$ of a 2×2 table, the "butterfly" in Fienberg geometry is mapped to the "jellyfish" in Haberman geometry using the logarithmic link function. Note that the componentwise logarithm transformation is applied with respect to the standard basis. Figure 2.10 links the two geometric approaches schematically. The independence model space for Fienberg geometry is in the bounded tetrahedron, while for Haberman geometry it is in the negative orthant. We move from one to the other via the logarithmic link function.

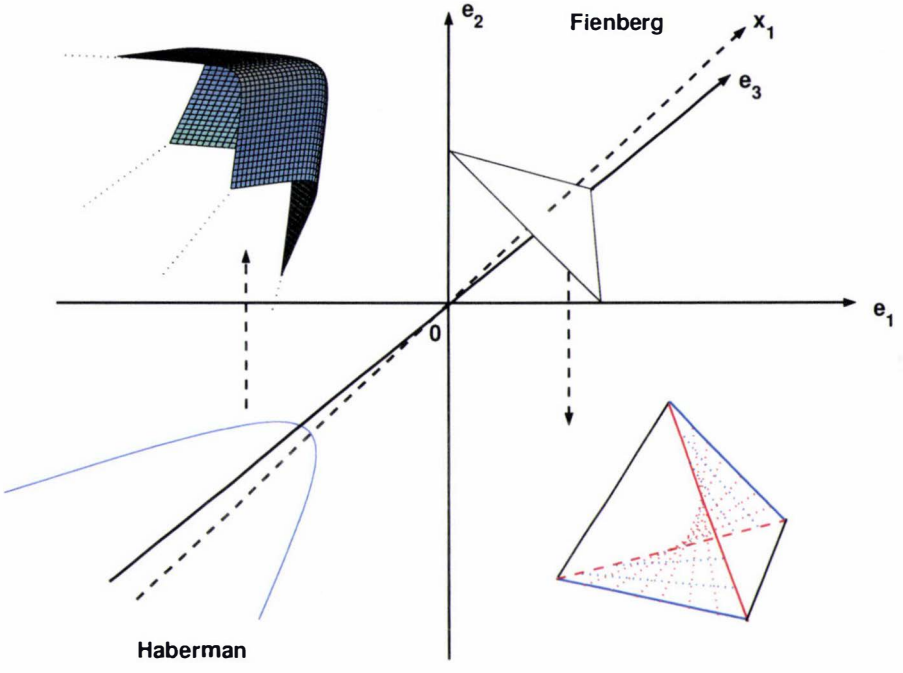


Figure 2.10: The Fienberg and Haberman geometries for the independence model, the former in the bounded tetrahedron and the latter in the negative orthant. We move from one to the other via the logarithmic link function.

2.4.2 The general case

In general, we consider the geometry of an m-way table $\{\pi_i\}$ where $i = (i_1, i_2, \ldots, i_m)$ with categorical variables $X_1, X_2, \ldots, X_m$, which have $n_1, n_2, \ldots, n_m$ levels indexed by $i_1, i_2, \ldots, i_m$ respectively. We denote the joint probability mass function by a column vector $\pi = [\pi_1, \pi_2, \ldots, \pi_n]^T$ where $n = n_1 n_2 \ldots n_m$. As with the geometry of the 2×2 table, we first construct the new basis in $\mathbf{R}^n$ motivated by the sufficient statistics for parameters in the saturated model. Then we discuss Fienberg and Haberman geometries for an saturated model and the unsaturated model respectively. Finally, the relationship between the two geometries is summarized in a commutative diagram, the core of this chapter.

Sufficient statistics and the new basis

For an observed relative frequency table $\{p_i\}$ with vector form $p = [p_1, p_2, \ldots, p_n]^T$, the sufficient statistics vector for the parameters vector in the saturated model $\log \pi = X\beta$ is obtained by projecting p onto the column vectors of X. Specifically, the sufficient statistics vector is

$$X^T p = \begin{bmatrix} p.x_1 \\ p.x_2 \\ \vdots \\ p.x_n \end{bmatrix}$$

where X is the design matrix of size $n \times n$, and $x_1, x_2, \ldots, x_n$ are the column vectors of the design matrix. Hence, if we denote the parameter vector β as $[\beta_1, \beta_2, \ldots, \beta_n]^T$, a sufficient statistic for β_i is $p.x_i$ and thus also $p.x_i/\|x_i\|$, the length of the projection of p onto x_i for $i = 1, 2, \ldots, n$. Motivated by the sufficient statistics and the linear independence of $x_1, x_2, \ldots, x_n$, we select $\{x_1, x_2, \ldots, x_n\}$ as a new basis in $\mathbf{R}^n$.

The saturated model

With respect to the standard basis, Fienberg (1968) pointed out that all possible joint probability mass functions, denoted by a column vectors $\pi = [\pi_1, \pi_2, \ldots, \pi_n]^T$ where $n = n_1 n_2 \ldots n_m$, can be represented by points within the $(n-1)$-dimensional simplex

$$S_{n-1} = \{(p_1, p_2, \ldots, p_n) \mid \sum_{i=1}^{n} p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i\} \subseteq \mathbf{R}^n$$

In Haberman geometry, however, all possible log-transformed joint probability mass functions $\{\log \pi_i\}$, with vector form $\log \pi$, can be represented by points within an extended simplex

$$\{(\log p_1, \log p_2, \ldots, \log p_n) \mid \sum_{i=1}^{n} p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i\}$$

Thus, on the standard basis, the saturated model spaces for the two geometries are linked by the componentwise logarithm transformation.

With respect to the new basis $\{x_1, x_2, \ldots, x_n\}$, in Fienberg geometry, π is decomposed as

$$\pi = \frac{\pi . x_1}{\|x_1\|^2} x_1 + \frac{\pi . x_2}{\|x_2\|^2} x_2 + \ldots + \frac{\pi . x_n}{\|x_n\|^2} x_n \tag{2.8}$$

while in Haberman geometry, $\log \pi$ is decomposed as

$$\log \pi = \frac{\log \pi . x_1}{\|x_1\|^2} x_1 + \frac{\log \pi . x_2}{\|x_2\|^2} x_2 + \ldots + \frac{\log \pi . x_n}{\|x_n\|^2} x_n \tag{2.9}$$

Thus in the saturated model, for the two geometries, the relationship between coordinates with respect to the new basis is clearly shown in Expressions (2.8) and (2.9).

The unsaturated model

For an unsaturated model $\log \pi = X\beta$ where X is the design matrix of size $n \times q$ $(q < n)$ and β the parameter vector of size $q \times 1$. Note that the column vectors in the design matrix for the unsaturated model are just a subset of the column vectors in the design matrix for the saturated model. Now the whole space $\mathbf{R}^n$ can be divided into a sufficiency space and an auxiliary space, specifically

$$\mathbf{R}^n = \mathcal{S} \ominus \mathcal{A}$$

where $\mathcal{S} = \text{span}\{x_1, x_2, \ldots, x_q\}$ and $\mathcal{A} = \mathcal{S}^{\perp} = \text{span}\{x_{q+1}, x_{q+2}, \ldots, x_n\}$. Here the new basis $\{x_1, x_2, \ldots, x_n\}$ is partitioned into the basis of the sufficiency space $\{x_1, x_2, \ldots, x_q\}$, the column vectors in the design matrix, and the basis of the auxiliary space $\{x_{q+1}, x_{q+2}, \ldots, x_n\}$, the remaining elements.

Now we denote an observed table and its maximum likelihood fit, under the unsaturated model, by column vectors p and $\hat{\pi}$ respectively. In Fienberg geometry, according to Birch's (1963) results, we have

$$X^T p = X^T \hat{\pi} \tag{2.10}$$

Geometrically, (2.10) indicates that the observation p and its fitted vector $\hat{\pi}$ have the same projection onto the basis of the sufficiency space, and these projections are the elements of sufficient statistics vector for β. In Haberman geometry, suppose that $\hat{\beta}$ is the estimated vector of β, then we have

$$\log \hat{\pi} = X\hat{\beta} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \ldots + \hat{\beta}_q x_q$$

Thus, $\log \hat{\pi}$ is a linear combination of the basis of the sufficiency space, in other words, $\log \hat{\pi} \in \mathcal{S}$, so

$$\log \hat{\pi} \cdot x_i = 0 \tag{2.11}$$

where $i = q+1, q+2, \ldots, n$. Geometrically, (2.11) indicates that the fitted vector $\log \hat{\pi}$ has zero projection onto the auxiliary space.

In summary, we have two critical properties:

1. In Fienberg geometry, the observation vector p and its fitted vector $\hat{\pi}$ have the same projection onto the sufficiency space, and this projection is sufficient for parameters in the model.

2. In Haberman geometry, the fitted vector $\log \hat{\pi}$ has zero projection onto the auxiliary space.

The relationship between the two geometries

From the above discussion, we find that the model space in Haberman geometry is the image of the model space in Fienberg geometry under the transformation of $\mathbf{R}^n$ which takes the logarithm of each coordinate with respect to the standard basis. Denoting the table with n cells and its log-transformation by column vectors π and $\log \pi$ respectively, the coordinate vector x^F of π with respect to the new basis in Fienberg geometry is related to the coordinate vector x^H of $\log \pi$ with respect to the same new basis in Haberman geometry as shown in the following commutative diagram:

$$\begin{array}{ccc} \pi & \xrightarrow{\quad A \quad} & x^F = A\pi \\ \Big\downarrow \log & & \Big\downarrow A \log A^{-1} \\ \log \pi & \xrightarrow{\quad A \quad} & x^H = A \log \pi \end{array}$$

where A maps coordinates with respect to the standard basis to coordinates with respect to the new basis.

When this commutative diagram is viewed vertically, the relationship between the two geometries is represented: the left side is with respect to the standard basis and the right side is with respect to the new basis. However, when viewed horizontally, the linkage between the two bases is revealed: the upper part is in Fienberg geometry and the lower part is in Haberman geometry.

For linear models, where the link function is the identity, the two geometries coalesce. This leads to fitting which combines the best of both worlds: the sufficient statistics preservation in the sufficiency space of Fienberg and the auxiliary space coefficient zeroing of Haberman. In Chapter 3 we will extend these ideas to generalized linear models, where in the commutative diagram the log link is replaced by the appropriate link function and A is determined by the design matrix.

2.4.3 Some examples

To illustrate the geometry of loglinear models, we study a three-way probability table $\{\pi_{ijk}\}$ with binary variables X_1, X_2 and X_3. The associated joint probability can be represented by a point (or vector) $\pi = [\pi_{111}, \pi_{112}, \pi_{121}, \pi_{122}, \pi_{211}, \pi_{212}, \pi_{221}, \pi_{222}]^T$ within a 7-dimensional simplex S_7 in $\mathbf{R}^8$. The geometry can be discussed with respect to the standard basis and a new basis. As with the geometry of a 2×2 table, a new basis $\{x_1, x_2, \ldots, x_8\}$ should be constructed in $\mathbf{R}^8$ using the column vectors of the design matrix in the saturated model. The design matrix is formed from the full

factorial contrasts involving three binary variables X_1, X_2, X_3, namely

$$X = \begin{array}{c} \begin{array}{cccccccc} 1 & X_1 & X_2 & X_1X_2 & X_3 & X_1X_3 & X_2X_3 & X_1X_2X_3 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \\ \begin{array}{cccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \end{array} \end{array} \tag{2.12}$$

where the annotations x_8 and $X_1X_2X_3$, for example, indicate that the corresponding column vector of X is element x_8 in the new basis and the associated contrast of the three way interaction of X_1, X_2 and X_3.

The saturated model

For the saturated model with respect to the new basis, in Fienberg geometry, the vector π can be written as the sum of projections onto the new basis elements

$$\pi = \frac{\pi . x_1}{\|x_1\|^2} x_1 + \frac{\pi . x_2}{\|x_2\|^2} x_2 + \ldots + \frac{\pi . x_8}{\|x_8\|^2} x_8$$

where

$$\begin{aligned} \frac{\pi . x_1}{\|x_1\|^2} &= \frac{1}{8} \\ \frac{\pi . x_2}{\|x_2\|^2} &= \frac{\pi_{1++} - \pi_{2++}}{8} \\ \frac{\pi . x_3}{\|x_3\|^2} &= \frac{\pi_{+1+} - \pi_{+2+}}{8} \end{aligned}$$

$$
\begin{aligned}
\frac{\pi . x_4}{\|x_4\|^2} &= \frac{\pi_{11+} - \pi_{12+} - \pi_{21+} + \pi_{22+}}{8} \\
\frac{\pi . x_5}{\|x_5\|^2} &= \frac{\pi_{+-1} - \pi_{++2}}{8} \\
\frac{\pi . x_6}{\|x_6\|^2} &= \frac{\pi_{1+1} - \pi_{1+2} - \pi_{2+1} + \pi_{2+2}}{8} \\
\frac{\pi . x_7}{\|x_7\|^2} &= \frac{\pi_{+11} - \pi_{+12} - \pi_{+21} + \pi_{+22}}{8} \\
\frac{\pi . x_8}{\|x_8\|^2} &= \frac{\pi_{111} - \pi_{112} - \pi_{121} + \pi_{122} - \pi_{211} + \pi_{212} + \pi_{221} - \pi_{222}}{8}
\end{aligned} \tag{2.13}
$$

For an observed relative frequency table $\{p_{ijk}\}$ with a corresponding vector p on the simplex S_7 in $\mathbf{R}^8$, we replace π by p in (2.13). It is then clear that the coordinates of the vector p with respect to the new basis are sufficient statistics for parameters in the saturated model, namely the marginal tables $\{p_{i++}\}$, $\{p_{+j+}\}$, $\{p_{ij+}\}$, $\{p_{++k}\}$, $\{p_{i+k}\}$, $\{p_{+jk}\}$ and $\{p_{ijk}\}$.

In Haberman geometry (after log-transformation), on the standard basis the log-transformed joint probability mass function

$$\log \pi = (\log \pi_{111}, \log \pi_{112}, \log \pi_{121}, \log \pi_{122}, \log \pi_{211}, \log \pi_{212}, \log \pi_{221}, \log \pi_{222})^T$$

corresponds to a point on an extended simplex S_7 in the negative orthant in $\mathbf{R}^8$. The vector $\log \pi$, however, can also be projected onto the new basis $\{x_1, x_2, \ldots, x_8\}$ as

$$\log \pi = \frac{\log \pi . x_1}{\|x_1\|^2} x_1 + \frac{\log \pi . x_2}{\|x_2\|^2} x_2 + \ldots + \frac{\log \pi . x_8}{\|x_8\|^2} x_8$$

where

$$
\begin{aligned}
\frac{\log \pi . x_1}{\|x_1\|^2} &= \frac{1}{8} \log \left(\pi_{111} \pi_{112} \pi_{121} \pi_{122} \pi_{211} \pi_{212} \pi_{221} \pi_{222} \right) \\
\frac{\log \pi . x_2}{\|x_2\|^2} &= \frac{1}{8} \log \left(\frac{\pi_{111} \pi_{112} \pi_{121} \pi_{122}}{\pi_{211} \pi_{212} \pi_{221} \pi_{222}} \right) \\
\frac{\log \pi . x_3}{\|x_3\|^2} &= \frac{1}{8} \log \left(\frac{\pi_{111} \pi_{112} \pi_{211} \pi_{212}}{\pi_{121} \pi_{122} \pi_{221} \pi_{222}} \right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\log \pi . \, x_4}{\|x_4\|^2} &= \frac{1}{8} \log \left(\frac{\pi_{111}\pi_{112}\pi_{221}\pi_{222}}{\pi_{121}\pi_{122}\pi_{211}\pi_{212}} \right) \\
\frac{\log \pi . \, x_5}{\|x_5\|^2} &= \frac{1}{8} \log \left(\frac{\pi_{111}\pi_{121}\pi_{211}\pi_{221}}{\pi_{112}\pi_{122}\pi_{212}\pi_{222}} \right) \\
\frac{\log \pi . \, x_6}{\|x_6\|^2} &= \frac{1}{8} \log \left(\frac{\pi_{111}\pi_{121}\pi_{212}\pi_{222}}{\pi_{112}\pi_{122}\pi_{211}\pi_{221}} \right) \\
\frac{\log \pi . \, x_7}{\|x_7\|^2} &= \frac{1}{8} \log \left(\frac{\pi_{111}\pi_{122}\pi_{211}\pi_{222}}{\pi_{112}\pi_{121}\pi_{212}\pi_{221}} \right) \\
\frac{\log \pi . \, x_8}{\|x_8\|^2} &= \frac{1}{8} \log \left(\frac{\pi_{111}\pi_{122}\pi_{212}\pi_{221}}{\pi_{112}\pi_{121}\pi_{211}\pi_{222}} \right)
\end{aligned} \tag{2.14}
$$

The unsaturated model

For a given unsaturated model, the whole space $\mathbf{R}^8$ can be partitioned into a sufficiency space and an auxiliary space. The elements of the new basis corresponding to column vectors of the design matrix in the model span the sufficiency space, and the remaining elements in the new basis span the auxiliary space. Here we demonstrate the geometry for the conditional independence model $X_1 \perp\!\!\!\perp X_2 \mid X_3$, and then summarize the geometry for other commonly used models.

Since the model $X_1 \perp\!\!\!\perp X_2 \mid X_3$ has model symbol $(1,\ X_1,\ X_2,\ X_3,\ X_1X_3,\ X_2X_3)$, the sufficient statistics are $\{p_{i++}\}$, $\{p_{+j+}\}$, $\{p_{++k}\}$, $\{p_{i+k}\}$ and $\{p_{+jk}\}$, the marginal tables of the observation table $\{p_{ijk}\}$. Cell probabilities of the model can be represented in terms of $\{\pi_{i+k}\}$ and $\{\pi_{+jk}\}$ as

$$
\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \qquad \text{for all } i,\ j,\ k \tag{2.15}
$$

noting that $\pi_{++k} = \sum_i \pi_{i+k} = \sum_j \pi_{+jk}$.

As elements associated with the columns vector of the design matrix in the new basis are x_1, x_2, x_3, x_5, x_6 and x_7, the whole model space can be partitioned as

$$
\mathbf{R}^8 = S \oplus \mathcal{A}
$$

where $\mathcal{S} = \text{span}\{x_1, x_2, x_3, x_5, x_6, x_7\}$ and $\mathcal{A} = \mathcal{S}^{\perp} = \text{span}\{x_4, x_8\}$.

In Fienberg geometry, referring to an observed table $\{p_{ijk}\}$ (corresponding to a vector p) and its maximum likelihood fitted table $\{\hat{\pi}_{ijk}\}$ (corresponding to a vector $\hat{\pi}$) for the model, the coordinates of p with respect to the new basis can be obtained by substituting p for π (2.13). Similarly, the coordinates of $\hat{\pi}$ with respect to the new basis can be obtained by substituting $\hat{\pi}$ for π in (2.13), using probability relationship (2.15). The results are shown in the following table.

p		$\hat{\pi}$
$\frac{1}{8}$	x_1	$\frac{1}{8}$
$\frac{p_{1++} - p_{2++}}{8}$	x_2	$\frac{\hat{\pi}_{1++} - \hat{\pi}_{2++}}{8}$
$\frac{p_{+1+} - p_{+2+}}{8}$	x_3	$\frac{\hat{\pi}_{+1+} - \hat{\pi}_{+2+}}{8}$
$\frac{p_{11+} - p_{12+} - p_{21+} + p_{22+}}{8}$	x_4	$\frac{(\hat{\pi}_{1++} - \hat{\pi}_{2++})(\hat{\pi}_{+1+} - \hat{\pi}_{+2+})}{8}$
$\frac{p_{++1} - p_{++2}}{8}$	x_5	$\frac{\hat{\pi}_{++1} - \hat{\pi}_{++2}}{8}$
$\frac{p_{1+1} - p_{1+2} - p_{2+1} + p_{2+2}}{8}$	x_6	$\frac{\hat{\pi}_{1+1} - \hat{\pi}_{1+2} - \hat{\pi}_{2+1} + \hat{\pi}_{2+2}}{8}$
$\frac{p_{+11} - p_{+12} - p_{+21} + p_{+22}}{8}$	x_7	$\frac{\hat{\pi}_{+11} - \hat{\pi}_{+12} - \hat{\pi}_{+21} + \hat{\pi}_{+22}}{8}$
$\frac{p_{111} - p_{112} - p_{121} + p_{122} - p_{211} + p_{212} + p_{221} - p_{222}}{8}$	x_8	$\frac{(\hat{\pi}_{1+1} - \hat{\pi}_{2+1})(\hat{\pi}_{+11} + \hat{\pi}_{+21})}{8\hat{\pi}_{++1}} - \frac{(\hat{\pi}_{1+2} - \hat{\pi}_{2+2})(\hat{\pi}_{+12} + \hat{\pi}_{+22})}{8\hat{\pi}_{++2}}$

After applying Birch's results we have that

$$p_{i++} = \hat{\pi}_{i++},\ p_{+j+} = \hat{\pi}_{+j+},\ p_{++k} = \hat{\pi}_{++k},\ p_{i+k} = \hat{\pi}_{i+k},\ \text{and}\ p_{+jk} = \hat{\pi}_{+jk}$$

Now we find that the coordinates of $\hat{\pi}$ are the same as the coordinates of p with respect to the sufficiency basis $\{x_1, x_2, x_3, x_5, x_6, x_7\}$ and that these coordinates are the sufficient statistics for parameters in the model.

Similarly, in Haberman geometry, for the model $X_1 \perp\!\!\!\perp X_2 \mid X_3$, the log-transformed table $\{\log p_{ijk}\}$ (represented by $\log p$) and its fitted table $\{\log \hat{\pi}_{ijk}\}$ (represented by $\log \hat{\pi}$) have coordinates with respect to the new basis of

$\log p$		$\log \hat{\pi}$
$\frac{1}{2\sqrt{2}} \log\left(p_{111}p_{112}p_{121}p_{122}p_{211}p_{212}p_{221}p_{222}\right)$	x_1	$\frac{1}{\sqrt{2}} \log\left(\frac{p_{1+1}p_{+11}p_{1+2}p_{+12}p_{2+1}p_{+21}p_{2+2}p_{+22}}{p_{++1}^2 p_{++2}^2}\right)$
$\frac{1}{2\sqrt{2}} \log\left(\frac{p_{111}p_{122}p_{121}p_{122}}{p_{211}p_{212}p_{221}p_{222}}\right)$	x_2	$\frac{1}{\sqrt{2}} \log\left(\frac{p_{1+1}p_{1+2}}{p_{2+1}p_{2+2}}\right)$
$\frac{1}{2\sqrt{2}} \log\left(\frac{p_{111}p_{112}p_{211}p_{212}}{p_{121}p_{122}p_{221}p_{222}}\right)$	x_3	$\frac{1}{\sqrt{2}} \log\left(\frac{p_{+11}p_{+12}}{p_{+21}p_{+22}}\right)$
$\frac{1}{2\sqrt{2}} \log\left(\frac{p_{111}p_{112}p_{221}p_{222}}{p_{121}p_{122}p_{211}p_{212}}\right)$	x_4	0
$\frac{1}{2\sqrt{2}} \log\left(\frac{p_{111}p_{121}p_{211}p_{221}}{p_{112}p_{122}p_{212}p_{222}}\right)$	x_5	$\frac{1}{\sqrt{2}} \log\left(\frac{p_{++2}^2(p_{1+1}p_{+11}p_{2+1}p_{+21})}{p_{++1}^2(p_{1+2}p_{+12}p_{2+2}p_{+22})}\right)$
$\frac{1}{2\sqrt{2}} \log\left(\frac{p_{111}p_{121}p_{212}p_{222}}{p_{112}p_{122}p_{211}p_{221}}\right)$	x_6	$\frac{1}{\sqrt{2}} \log\left(\frac{p_{1+1}p_{2+2}}{p_{2+1}p_{1+2}}\right)$
$\frac{1}{2\sqrt{2}} \log\left(\frac{p_{111}p_{122}p_{211}p_{222}}{p_{112}p_{121}p_{212}p_{221}}\right)$	x_7	$\frac{1}{\sqrt{2}} \log\left(\frac{p_{+11}p_{+22}}{p_{+12}p_{+21}}\right)$
$\frac{1}{2\sqrt{2}} \log\left(\frac{p_{111}p_{122}p_{212}p_{221}}{p_{112}p_{121}p_{211}p_{222}}\right)$	x_8	0

Now we find that the coordinates of $\log \hat{\pi}$ with respect to the basis of the auxiliary space $\{x_4, x_8\}$ are zeros.

Properties of the coordinates of the fitted vector for certain models, in the two geometries, are summarized in Table 2.2.

To interpret the results in Table 1, consider the model $X_1 \perp\!\!\!\perp X_2 \mid X_3$, symbolized as $(X_1X_3,\ X_2X_3)$ in Agresti (1990). In the Fienberg geometry the fitted vector will have the same coordinates as the data vector with respect to the sufficiency basis $\{x_1, x_2, x_3, x_5, x_6, x_7\}$, while in the Haberman geometry the fitted vector will zero the coordinates with respect to the basis of the auxiliary space $\{x_4, x_8\}$, as shown in Table 2.2.

2.5 Conclusion

We have described two geometric approaches to categorical data models (Fienberg and Haberman geometries) from the simplest case to the general case. As with the geometry of linear models the whole space can be split into a sufficiency space and

Model		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Saturated	F	$(c_1$	c_2	c_3	c_4	c_5	c_6	c_7	$c_8)$
$(X_1X_2,\ X_1X_3,\ X_2X_3)$	F	$(c_1$	c_2	c_3	c_4	c_5	c_6	c_7	$*)$
	H	$(*$	$*$	$*$	$*$	$*$	$*$	$*$	$0)$
$(X_1X_3,\ X_2X_3)$	F	$(c_1$	c_2	c_3	$*$	c_5	c_6	c_7	$*)$
or $X_1 \perp\!\!\!\perp X_2 \mid X_3$	H	$(*$	$*$	$*$	0	$*$	$*$	$*$	$0)$
$(X_1,\ X_2X_3)$	F	$(c_1$	c_2	c_3	$*$	c_5	$*$	c_7	$*)$
or $X_1 \perp\!\!\!\perp (X_2, X_3)$	H	$(*$	$*$	$*$	0	$*$	0	$*$	$0)$
$(X_1,\ X_2,\ X_3)$	F	$(c_1$	c_2	c_3	$*$	c_5	$*$	$*$	$*)$
or $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3$	H	$(*$	$*$	$*$	0	$*$	0	0	$0)$

Table 2.2: The coordinates of the fitted vector for five commonly seen models for variables X_1, X_2 and X_3 in each of the two geometries. Here c_i $(i = 1, 2, \ldots, 8)$ are the coordinates of the new basis in Fienberg geometry, "F" refers to Fienberg geometry, "H" refers to Haberman geometry, and an asterisk denotes a coordinate which cannot be obtained from the given data and model immediately. The sufficient statistics are in each case the marginal distributions of the terms occurring in the Agresti model symbol. We find that the fitting preserves the coordinates on the basis of the sufficiency space in Fienberg geometry, while zeroes the coordinates on the basis of the auxiliary space in Haberman geometry.

an auxiliary space through a change of basis which is determined by the sufficient statistics. In Fienberg geometry the coordinates of the basis for the sufficiency space are preserved during maximum likelihood model fitting, while in Haberman geometry the coordinates of the basis for the auxiliary space are zeroed. The relationship between the two geometries is summarized by a commutative diagram.

Chapter 3

The geometry of GLMs

3.1 Introduction

Generalized linear models were introduced by Nelder and Wedderburn in 1972 and became popular gradually during the 1980s. The response variable in a generalized linear model is allowed to follow a distribution from an exponential family, rather than specifically the normal distribution, as in a linear model. Furthermore, the mean of the response variable is linearly related to explanatory variables through a link function. The linear model and loglinear model are special cases of the generalized linear model with an identity link and a log link respectively. In Chapter 1 the geometry of linear models was discussed from Fisher's point view, while in Chapter 2 the geometry of loglinear models was revealed by combining the two distinct geometric views contributed by Fienberg (1968, 1970) and Haberman (1974) respectively. In 1997 the geometry of generalized linear models was described by Kass and Vos using the language of differential geometry.

In this chapter, however, we elucidate a geometry underlying generalized linear models by extending and melding the two geometries developed for loglinear models.

In doing so we provide a succinct framework for conceptualizing such models, in which the geometries of linear models and loglinear models are seen as special cases. Sufficiency and linearity play key roles, with the general framework extending and linking the loglinear model–specific approaches of Fienberg and Haberman.

This chapter is organized as follows. In the next section, generalized linear models are briefly described and key results about sufficient statistics are highlighted. Section 3 then describes an existing geometric framework, contributed by Kass and Vos, underlying generalized linear models. In Section 4 an alternative geometry is obtained by generalizing the geometry of categorical data models. Section 5 illustrates the alternative geometry using a logistic regression example. Finally, a conclusion completes this chapter.

3.2 GLMs and sufficient statistics

A generalized linear model comprises three parts (McCullagh and Nelder, 1983):

i) A random vector Y with a realized value $y = [y_1, y_2, \ldots, y_n]^T$ whose elements are assumed to be independent realizations from a single exponential family, so the ith observation has density function

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\} \text{ for } i = 1, 2, \ldots, n$$

where θ_i is called the natural parameter, and ϕ is called the dispersion parameter, assumed given or estimated.

ii) Parameters $\beta = [\beta_1, \beta_2, \ldots, \beta_q]^T$ and the $n \times q$ design matrix $X = [x_1\ x_2\ \ldots\ x_q]$ where x_j is the jth independent variable with elements $[x_{1j}, x_{2j}, \ldots, x_{nj}]^T$ for $j = 1, 2, \ldots, q$.

iii) A monotone differentiable link function g for which $\eta_i = g(\mu_i)$ for all i, where $\mu = [\mu_1, \mu_2, \ldots, \mu_n]^T = E(Y)$ and $\eta = [\eta_1, \eta_2, \ldots, \eta_n]^T = X\beta$. A link function is termed canonical if $g(\mu_i) = \theta_i$ for all i. For convenience, we denote the link function in vector form as $\eta = g(\mu)$ where $g(\mu) = [g(\mu_1), g(\mu_2), \ldots, g(\mu_n)]^T$.

To avoid later confusion, we point out now that, for the binomial distribution, the canonical link function has the form

$$g(\mu_i) = \log\left(\frac{\frac{\mu_i}{n_i}}{1 - \frac{\mu_i}{n_i}}\right)$$

where $\mu_i = E(Y_i)$ and n_i is the number of observations associated with the y_i outcome.

Standard theory for this situation (Agresti 1990, p.446-447) gives expressions for the mean and variance of Y as

$$\mu_i = E(Y) = b'(\theta_i) \qquad \text{and} \qquad Var(Y) = b''(\theta_i)a(\phi)$$

Examples of generalized linear models include loglinear models, Poisson regression and logistic regression. Two properties of such models, pertinent to our geometric development in the next section, are presented in the next theorem.

Theorem 3.1. *For a generalized linear model with canonical link function, observation vector y and $\hat{\mu}$, the maximum likelihood estimator of the mean vector μ,*

i) $y.x_j$ is sufficient for β_j, and

ii) $\hat{\mu}.x_j = y.x_j$

for $j = 1, 2, \ldots, q$.

Proof. i) Since $y = [y_1, y_2, \ldots, y_n]^T$ with

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}$$

for all i, the joint probability density function of $Y_1, Y_2, \ldots, Y_n$ is

$$f(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n, \phi) = \exp\{\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)]/a(\phi) + \sum_{i=1}^{n} c(y_i, \phi)\} \qquad (3.1)$$

For a generalized linear model with canonical link we have

$$\theta_i = g(\mu_i) = \sum_{j=1}^{q} \beta_j x_{ij}$$

so the right hand side of Expression (3.1) becomes

$$\exp\{[\sum_{j=1}^{q}(\sum_{i=1}^{n} x_{ij}y_i)\beta_j - \sum_{i=1}^{n} b(\sum_{j=1}^{q}\beta_j x_{ij})]/a(\phi) + \sum_{i=1}^{n} c(y_i, \phi)\}$$

$$= h(y)d(\beta)\exp\left[\sum_{j=1}^{q}(\sum_{i=1}^{n} x_{ij}y_i)\beta_j/a(\phi)\right]$$

where $h(y) = \exp\left[\sum_{i=1}^{n} c(y_i, \phi)\right]$ and $d(\beta) = \exp\left[-\sum_{i=1}^{n} b(\sum_{j=1}^{q}\beta_j x_{ij})\right]/a(\phi)$. Thus, by the Factorization Theorem (George and Roger 1990, p.250), a sufficient statistic for β_j is

$$\sum_{i=1}^{n} x_{ij}y_i = y.\, x_j \qquad \text{for} \qquad j = 1, 2, \ldots, q$$

ii) Recall that $\eta_i = g(\mu_i)$, so with the canonical link we have $\eta_i = g(\mu_i) = \theta_i$. Thus

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} \Big/ \frac{\partial \eta_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i)$$

Furthermore

$$\frac{\partial \ell(\theta : y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \qquad \text{(Agresti 1990, p.448)}$$

where $\theta = [\theta_1, \theta_2, \ldots, \theta_n]^T$, $\ell(\theta : y) = \sum_{i=1}^{n} \{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}$ is the log likelihood function and x_{ij} is the ijth entry in the design matrix. Thus

$$\frac{\partial \ell(\theta : y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{a(\phi)}$$

which in vector form becomes

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{X^T(y - \mu)}{a(\phi)}$$

where $a(\phi)$ is known and identical for all observations.

The maximum likelihood estimate $\hat{\mu}$ of μ is obtained by solving $X^T(y - \mu) = 0$, hence we have that $X^T\hat{\mu} = X^T y$, as required. Thus

$$\hat{\mu}.\, x_j = y.\, x_j \qquad \text{for} \qquad j = 1, 2, \ldots, q$$

□

From Theorem 3.1 we know that $\hat{\mu}.\, x_j = y.\, x_j$, so

$$\hat{\mu}.\, \frac{x_j}{\|x_j\|} = y.\, \frac{x_j}{\|x_j\|} \qquad \text{for} \qquad j = 1, 2, \ldots, q$$

Since $y.\, x_j$ is a sufficient statistic for the parameter β_j for all j, then $y.\, x_j/\|x_j\|$ also is a sufficient statistic for β_j due to the one-to-one relationship between $y.\, x_j$ and $y.\, x_j/\|x_j\|$ for all j. Thus, in Euclidean space $\mathbf{R}^n$, the sufficient statistics for β in a generalized linear model are the lengths (ignoring the sign) of projection of the observation vector onto the directions specified by the column vectors of the design matrix, and these sufficient statistics are preserved in the fitting process. The associated ideas are represented in Figure 3.1.

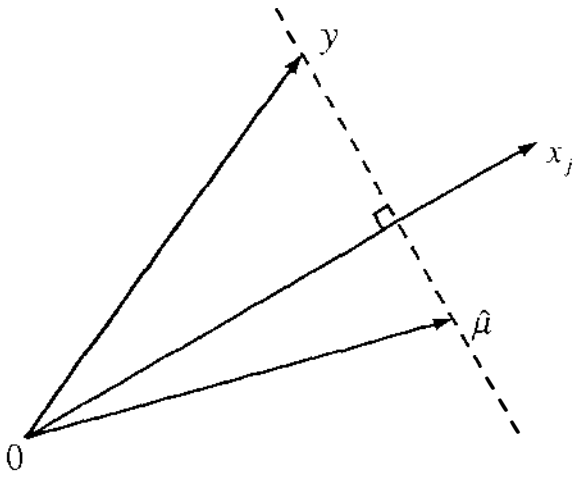


Figure 3.1: For a generalized linear model with canonical link the observation vector y and its fitted vector $\hat{\mu}$ have the same projection onto the direction specified by a column vector of the design matrix x_j, for $j = 1, 2, \ldots, q$.

3.3 Kass and Vos approach

There is an existing geometric framework for generalized linear models, discussed by Kass and Vos, using differential geometry. Here we first review the main ideas considered by Kass and Vos, then construct an alternative geometric framework for generalized linear models by developing the geometry of categorical data models.

Differential geometry

From a differential geometry point of view, Kass and Vos (1997) built a geometric framework for generalized linear models. In this geometric framework an inner product was defined to reveal an important property of the maximum likelihood estimate. To develop an understanding of the Kass and Vos ideas, the independence model of a 2×2 table is now illustrated.

Geometric structure

Kass and Vos viewed a realized value of random variable Y, $y = [y_1, y_2, \ldots, y_n]^T$, as a vector in the vector space $\mathbf{R}^n$ equipped with an inner product defined by the Fisher information matrix. There are two families of density functions, then, relating to generalized linear models. One is an n-dimensional exponential family distribution set

$$\mathcal{F} = \left\{ f : f(y; \theta, \phi) = \exp\left\{ [y^T\theta - b(\theta)]/a(\phi) + c(y, \phi) \right\} \right\}$$

where $\theta = [\theta_1, \theta_2, \ldots, \theta_n]^T$ is in the n-dimensional natural parameter space Θ of the exponential family, and $\phi = [\phi_1, \phi_2, \ldots, \phi_n]^T$ is the dispersion parameter, assumed given or estimated. Another is a q-dimensional subset of $\mathcal{F}$ determined by the design matrix X and the link function g via

$$\mathcal{M} = \left\{ f \in \mathcal{F} : \mu(f) = g^{-1}(X\beta) \right\}$$

where $\mu : f \mapsto \mu(f)$ from $\mathcal{F}$ to $\mathbf{R}^n$ provides the mean of f, g^{-1} is the inverse of the link function g, and $\beta = [\beta_1, \beta_2, \ldots, \beta_q]^T$ is in the q-dimensional parameter space B. Note that B specifies a set of n-dimensional exponential densities whose mean is determined by β and the design matrix X. Corresponding to these two families of density functions, there are two mean sets $\mathcal{F}_\mathbf{R} = \{\mu(f) : f \in \mathcal{F}\}$ and $\mathcal{M}_\mathbf{R} = \{\mu(f) : f \in \mathcal{M}\}$, so $\mathcal{M}_\mathbf{R} \subset \mathcal{F}_\mathbf{R}$.

Since $\mathcal{M} \subset \mathcal{F}$, then any point f (a density function) in $\mathcal{M}$, is also in $\mathcal{F}$. When f is viewed as a point on $\mathcal{F}$, there is a tangent space of $\mathcal{F}$ at f denoted by $T_f\mathcal{F}$ which is defined as the span of the score functions U_i with respect to the mean μ, so

$$U_i = \frac{\partial \ell(\theta : Y)}{\partial \mu_i} \qquad \text{for } i = 1, 2, \ldots, n$$

where

$$\ell(\theta : Y) = [Y^T\theta - b(\theta)]/a(\phi) + c(Y, \phi)$$

and θ relates to μ through $\theta_i = b'^{-1}(\mu_i)$ for all i.

When f, however, is viewed as a point in $\mathcal{M}$, the associated tangent space at f, denoted by $T_f\mathcal{M}$, is spanned by the score functions with respect to the parameter β

$$V_j = \frac{\partial \ell(\theta : Y)}{\partial \beta_j} \qquad \text{for } j = 1, 2, \ldots, q$$

here θ relates to μ through $\theta_i = b'^{-1}(\mu_i)$ for all i and μ relates to β through $\mu = g^{-1}(X\beta)$. Thus, it can be shown that $T_f\mathcal{M}$ is a linear subspace of $T_f\mathcal{F}$ using the chain rule (Kass and Vos 1997, p.124).

Note that the mean sets $\mathcal{F}_{\mathbf{R}}$ and $\mathcal{M}_{\mathbf{R}}$ are the image of the density sets $\mathcal{F}$ and $\mathcal{M}$ respectively under the map $\mu(f) : \mathcal{F} \mapsto \mathbf{R}^n$ where μ is the mean of a density f. Since all of these sets may not be vector spaces for a generalized linear model, $T_f\mathcal{F}$ and $T_f\mathcal{M}$ are constructed as vector spaces approximating $\mathcal{F}$ and $\mathcal{M}$ at $f \in \mathcal{M}$ locally. The relationship among the families of density functions, mean sets and tangent spaces is shown in Figure 3.2.

Geometric property of maximum likelihood estimate

Suppose $\hat{\mu} = [\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n]^T$ is the maximum likelihood estimate for the mean of a generalized linear model with observation vector $y = [y_1, y_2, \ldots, y_n]^T$. The observation vector y, then, can be linked with the geometric structure discussed above by constructing vectors in $T_f\mathcal{F}$ such as $y - \mu = \sum_i (y_i - \mu_i)U_i$ where $\mu = [\mu_1, \mu_2, \ldots, \mu_n]^T \in \mathcal{M}_{\mathbf{R}}$ and U_i is the score functions with respect to the μ_i for all i.

Next, an inner product is defined in $T_f\mathcal{F}$, and thus in $T_f\mathcal{M}$, via

$$\langle U_i, U_j \rangle = \mathrm{Cov}(U_i, U_j) = E(U_iU_j)$$

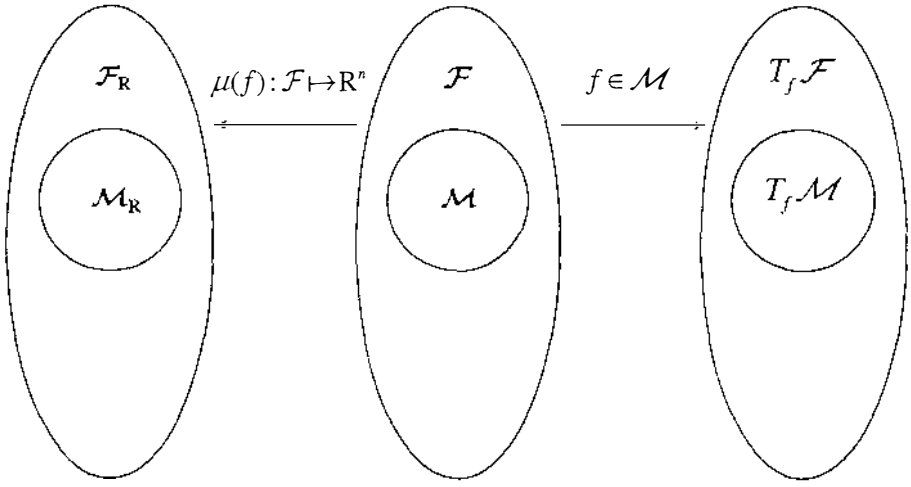


Figure 3.2: The graphic shows that the mean sets $\mathcal{F}_{\mathbf{R}}$ and $\mathcal{M}_{\mathbf{R}}$ are the image of the families of density functions $\mathcal{F}$ and $\mathcal{M}$ respectively under the map $\mu(f) : \mathcal{F} \mapsto \mathbf{R}^n$ where μ is the mean of a density f. For each point $f \in \mathcal{M}$, there are tangent spaces $T_f\mathcal{F}$ and $T_f\mathcal{M}$ approximating $\mathcal{F}$ and $\mathcal{M}$ locally at f.

Then for any $a, b \in T_f\mathcal{F}$ we have

$$\begin{aligned} \langle a, b \rangle &= E\Big(\sum_{i,j} a_i b_j U_i U_j\Big) \\ &= \sum_{i,j} a_i b_j E(U_i U_j) = a^T I_n(\mu) b \end{aligned}$$

where $a = \sum_i a_i U_i$, $b = \sum_j b_j U_j$ and $I_n(\mu)$ is an $n \times n$ matrix with ijth element $E(U_i U_j)$. Note that $I_n(\mu)$ is the Fisher information matrix for the mean parameter at $\mu = \mu(f)$ (the detail see Kass and Vos (1997, p.17–18)).

Kass and Vos (1997, p.127) shown that the maximum likelihood estimate $\hat{\mu} = \mu(\hat{f})$ is obtained by satisfying

$$(y - \hat{\mu}) \perp T_{\hat{f}}\mathcal{M}$$

where $y - \hat{\mu} \in T_{\hat{f}}\mathcal{F}$ under the inner product defined above.

Example

For example, a 2×2 frequency table with data vector form $y = [y_1, y_2, y_3, y_4]^T$ (where each y_i is a count) is observed from a Poisson sampling scheme. Suppose the underlying mean of Y is $\mu = [\mu_1, \mu_2, \mu_3, \mu_4]^T$, and for the independence model the maximum likelihood estimate of μ is $\hat{\mu} = [\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4]^T$. Then

$$\mathcal{F} = \left\{ f : f(y;\theta) = \exp\left\{ y^T\theta - \sum_i \left(\exp(\theta_i)\right) - \sum_i \log(y_i!) \right\},\ \theta \in \mathbf{R}^n \right\}$$

where $\theta = [\log \mu_1, \log \mu_2, \log \mu_3, \log \mu_4]^T$ with $\theta_i = \log \mu_i$. If $\mu(f) : \mathcal{F} \mapsto \mathbf{R}^4$ is the mean of f, then

$$\mathcal{M} = \left\{ f \in \mathcal{F} : \mu_i(f) = \exp\left(\sum_{j=1}^{3} \beta_j x_{ij} \right),\ \text{for } i = 1, 2, 3, 4 \right\}$$

where x_{ij} is the ijth entry of the following matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

Now we obtain that

$$\mathcal{F}_{\mathbf{R}} = \left\{ \mu : f(y;\mu) = \exp\left\{ y^T \log \mu - \sum_i \mu_i - \sum_i \log(y_i!) \right\},\ \mu \in (\mathbf{R}^+)^n \right\}$$

and

$$\mathcal{M}_{\mathbf{R}} = \left\{ \mu :\ \frac{\mu_1\mu_4}{\mu_2\mu_3} = 1 \text{ and } \mu \in \mathcal{F}_{\mathbf{R}} \right\}$$

and it is clear that $\mathcal{M}_{\mathbf{R}} \subset \mathcal{F}_{\mathbf{R}}$.

Since the log likelihood function for y is

$$l(\mu;\ y) = y^T \log \mu - \sum_i \mu_i - \sum_i \log(y_i!)$$

we obtain that

$$\frac{\partial l(\mu;\ y)}{\partial \mu_i} = \frac{y_i - \mu_i}{\mu_i} \tag{3.2}$$

Now, for any point f in $\mathcal{M}$, and thus in $\mathcal{F}$, there is a tangent space of $\mathcal{F}$ at f given by

$$T_f\mathcal{F} = \text{span}\{U_i\}_{i=1}^4$$

where

$$U_i = \frac{\partial l(\mu;\ Y)}{\partial \mu_i} = \frac{Y_i - \mu_i}{\mu_i} \qquad \text{for } Y_i \in (0, \infty)$$

for $i = 1, 2, 3, 4$.

On the other hand, from (3.2) and $\mu_i = \exp\left(\sum_{j=1}^3 \beta_j x_{ij}\right)$ for all i we have

$$\begin{aligned} \frac{\partial l(\mu;\ y)}{\partial \beta_j} &= \sum_{i=1}^4 \frac{\partial\left[\exp\left(\sum_{j=1}^3 \beta_j x_{ij}\right)\right]}{\partial \beta_j} \frac{\partial l(\mu;\ y)}{\partial \mu_i} \\ &= \sum_{i=1}^4 \mu_i\, x_{ij} \frac{\partial l(\mu;\ y)}{\partial \mu_i} \end{aligned}$$

Thus on $\mathcal{M}$, the associated tangent space at f will be

$$T_f\mathcal{M} = \text{span}\{V_j\}_{j=1}^3$$

where

$$V_j = \frac{\partial l(\mu;\ Y)}{\partial \beta_j} = \sum_{i=1}^4 \mu_i\, x_{ij}\, U_i \qquad \text{for } j = 1, 2, 3$$

Specifically, we obtain

$$V_1 = [\mu_1, \mu_2, \mu_3, \mu_4]^T,\ V_2 = [\mu_1, \mu_2, -\mu_3, -\mu_4]^T,\ \text{and } V_3 = [\mu_1, -\mu_2, \mu_3, -\mu_4]^T$$

with respect to the basis $\{U_1, U_2, U_3, U_4\}$. Thus $T_f\mathcal{M} \subset T_f\mathcal{F}$.

In $T_f\mathcal{F}$, an inner product is defined by

$$\langle a, b\rangle = a^T I_4(\mu) b \qquad \text{for } a,\ b \in T_f\mathcal{F}$$

where $I_4(\mu)$ is a 4×4 matrix with ijth element $E(U_iU_j)$ such as

$$I_4(\mu) = \begin{bmatrix} \frac{1}{\mu_1} & 0 & 0 & 0 \\ 0 & \frac{1}{\mu_2} & 0 & 0 \\ 0 & 0 & \frac{1}{\mu_3} & 0 \\ 0 & 0 & 0 & \frac{1}{\mu_4} \end{bmatrix}$$

Let

$$l(\mu : y) = \sum_i y_i \log \mu_i - \sum_i \mu_i$$

where $\log \mu_i = x_{i1}\beta_1 + x_{i2}\beta_1 + x_{i3}\beta_3$ for $i = 1, 2, 3, 4$, then

$$\begin{aligned} \frac{\partial \ell(\mu : y)}{\partial \beta_j} &= \sum_i (y_i - \mu_i) x_{ij} \\ &= V_j^T I_4(\mu)(y - \mu) \\ &= \langle y - \mu, V_j \rangle \qquad \text{for } j = 1, 2, 3 \end{aligned}$$

where

$$y - \mu = [(y_1 - \mu_1), (y_2 - \mu_2), (y_3 - \mu_3), (y_4 - \mu_4)]^T \in T_f\mathcal{F}$$

Now the maximum likelihood estimate $\hat{\mu}$ will be the point in $\mathcal{M}_{\mathbf{R}}$ such that

$$\langle y - \hat{\mu}, \hat{V}_j \rangle = 0 \qquad \text{for} \qquad j = 1, 2, 3$$

Thus, $y - \hat{\mu}$ is orthogonal to $T_{\hat{f}}\mathcal{M}$ at $\hat{f}$ with $\hat{\mu} = \mu(\hat{f})$.

3.4 An alternative geometric approach

In the geometry of categorical data models, Fienberg and Haberman geometries were embedded in a unified framework where the whole space was partitioned based on

the new basis constructed from sufficient statistics. In the general context, "Fienberg geometry" will now refer to the framework before canonical link transformation and "Haberman geometry" to the framework after canonical link transformation. The construction is detailed next.

Geometric structure

As in the Kass and Vos differential geometric approach to generalized linear models, in the alternative geometric structure we shall consider the two density sets $\mathcal{F}$, $\mathcal{M}$ and the associated mean sets $\mathcal{F}_{\mathbf{R}}$, $\mathcal{M}_{\mathbf{R}}$. The density sets $\mathcal{F}$, $\mathcal{M}$, however, are not germane to our development. Here, what we are really concerned with is the mean sets $\mathcal{F}_{\mathbf{R}}$, $\mathcal{M}_{\mathbf{R}}$ in $\mathbf{R}^n$. Note that $\mathcal{F}_{\mathbf{R}}$ generalizes the tetrahedron discussed for a 2×2 contingency table, and $\mathcal{M}_{\mathbf{R}}$ generalizes the "butterfly" of the independence model. Furthermore, after the link function transformation associated with $\mathcal{F}_{\mathbf{R}}$ and $\mathcal{M}_{\mathbf{R}}$ we obtain two transformed mean sets $g(\mathcal{F}_{\mathbf{R}}) = \{g(\mu) : \mu \in \mathcal{F}_{\mathbf{R}}\}$ and $g(\mathcal{M}_{\mathbf{R}}) = \{g(\mu) : \mu \in \mathcal{M}_{\mathbf{R}}\}$ such that $g(\mathcal{M}_{\mathbf{R}}) \subset g(\mathcal{F}_{\mathbf{R}}) \subset \mathbf{R}^n$. Note that $g(\mathcal{F}_{\mathbf{R}})$ generalizes the extended tetrahedron discussed for a 2×2 contingency table, and $g(\mathcal{M}_{\mathbf{R}})$ generalizes the "jellyfish" for the independence model. Thus, the sets $\mathcal{F}_{\mathbf{R}}$ and $\mathcal{M}_{\mathbf{R}}$ generalize Fienberg structures, while the sets $g(\mathcal{F}_{\mathbf{R}})$ and $g(\mathcal{M}_{\mathbf{R}})$ generalize Haberman structures for loglinear models. The relationship among these sets is shown in Figure 3.3.

The partition of $\mathbf{R}^n$

To reveal the geometric properties of maximum likelihood estimates for a generalized linear model, we consider a partition of the space $\mathbf{R}^n$ in which $\mathcal{F}_{\mathbf{R}}$, $\mathcal{M}_{\mathbf{R}}$, $g(\mathcal{F}_{\mathbf{R}})$ and $g(\mathcal{M}_{\mathbf{R}})$ lie. Since the column vectors of the design matrix $x_1, x_2, \ldots, x_q$ are linearly

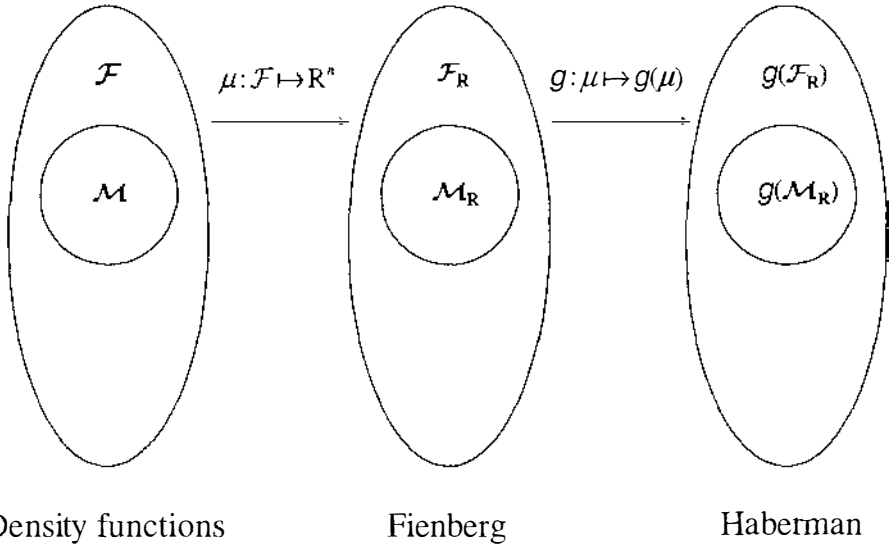


Figure 3.3: The relationships among the sets $\mathcal{F}$, $\mathcal{M}$, $\mathcal{F}_{\mathbf{R}}$, $\mathcal{M}_{\mathbf{R}}$, $g(\mathcal{F}_{\mathbf{R}})$ and $g(\mathcal{M}_{\mathbf{R}})$.

independent (assuming non-collinearity), the variation of the Gram-Schmidt process allows us to construct a basis in $\mathbf{R}^n$ as

$$\{x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_n\}$$

for which each new basis vector is orthogonal to all column vectors in the design matrix. Now the space $\mathbf{R}^n$ is split naturally into two orthogonal spaces, the "sufficiency space" $\mathcal{S}$ and the "auxiliary space" $\mathcal{A}$, with

$$\mathcal{S} = span\{x_1, \ldots, x_q\} \quad \text{and} \quad \mathcal{A} = \mathcal{S}^{\perp} = span\{x_{q+1}, \ldots, x_n\}$$

whence

$$\mathbf{R}^n = \mathcal{S} \oplus \mathcal{A}$$

In traditional linear model geometry, the data enters $\mathbf{R}^n$ as a vector. In this more general situation, where iteration is needed in model fitting, it is appropriate for the data to determine the "sufficiency" affine plane, $\mathcal{T} = s + \mathcal{A}$, where

$s = \{y.x_1, \ldots, y.x_q, 0, \ldots, 0\}$. Thus

$$\mathcal{T} = \{(y.x_1, \ldots, y.x_q, z_{q+1}, \ldots, z_n) \ : \ z_{q+1}, \ldots, z_n \in \mathbf{R}\}$$

The result which follows will determine the relationship between model and observations for the maximum likelihood fit $\hat{\mu}$.

Theorem 3.2. *Let $\hat{\mu}$ be the maximum likelihood estimator of μ. Then*

i) $\hat{\mu} \in \mathcal{T}$, and

ii) $g(\hat{\mu}) \in \mathcal{S}$.

Proof. Statement i) follows immediately from the fact that

$$\hat{\mu}.x_i = y.x_i \text{ for } i = 1, \ldots, q \qquad \text{(Theorem 3.1)}$$

For ii), simply note that

$$g(\hat{\mu}) = X\hat{\beta} \in \mathcal{S}$$

where $\hat{\beta}$ is the vector of fitted values of the parameters in β. □

Geometric properties of the maximum likelihood estimate

In Fienberg geometry, we consider the decomposition of a vector in the sets $\mathcal{F}_{\mathbf{R}}$ and $\mathcal{M}_{\mathbf{R}}$ with respect to the basis $\{x_1, x_2, \ldots, x_n\}$. Thus for a given generalized linear model, the observation vector $y \in \mathcal{F}_{\mathbf{R}}$ and its fitted mean vector $\hat{\mu} \in \mathcal{M}_{\mathbf{R}}$ have the form

$$y = (y.\frac{x_1}{\|x_1\|})\frac{x_1}{\|x_1\|} + (y.\frac{x_2}{\|x_2\|})\frac{x_2}{\|x_2\|} + \cdots + (y.\frac{x_q}{\|x_q\|})\frac{x_q}{\|x_q\|}$$

$$+(y.\frac{x_{q+1}}{\|x_{q+1}\|})\frac{x_{q+1}}{\|x_{q+1}\|} + \cdots + (y.\frac{x_n}{\|x_n\|})\frac{x_n}{\|x_n\|}$$

and

$$\hat{\mu} = (\hat{\mu}.\frac{x_1}{\|x_1\|})\frac{x_1}{\|x_1\|} + (\hat{\mu}.\frac{x_2}{\|x_2\|})\frac{x_2}{\|x_2\|} + \cdots + (\hat{\mu}.\frac{x_q}{\|x_q\|})\frac{x_q}{\|x_q\|}$$

$$+(\hat{\mu}.\frac{x_{q+1}}{\|x_{q+1}\|})\frac{x_{q+1}}{\|x_{q+1}\|} + \cdots + (\hat{\mu}.\frac{x_n}{\|x_n\|})\frac{x_n}{\|x_n\|}$$

respectively. Theorem 3.2 indicates that $y.x_i = \hat{\mu}.x_i$ for $i = 1, 2, \ldots, q$, so the observation vector y and its fitted vector $\hat{\mu}$ have the same coefficients when represented in terms of the sufficiency space basis $\{x_1, x_2, \ldots, x_q\}$; these coefficients are the sufficient statistics for the components of the parameter vector β.

In Haberman geometry, on the other hand, we consider the decomposition of a vector in the sets $g(\mathcal{F}_{\mathbf{R}})$ and $g(\mathcal{M}_{\mathbf{R}})$ with respect to the new basis. Thus for a given generalized linear model, the transformed observation vector $g(y) \in g(\mathcal{F}_{\mathbf{R}})$ and its fitted vector $g(\hat{\mu}) \in g(\mathcal{M}_{\mathbf{R}})$ have the form

$$g(y) = (g(y).\frac{x_1}{\|x_1\|})\frac{x_1}{\|x_1\|} + (g(y).\frac{x_2}{\|x_2\|})\frac{x_2}{\|x_2\|} + \cdots + (g(y).\frac{x_q}{\|x_q\|})\frac{x_q}{\|x_q\|}$$

$$+(g(y).\frac{x_{q+1}}{\|x_{q+1}\|})\frac{x_{q+1}}{\|x_{q+1}\|} + \cdots + (g(y).\frac{x_n}{\|x_n\|})\frac{x_n}{\|x_n\|}$$

and

$$g(\hat{\mu}) = (g(\hat{\mu}).\frac{x_1}{\|x_1\|})\frac{x_1}{\|x_1\|} + (g(\hat{\mu}).\frac{x_2}{\|x_2\|})\frac{x_2}{\|x_2\|} + \cdots + (g(\hat{\mu}).\frac{x_q}{\|x_q\|})\frac{x_q}{\|x_q\|}$$

$$+(g(\hat{\mu}).\frac{x_{q+1}}{\|x_{q+1}\|})\frac{x_{q+1}}{\|x_{q+1}\|} + \cdots + (g(\hat{\mu}).\frac{x_n}{\|x_n\|})\frac{x_n}{\|x_n\|}$$

respectively. Theorem 3.2 indicates that $g(\hat{\mu}).x_i = 0$ for $i = q+1, \ldots, n$, so the coefficients of the transformed fitted vector $g(\hat{\mu})$ are zeroed on the auxiliary space

basis $\{x_{q+1}, x_{q+2}, \ldots, x_n\}$. This property holds for any element in $g(\mathcal{M}_{\mathbf{R}})$, which implies $g(\mathcal{M}_{\mathbf{R}}) \subset \mathcal{S}$.

Figure 4.2 provides a schematic illustration of the geometric components of a generalized linear model. Both $\mathcal{S}$ and $\mathcal{A}$ need to be at least two-dimensional in order to avoid degeneracy, so they are pictured, necessarily in three-space, intersecting at the origin alone. The untransformed mean space $\mathcal{M}_{\mathbf{R}}$ in general cuts across all dimensions of the space, while the transformed mean space $g(\mathcal{M}_{\mathbf{R}})$ lies in $\mathcal{S}$.

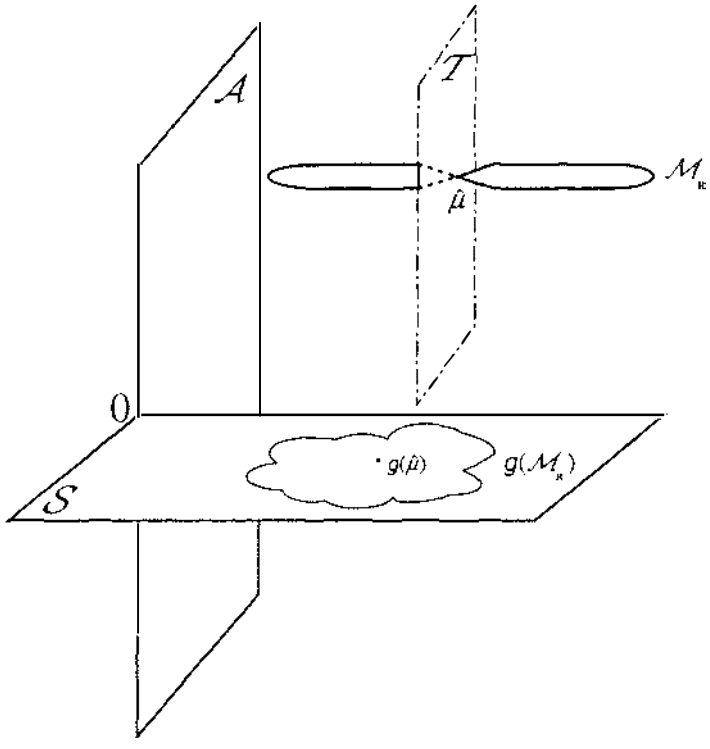


Figure 3.4: The geometric components of a generalized linear model. For a sample of size n, $\mathbf{R}^n$ splits into an orthogonal direct sum of the sufficiency space $\mathcal{S}$ and the auxiliary space $\mathcal{A}$. The maximum likelihood estimator of the mean $\hat{\mu}$ lies in the intersection of the sufficiency affine plane $\mathcal{T} = s + \mathcal{A}$ and the untransformed model space $\mathcal{M}_{\mathbf{R}}$. The link transformed mean vector $g(\hat{\mu})$ lies in the transformed mean space $g(\mathcal{M}_{\mathbf{R}})$.

Therefore in this geometric view of generalized linear models, the columns of

the design matrix determine the sufficient statistics $y.x_1, y.x_2, \ldots, y.x_q$ for β and hence μ. These values are preserved in maximum likelihood fitting, providing the first geometric result in Theorem 4.3.1: y and $\hat{\mu}$ have the same projection onto any direction in the sufficiency space. The columns of the design matrix determine the transformed model mean manifold (sometimes a linear space, but not necessarily so), providing the second geometric result in Theorem 4.3.1: $g(\hat{\mu})$ is orthogonal to every direction in the auxiliary space.

In the geometry of generalized linear models, Fienberg geometry focuses on the untransformed model manifolds $\mathcal{F}_{\mathbf{R}}$ and $\mathcal{M}_{\mathbf{R}}$ while Haberman geometry focuses on the transformed model manifolds $g(\mathcal{F}_{\mathbf{R}})$ and $g(\mathcal{M}_{\mathbf{R}})$. Let A be the $n \times n$ matrix which changes a vector from representation with respect to the standard basis in $\mathbf{R}^n$ to representation with respect to the new basis. Then

$$A = [x_1\ x_2\ \ldots\ x_q\ x_{q+1}\ x_{q+2}\ \ldots\ x_n]^{-1}$$

Hence Ay is the untransformed observation vector with respect to the new basis, and $Ag(y)$ is the transformed observation vector with respect to the new basis. Thus the untransformed and transformed views are related as shown in the following commutative diagram:

$$\begin{array}{ccccc} y & \xrightarrow{A} & Ay & \xrightarrow{P_T} & P_T Ay \\ \Big\downarrow g & & \Big\downarrow AgA^{-1} & & \\ g(y) & \xrightarrow{A} & Ag(y) & \xrightarrow{P_S} & P_S Ag(y) \end{array}$$

noting that g is the canonical link function, $P_{\mathcal{T}}$ is projection onto the sufficiency affine plane $\mathcal{T}$, so, for a point expressed with respect to the new basis,

$$P_{\mathcal{T}}(z_1, \ldots, z_q, z_{q+1}, \ldots, z_n) = (y.x_1, \ldots, y.x_q, z_{q+1}, \ldots, z_n)$$

and P_S is non-orthogonal projection onto the transformed model space $g(\mathcal{M}_{\mathbf{R}})$ (discussed in Section 4.3).

The commutative diagram highlights that Fienberg and Haberman geometries are connected by the canonical link function g. Meanwhile, the results with respect to the standard basis and the new basis are related by the matrix A. In addition, the commutative diagram motivates a new algorithm, to be discussed in the next chapter, for fitting generalized linear models.

For linear models, where the link function is the identity, the Fienberg and Haberman geometries coalesce. This leads to fitted value combining the best of both worlds: the minimal sufficient statistics are preserved on the basis of the sufficiency space, while the coefficients are zeroed on the basis of the auxiliary space.

3.5 Example

To illustrate this alternative geometry of generalized linear models we consider fitting a logistic regression for an artificial data set with three observations, shown in Table 3.1.

y_i (Response)	n_i (Total)	x_i (Covariate)	$\log\left(\frac{y_i}{n_i}/(1-\frac{y_i}{n_i})\right)$
21	23	1	2.3514
10	45	2	-1.2528
8	12	3	0.6931

Table 3.1: The artificial logistic regression data set, with three observations.

Here we have the design matrix $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$, so the new basis in $\mathbf{R}^3$ is

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{and} \quad x_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

and the change of basis matrix is

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & -2 \\ 1 & 3 & 1 \end{bmatrix}^{-1}$$

Now the whole space $\mathbf{R}^3$ can be split into sufficiency space

$$\mathcal{S} = \text{span}\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right\}$$

and auxiliary space

$$\mathcal{A} = \text{span}\left\{ \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \right\}$$

Fitting a logistic regression model

$$\log\left(\frac{\mu_i}{n_i}/(1-\frac{\mu_i}{n_i})\right) = \beta_1 + \beta_2 x_{i2}$$

where x_{i2} is the $i2$th entry in the design matrix X for $i = 1, 2, 3$, we obtain the estimates $\hat{\beta}_1 = 1.7757$, $\hat{\beta}_2 = -0.9840$ for the parameters β_1 and β_2 using the new algorithm (discussed in Section 4.4). The associated fitted values are shown in Table 3.2.

$\hat{\mu}$ (Response)	$\log\left(\frac{\hat{\mu}_i}{n_i}/(1-\frac{\hat{\mu}_i}{n_i})\right)$
15.8285	0.7917
20.3430	-0.1923
2.8285	-1.1764

Table 3.2: The fitted values for the logistic regression $\log\left(\frac{\mu_i}{n_i}/(1-\frac{\mu_i}{n_i})\right) = \beta_1 + \beta_2 x_{i2}$ for $i = 1, 2, 3$.

In Fienberg geometry the given data y can be represented as

$$\begin{bmatrix} 21 \\ 10 \\ 8 \end{bmatrix} = 26 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 6.5 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + 1.5 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

hence y has coefficients with respect to the new basis of

$$Ay = [26,\ -6.5,\ 1.5]^T$$

Similarly, the corresponding fitted value $\hat{\mu}$ has coefficients with respect to the new basis of

$$\begin{bmatrix} 15.8285 \\ 20.3430 \\ 2.8285 \end{bmatrix} = 26 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 6.5 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - 3.6715 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

or

$$A\hat{\mu} = [26,\ -6.5,\ -3.6715]^T$$

It is clear that Ay and $A\hat{\mu}$ have the same projection onto the sufficiency space. Figure 3.5 shows the result of fitting the logistic regression model with three observations in Fienberg geometry.

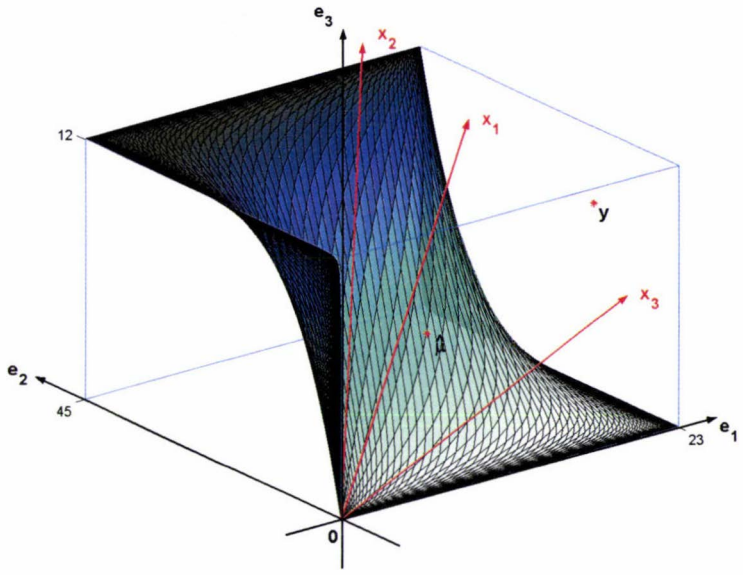


Figure 3.5: For the logistic regression model with three observations and Fienberg geometry, the saturated model space $\mathcal{F}_{\mathbf{R}}$ is a cube. The unsaturated model space $\mathcal{M}_{\mathbf{R}}$ is a curved surface within the hexahedron. In this example, the observation y and its fitted value $\hat{\mu}$ have the same projection onto each of the directions x_1 and x_2.

On the other hand, in the Haberman geometry, we consider the world transformed by the link function. Thus, a logit link function is applied to the proportion y_i/n_i for logistic regression, so $\log\left(\frac{y_i}{n_i}/(1-\frac{y_i}{n_i})\right)$, and the corresponding fitted value $\log\left(\frac{\hat{\mu}_i}{n_i}/(1-\frac{\hat{\mu}_i}{n_i})\right)$ can be expressed with respect to the new basis as

$$\begin{bmatrix} 2.3514 \\ -1.2528 \\ 0.6931 \end{bmatrix} = 2.2555 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0.8291 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + 0.9250 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

or

$$Ag(y) = [2.2555,\ -0.8291,\ 0.9250]^T$$

and

$$\begin{bmatrix} 0.7917 \\ -0.1923 \\ -1.1764 \end{bmatrix} = 1.7757 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0.9840 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - 0 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

or

$$Ag(\hat{\mu}) = [1.7757,\ -0.9840,\ 0]^T$$

This confirms that in the Haberman geometry $g(\hat{\mu})$ lies on the auxiliary space for all i. Figure 3.6 shows the result of fitting the logistic regression model with three observations in Haberman geometry.

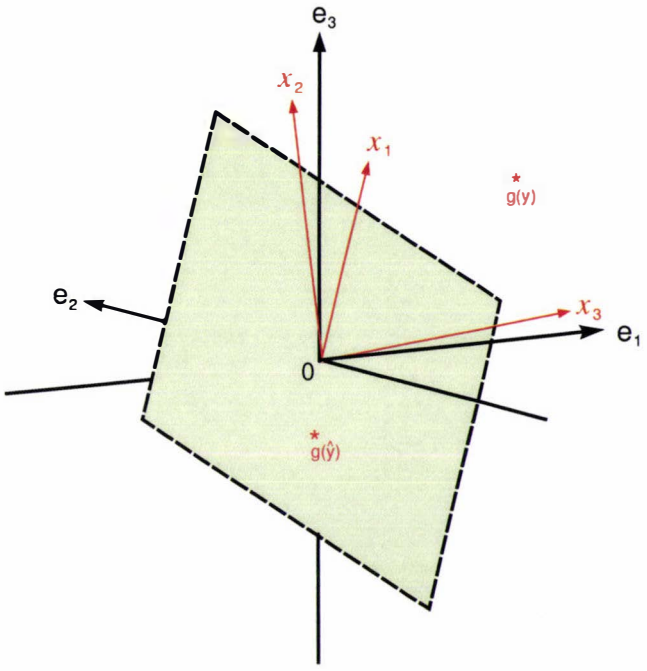


Figure 3.6: For the logistic regression model with three observations and Haberman geometry, the saturated model space $g(\mathcal{F}_{\mathbf{R}})$ is Euclidean space $\mathbf{R}^n$. The unsaturated model space $g(\mathcal{M}_{\mathbf{R}})$ is a plane in $\mathbf{R}^n$. The fitted value $g(\hat{\mu})$ has zero projection onto the direction x_3.

In the example we show the untransformed model space $\mathcal{M}_{\mathbf{R}}$ (see Figure 3.5)

and transformed model space $g(\mathcal{M}_{\mathbf{R}})$ (see Figure 3.6) for the logistic regression (a generalized linear model with logistic link). Using the same design matrix in the example above, the model spaces $\mathcal{M}_{\mathbf{R}}$ and $g(\mathcal{M}_{\mathbf{R}})$ are drawn for a generalized linear model with other canonical links in Figure 3.7 to Figure 3.10.

To summarize, the differential geometric approach to generalized linear models focuses on the families of density functions and relies on the likelihood function and the inner product defined by the Fisher information matrix in $\mathbf{R}^n$. In the fitting process, the residual vector is perpendicular to the tangent plane spanned by the scores functions under the inner product defined by the Fisher information matrix. On the other hand, the alternative geometric approach to generalized linear models depends on sufficiency and a partition of $\mathbf{R}^n$ using the new basis. Here all discussions are based on the mean sets. In the fitting process, before link transformation the coefficients with respect to the basis of the sufficiency space are preserved and after link transformation the coefficients with respect to the basis of the auxiliary space are zeroed.

3.6 Conclusion

In this chapter we have set up a geometric framework for generalized linear models in three stages. First, the statistical model is identified with a subset in Euclidean space. Two geometric objects relate to this subset: an untransformed one in the Fienberg geometry and a link transformed one in the Haberman geometry. Second, the observations are viewed as a vector in this space. Third, the whole space is split into a sufficiency space and an auxiliary space. In the Fienberg geometry the coefficients of the basis of the sufficiency space are preserved during model fitting,

while in the Haberman geometry the coefficients of the basis of the auxiliary space are zeroed. Now the geometries of linear models and loglinear models, discussed in Chapter 1 and Chapter 2, are special cases of the geometry of generalized linear models with identity link and log link respectively.

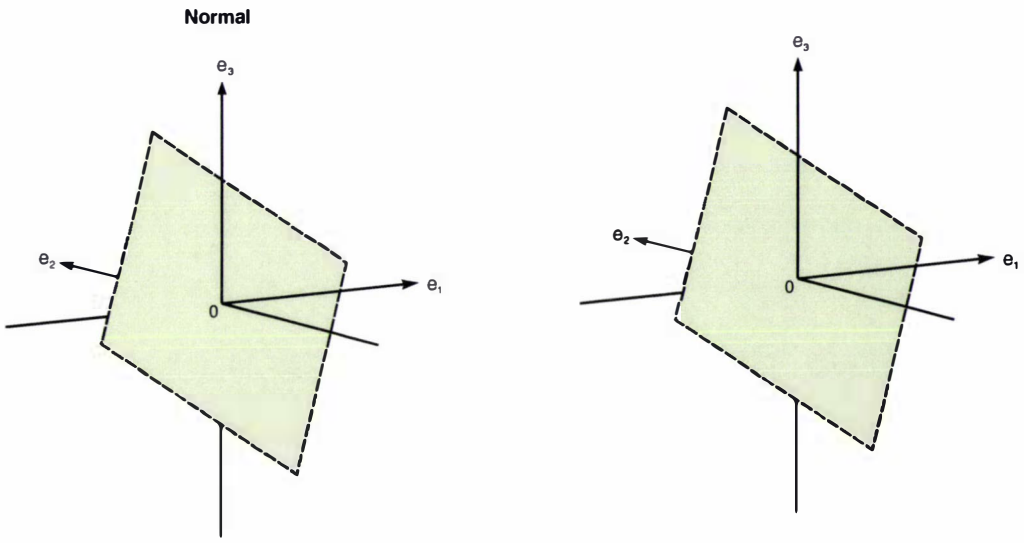


Figure 3.7: This graphic shows the model spaces $\mathcal{M}_{\mathbf{R}}$ (the left panel) and $g(\mathcal{M}_{\mathbf{R}})$ (the right panel) for the generalized linear model with identity link. Here elements of the observation vector Y independently follow a Normal distribution.

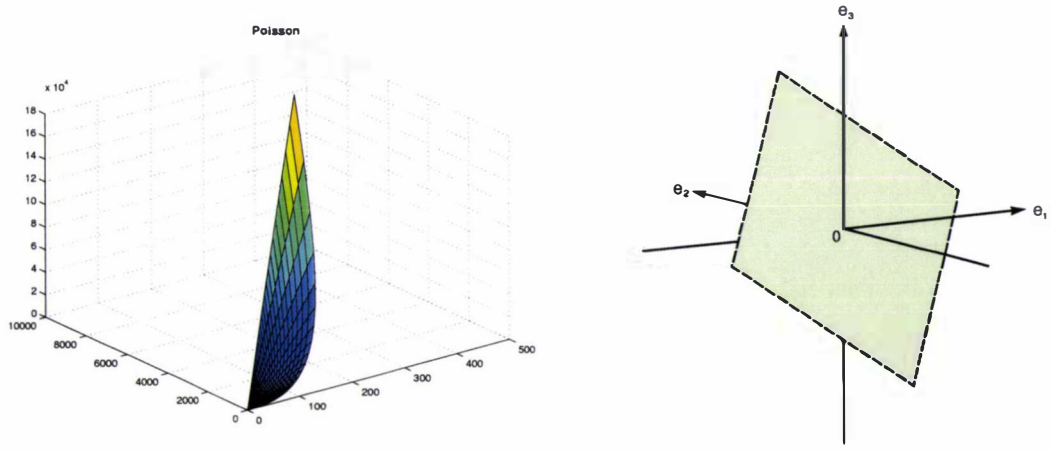


Figure 3.8: This graphic shows the model spaces $\mathcal{M}_{\mathbf{R}}$ (the left panel) and $g(\mathcal{M}_{\mathbf{R}})$ (the right panel) for the generalized linear model with log link. Here elements of the observation vector Y independently follow a Poisson distribution.

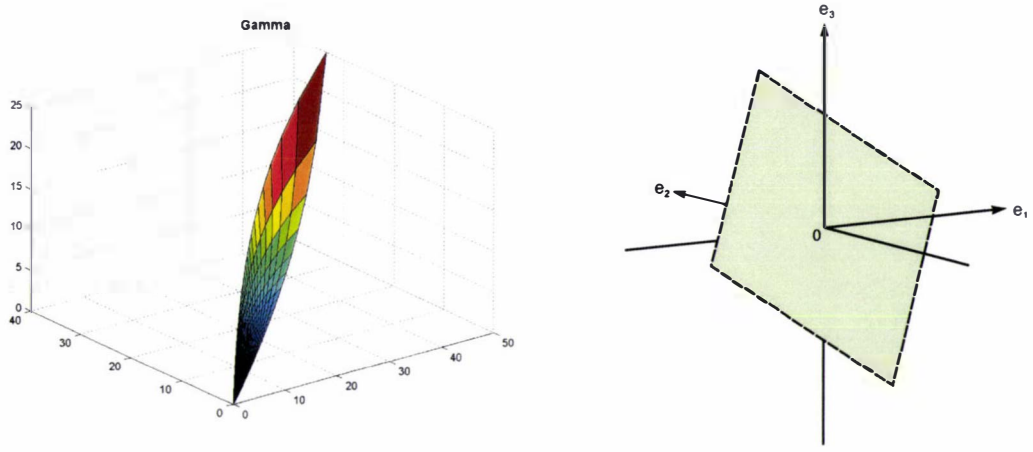


Figure 3.9: This graphic shows the model spaces $\mathcal{M}_{\mathbf{R}}$ (the left panel) and $g(\mathcal{M}_{\mathbf{R}})$ (the right panel) for the generalized linear model with reciprocal link and positive parameters. Here elements of the observation vector Y independently follow a Gamma distribution.

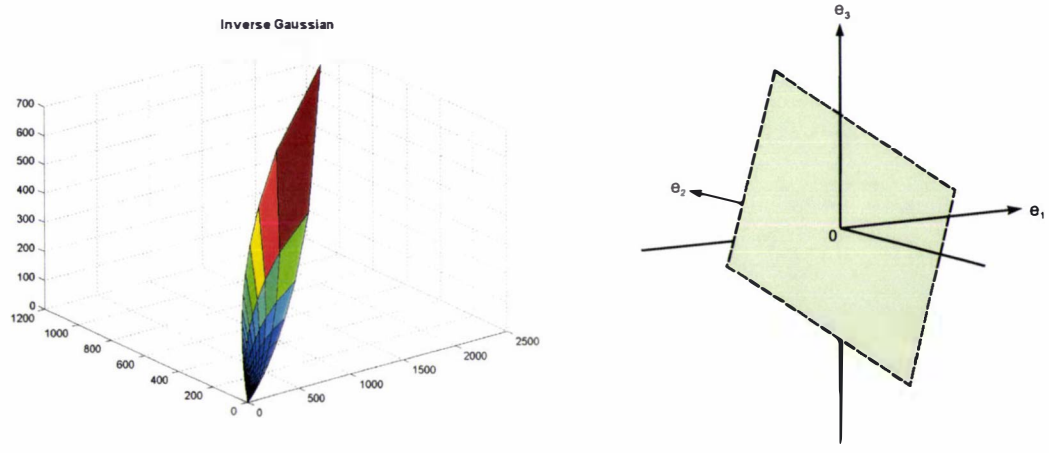


Figure 3.10: This graphic shows the model spaces $\mathcal{M}_{\mathbf{R}}$ (the left panel) and $g(\mathcal{M}_{\mathbf{R}})$ (the right panel) for the generalized linear model with squared reciprocal link and positive parameters. Here elements of the observation vector Y independently follow a inverse Gaussian distribution.

Chapter 4

A new algorithm for fitting GLMs

4.1 Introduction

A geometrical view of statistical models can assist in providing an overall understanding of statistical concepts, through allowing us to visualize a set of ideas. Geometry, however, can also catalyze the development of some new methodologies in statistics. In this chapter a new algorithm for fitting generalized linear models is constructed using the alternative geometry discussed in the last chapter.

Estimation of parameters in a generalized linear model is usually performed by the method of maximum likelihood, effected by a modified Newton-Raphson method named the scoring method. The scoring method was first considered by Fisher (1935) in the context of probit analysis and extended to find maximum likelihood estimation for generalized linear models by Nelder and Wedderburn in 1972. In the scoring method the estimates have to be obtained numerically by an iterative procedure. Recently, this can be interpreted using differential geometry, as discussed by Kass and Vos (1997).

In this chapter a new algorithm for fitting generalized linear models will be constructed using the alternative geometry. We first, however, discuss the scoring method using differential geometry in Section 2, then in Section 3 construct a new algorithm using the alternative geometry with a detailed example in Section 4. Section 5 links the two algorithms and is followed by a numerical comparison between the two methods in Section 6. This chapter is concluded by Section 7.

4.2 Geometry of the scoring method

In Chapter 3, we discussed the work of Kass and Vos (1997) in which they presented a geometry of generalized linear models using the language of differential geometry. This point of view leads to the scoring method. Here we briefly review the main points of Section 3.3 and summarize the algorithm suggested by Kass and Vos, then show that this algorithm turns out to be the scoring method.

In Section 3.3 we considered an n-dimensional set of exponential family density functions $\mathcal{F}$ and its q-dimensional subset $\mathcal{M}$ determined by the link function g and the design matrix X. The data y is inserted into the tangent space $T_f\mathcal{F}$, spanned by the score functions with respect to the mean parameter at $\mu = \mu(f)$ (the mean of f), by considering the vector with coordinates $y - \mu$. The maximum likelihood estimate $\hat{\mu}$ of μ is then obtained by satisfying

$$(y - \hat{\mu}) \perp T_{\hat{f}}\mathcal{M}$$

under the inner product defined by the Fisher information matrix in $T_{\hat{f}}\mathcal{F}$. Here $T_{\hat{f}}\mathcal{M}$, spanned by the score functions with respect to the parameter β, is the tangent space of $\mathcal{M}$ at $\hat{f}$. Note that $T_{\hat{f}}\mathcal{M}$ is a subset of $T_{\hat{f}}\mathcal{F}$. This leads to an algorithm for fitting

generalized linear models.

The algorithm starts with the initial density function estimate $f_0 \in \mathcal{M}$. We then obtain the initial mean estimate $\mu_0 = \mu(f_0)$, the mean of f_0, and initial parameter estimate β_0 of β through $\beta_0 = (X^T X)^{-1} X^T g(\mu_0)$. Since $f_0 \in \mathcal{M}$, so $f_0 \in \mathcal{F}$, there are two tangent spaces at f_0, namely $T_{f_0}\mathcal{M}$ for $\mathcal{M}$ and $T_{f_0}\mathcal{F}$ for $\mathcal{F}$. We insert the data y by constructing a vector $y - \mu_0 \in T_{f_0}\mathcal{F}$, and then project $y - \mu_0$ onto $T_{f_0}\mathcal{M}$ orthogonally to obtain the projection $v_0 \in T_{f_0}\mathcal{M}$. Next, we map $v_0 \in T_{f_0}\mathcal{M}$ into $\mathcal{M}$ using a mapping $R : v \mapsto R(v)$, where $v \in T_f\mathcal{M}$ and $R(v) \in \mathcal{M}$, to provide a new estimate of f. The mapping R is made up of the following steps

Step 1 Start with $v_0 \in T_f\mathcal{M}$, and then find an estimate β_1 of β using $\beta_1 = \beta_0 + v_0$,

Step 2 Obtain the new estimate μ_1 of μ by $\mu_1 = g^{-1}(X\beta_1)$,

Step 3 Achieve the new estimate $f_1 \in \mathcal{M}$ of f through μ_1. Here we first obtain an estimate θ_1 by $\theta_1 = b'^{-1}(\mu_1)$, and then gain

$$f_1(y; \theta_1, \phi) = \exp\{y^T\theta_1 - b(\theta_1) + c(y, \phi)\}$$

where ϕ is a dispersion parameter, assumed given or estimated.

The algorithm is repeated, using f_1 now instead of f_0. The maximum likelihood estimate $\hat{\mu}$ of μ will be found when $(y - \hat{\mu}) \perp T_{\hat{f}}\mathcal{M}$. The process is shown in Figure 4.1.

Formally, the algorithm can be presented as follows.

Kass and Vos algorithm

Step 1 Set $k = 0$. Take an initial estimate f_0 of f in $\mathcal{M}$, and produce the mean estimate μ_0 by $\mu_0 = \mu(f_0)$ and the parameter estimate β_0 by $\beta_0 = (X^T X)^{-1} X^T g(\mu_0)$.

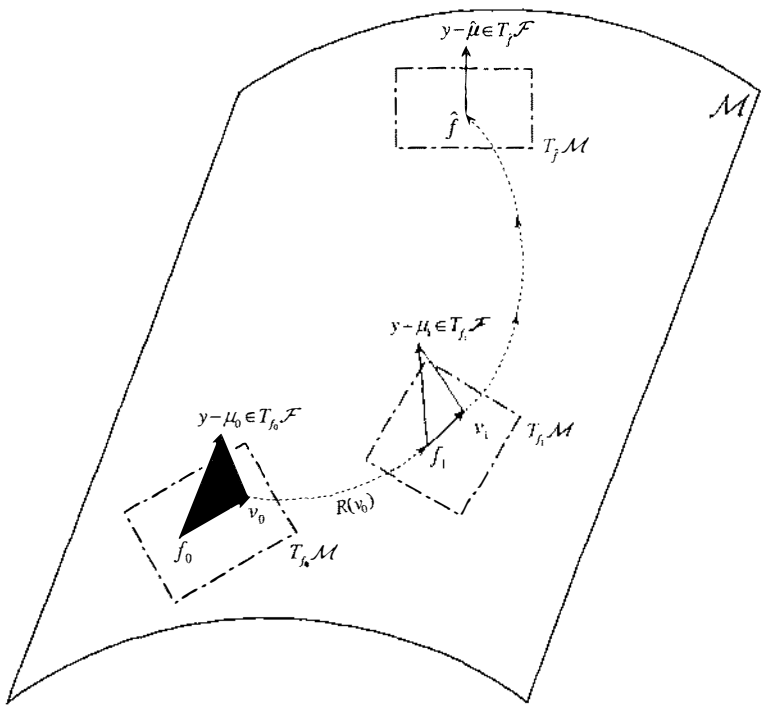


Figure 4.1: The graphic shows the process of fitting generalized linear models using the Kass and Vos algorithm. The algorithm starts with the initial estimate f_0 of f, then projects $y - \mu_0 \in T_{f_0}\mathcal{F}$ (where μ_0 is the mean of f_0) onto the tangent space $T_{f_0}\mathcal{M}$ orthogonally. The projection v_0 is then taken back into $\mathcal{M}$ by the mapping R, giving f_1. The algorithm is repeated using f_1 instead of f_0. The maximum likelihood estimate $\hat{\mu}$ of μ is found when $(y - \hat{\mu}) \perp T_{\hat{f}}\mathcal{M}$.

Step 2 Project $y - \mu_k \in T_{f_k}\mathcal{F}$ onto the tangent space $T_{f_k}\mathcal{M}$ orthogonally at f_k under the inner product defined by the Fisher information matrix in $T_{f_k}\mathcal{F}$. Map this projection v_k back into $\mathcal{M}$ by the mapping $R : v \mapsto R(v)$ to obtain f_{k+1}.

Step 3 If a stopping criterion is met, stop. Otherwise, increment k and return to Step 2.

This is now shown to provide the scoring algorithm.

For a generalized linear model with canonical link and $a(\phi) = 1$, we have

$$\mathcal{F} = \left\{ f : f(y; \theta, \phi) = \exp\{y^T\theta - b(\theta) + c(y, \phi)\} \text{ for some } \theta \in \Theta \right\}$$

where Θ is the n-dimensional natural parameter space, and

$$\mathcal{M} = \left\{ f \in \mathcal{F} : \mu(f) = g^{-1}(X\beta),\ \beta \in B \right\}$$

where $\mu : f \mapsto \mu(f)$ from $\mathcal{F}$ to $\mathbf{R}^n$ provides the mean of f, g is the canonical link function, and B is the q-dimensional parameter space.

Then for any $f \in \mathcal{M}$ the tangent space for $\mathcal{F}$ is given by

$$T_f\mathcal{F} = \text{span}\{U_i\}_{i=1}^n$$

where

$$\begin{aligned} U_i &= \frac{\partial l(\mu; Y)}{\partial \mu_i} \qquad \text{(where } l(\mu; Y) = \log f(\mu; Y)\text{)} \\ &= \frac{\partial \left[Y^T g(\mu) - b(g(\mu)) + c(Y, \phi) \right]}{\partial \mu_i} \\ &= \left[Y_i - b'(g(\mu_i)) \right] g'(\mu_i) \\ &= (Y_i - \mu_i) g'(\mu_i) \qquad \text{(since } \mu_i = b'(\theta_i) \text{ and } \theta_i = g(\mu_i)\text{)} \end{aligned}$$

for $i = 1, 2, \ldots, n$.

For any $f \in \mathcal{M}$ the tangent space for $\mathcal{M}$ is given by

$$T_f\mathcal{M} = \text{span}\{V_j\}_{j=1}^q$$

where

$$\begin{aligned} V_j &= \frac{\partial l(\beta; Y)}{\partial \beta_j} \\ &= \sum_{i=1}^n \frac{\partial l(\mu; Y)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^n \frac{\partial \mu_i}{\partial \eta_i} x_{ij} U_i \qquad \text{(since } \mu = g^{-1}(\eta) \text{ and } \eta = X\beta\text{)} \end{aligned}$$

Specifically,

$$V_j = [\frac{\partial \mu_1}{\partial \eta_1} x_{1j}, \frac{\partial \mu_2}{\partial \eta_2} x_{2j}, \ldots, \frac{\partial \mu_n}{\partial \eta_n} x_{nj}]^T$$

with respect to the basis $\{U_1, U_2, \ldots, U_n\}$ for $j = 1, 2, \ldots, q$.

In other words, $T_f\mathcal{M}$ is spanned by the column vectors in the matrix $F = WX$ where W is the diagonal matrix with elements $w_i = \partial \mu_i / \partial \eta_i$ for $i = 1, 2, \ldots, n$ on the main diagonal. It is clear that $T_f\mathcal{M} \subset T_f\mathcal{F}$.

Next, in $T_f\mathcal{F}$ an inner product is defined by

$$\langle a, b \rangle = a^T W^{-1} b \qquad \text{for } a,\ b \in T_f\mathcal{F}$$

where W^{-1} is the Fisher information matrix for the mean parameter at $\mu = \mu(f)$.

Now, the maximum likelihood estimate $\hat{\mu}$ of μ can be obtained in the following three steps.

Step 1 Set $k = 0$. Take any $\mu_0 \in \mathcal{M}_{\mathbf{R}}$, the mean set of any $f \in \mathcal{M}$, and define f_0 by $\mu_0 = \mu(f_0)$ and β_0 by $\beta_0 = (X^T X)^{-1} X^T g(\mu_0)$.

Step 2 Project $y - \mu_k$ onto the tangent space $T_{f_k}\mathcal{M}$ at f_k, spanned by the column vectors in the matrix F, orthogonally under the inner product defined by $\langle a, b \rangle = a^T W_k^{-1} b$ for any $a, b \in T_{f_k}\mathcal{F}$. Doing this we obtain the projection

$$v_k = (F_k^T W_k^{-1} F_k)^{-1} F_k^T W_k^{-1} (y - \mu_k) \in T_f\mathcal{M}$$

Update the estimate of β by

$$\beta_{k+1} = \beta_k + v_k$$

and then obtain μ_{k+1} as the value of mean μ evaluated at $\beta = \beta_{k+1}$.

Step 3 If a stopping criterion is met, stop. Otherwise, increment k and return to Step 2.

To summarize, the algorithm with respect to the parameter β is

$$\begin{aligned} \beta_{k+1} &= \beta_k + (F_k^T W_k^{-1} F_k)^{-1} F_k^T W_k^{-1} (y - \mu_k) \\ &= \beta_k + (X^T W_k X)^{-1} X^T (y - \mu_k) \end{aligned}$$

the familiar scoring method. Note that F_k and W_k are F and W estimated at $\beta = \beta_k$ respectively.

4.3 The new algorithm

In this section, we establish a new algorithm for fitting generalized linear models with canonical link using the alternative geometry. Here, we first consider a theorem for characterizing maximum likelihood estimation for generalized linear models with canonical link. This theorem and the commutative diagram discussed in Chapter 3 serve as motivation for the new algorithm which is, then, constructed in the alternative geometric framework. Next, a simplification of the new algorithm is given with its convergence discussed.

Theorem 4.1. *For a generalized linear model with canonical link, the maximum likelihood estimate $\hat{\mu}$ of the mean vector μ is uniquely determined by*

i) $\hat{\mu} \in \mathcal{M}_{\mathbf{R}}$,

ii) $X^T \hat{\mu} = X^T y$.

Proof. The exponential family of distributions satisfy enough regularity conditions to ensure that the maximum likelihood estimate $\hat{\mu}$ is unique for a generalized linear model (Cox and Hinkley 1974, p.245). This solution is determined by the following equation, for a generalized linear model with canonical link,

$$\frac{\partial \ell}{\partial \beta} = \frac{X^T(y-\mu)}{a(\phi)} = 0 \qquad \text{(see Theorem 3.1 ii))} \tag{4.1}$$

where $a(\phi)$ is known and identical for all observations. This equation is solved by the scoring method to obtain the maximum likelihood estimate $\hat{\beta}$ of β, then $\hat{\mu} = g^{-1}(X\hat{\beta}) \in \mathcal{M}_{\mathbf{R}}$, the model space. Thus i) holds.

Furthermore, $\hat{\mu}$ is the solution of (4.1), so $\hat{\mu}$ should satisfy $X^T(y-\mu) = 0$. The equation $X^T\hat{\mu} = X^T y$, then, is required. Thus ii) holds. □

Theorem 4.1 indicates that the maximum likelihood estimate for a generalized linear model with canonical link is unique and both satisfies the model and matches the observations in the sufficient statistics.

A new algorithm for fitting generalized linear models with canonical link is suggested by Theorem 4.1 and the commutative diagram in Chapter 3. To find the estimate $\hat{\mu}$, the unique point which is the intersection of the model space $\mathcal{M}_{\mathbf{R}}$ and the sufficiency affine plane $\mathcal{T}$, the algorithm starts with an initial estimate μ_0. To match sufficient statistics, we project μ_0 onto the affine sufficiency plane $\mathcal{T}$ orthogonally to obtain a point μ_T which matches the observation vector y in the sufficient statistics but may not satisfy the model, that is, μ_T may not be in $\mathcal{M}_{\mathbf{R}}$ and thus $g(\mu_T)$ may not be in $g(\mathcal{M}_{\mathbf{R}})$. To satisfy the model, we move to the link transformed point $g(\mu_T)$, then project $g(\mu_T)$ onto the transformed model space $g(\mathcal{M}_{\mathbf{R}})$ non-orthogonally (a weighted least squares) to obtain a point $g(\mu_M) \in g(\mathcal{M}_{\mathbf{R}})$. We then inverse link

transform to obtain the point μ_M in $\mathcal{M}_{\mathbf{R}}$. The point μ_M satisfies the model but may not match y in the sufficient statistics, that is μ_M may not be in $\mathcal{T}$. Let $\mu_1 = \mu_M$, then start again with μ_1 in the role of μ_0. The process is shown in Figure 4.2.

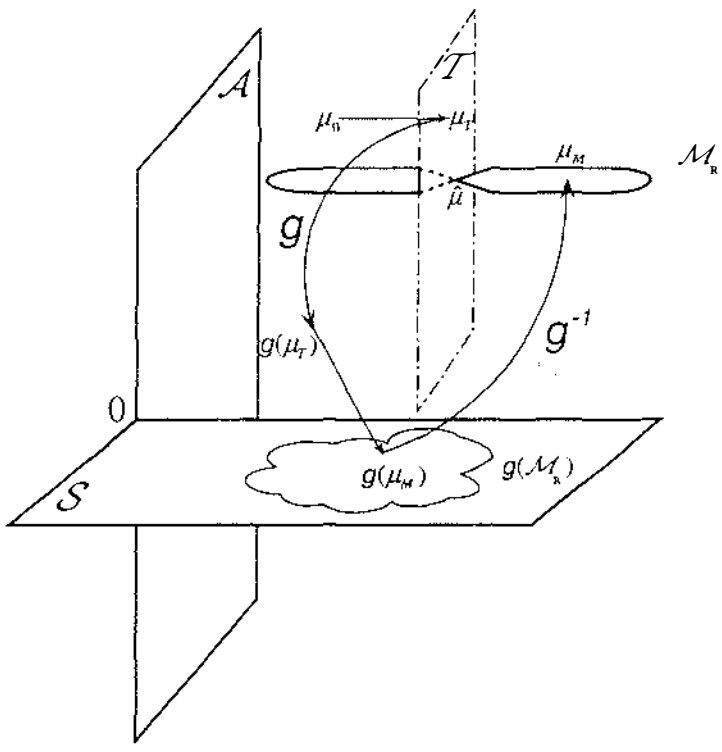


Figure 4.2: The graphic shows the process of fitting a generalized linear model using the new algorithm. The algorithm starts with the initial estimate μ_0 of μ, then projects $\mu_0 \in \mathcal{F}_{\mathbf{R}}$ onto the sufficiency affine plane $\mathcal{T}$ orthogonally to obtain point μ_T. After the link transformation we project the point $g(\mu_T)$ onto $g(\mathcal{M}_{\mathbf{R}})$ non-orthogonally to obtain the point $g(\mu_M)$, and then find the point μ_M by back transformation. The algorithm starts again using $\mu_1 = \mu_M$ instead of μ_0.

Note that in the algorithm each projection is with respect to the new basis $\{x_1, x_2, \ldots, x_n\}$, while each canonical link transformation is with respect to the standard basis. Informally, projection onto $\mathcal{T}$ improves the first q coordinates while projection onto $g(\mathcal{M}_{\mathbf{R}})$ improves the last $n-q$ coordinates. For linear models, generalized linear models with identity link, $\mathcal{M}_{\mathbf{R}}$ and $g(\mathcal{M}_{\mathbf{R}})$ fuse together, the non-orthogonal

projection onto $\mathcal{S}$ becomes an orthogonal projection. Thus both model space and sufficient statistics are satisfied by the estimate of μ in the first iteration of the algorithm. Here we should point out that initially projection onto $\mathcal{S}$ in the algorithm was performed orthogonally, but the algorithm then only worked for some cases of generalized linear models with non-identity link. After using non-orthogonal projection onto $\mathcal{S}$, a weighted least squares with the weight chosen to be the variance of estimate of μ, the algorithm has been found to work for all cases.

Denote the projection onto the sufficiency affine plane by P_T, the projection onto the sufficiency space by P_S and the change of basis matrix by A. The algorithm is now described.

A new fitting algorithm

Step 1 Set $k = 0$. Take $\mu_0 = y$, the observation vector.

Step 2 Set $\mu_{k+1} = g^{-1}A^{-1}P_S A g A^{-1} P_T A(\mu_k)$.

Step 3 If a stopping criterion is met, stop. Otherwise, increment k and return to Step 1.

To consider the new algorithm in more detail, we can represent it algebraically as

$$\begin{aligned} \mu_{k+1} &= g^{-1}(X(X^T W_k X)^{-1} X^T W_k z_k) \\ \text{with} \quad z_k &= g\left\{ A^{-1}\left[\begin{pmatrix} 0 & 0 \\ 0 & I_{n-q} \end{pmatrix} A\mu_k + \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} Ay \right] \right\} \end{aligned} \tag{4.2}$$

where

μ_k denotes the kth approximation for the maximum likelihood estimate $\hat{\mu}$.

W is the $n \times n$ diagonal matrix with $w_i = \partial\mu_i/\partial\eta_i$ for all i on the main diagonal, and W_k is W estimated at $\mu = \mu_k$,

g and g^{-1} represent the canonical link function and its inverse,

$A = [X\ X_c]^{-1}$ is the change of basis matrix with size $n \times n$ generated by the design matrix X through the variation of the Gram-Schmidt process, so X_c is the complementary matrix of X in $\mathbf{R}^n$,

I_q and I_{n-q} are the identity matrixes of size $q \times q$ and $(n-q) \times (n-q)$ respectively.

The interpretation of the algebraic representation of the new algorithm is shown in the following table.

Operation in Step 2 of the new algorithm	Algebraic representation
μ_k	μ_k
$A(\mu_k)$	$A\mu_k$
$P_T A(\mu_k)$	$\begin{pmatrix} 0 & 0 \\ 0 & I_{n-q} \end{pmatrix} A\mu_k + \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} Ay$
$A^{-1}P_T A(\mu_k)$	$A^{-1}\left[\begin{pmatrix} 0 & 0 \\ 0 & I_{n-q} \end{pmatrix} A\mu_k + \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} Ay\right]$
$gA^{-1}P_T A(\mu_k)$	$z_k = g\left\{A^{-1}\left[\begin{pmatrix} 0 & 0 \\ 0 & I_{n-q} \end{pmatrix} A\mu_k + \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} Ay\right]\right\}$
$P_S A g A^{-1} P_T A(\mu_k)$	$(X^T W_k X)^{-1} X^T W_k z_k$
$A^{-1} P_S A g A^{-1} P_T A(\mu_k)$	$X(X^T W_k X)^{-1} X^T W_k z_k$
$g^{-1} A^{-1} P_S A g A^{-1} P_T A(\mu_k)$	$g^{-1}(X(X^T W_k X)^{-1} X^T W_k z_k)$

This table just shows the operations of Step 2 in the new algorithm and the associated algebraic representation. To fully understand the contents in the table reader is referred to the next section, where a detailed example is given.

In the event that the design matrix X is orthonormal, we now consider a simpler form of the new algorithm. Since X is orthonormal, then the matrix $[X\ X_c]$, extended

by X through the variation of the Gram-Schmidt process, is also orthonormal. Now the change of basis matrix is

$$A = \begin{bmatrix} X^T \\ X_c^T \end{bmatrix}$$

where

$$X^T X = I_q \qquad \text{and} \qquad X_c X_c^T = I_n - XX^T$$

Then, (4.2) becomes

$$\begin{aligned} z_k &= g\left[[X\ X_c] \begin{pmatrix} 0 & 0 \\ 0 & I_{n-q} \end{pmatrix} \begin{bmatrix} X^T \\ X_c^T \end{bmatrix} \mu_k + [X\ X_c] \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \begin{bmatrix} X^T \\ X_c^T \end{bmatrix} y \right] \\ &= g\left[X_c X_c^T \mu_k + XX^T y \right] \\ &= g\left[(I_n - XX^T)\mu_k + XX^T y \right] \\ &= g\left[\mu_k + XX^T (y - \mu_k) \right] \end{aligned}$$

Thus, the algorithm simplifies to

$$\begin{aligned} \mu_{k+1} &= g^{-1}(X(X^T W_k X)^{-1} X^T W_k z_k) \\ \text{with} \quad z_k &= g\left[\mu_k + XX^T (y - \mu_k) \right] \end{aligned}$$

where column vectors in the design matrix X are orthonormalized.

The change of basis in $\mathbf{R}^n$ has an influence on the coordinates of a vector in $\mathbf{R}^n$, but has no influence on the vector itself. Thus, geometrically, for the simplified algorithm and non-simplified algorithm, the fitting process will be the same, but the coordinates of a vector with respect to the new basis will be different. Since the fitted value $\hat{\mu}$ is with respect to the standard basis, the simplified algorithm and non-simplified algorithm have the same result of $\hat{\mu}$. The estimate $\hat{\beta}$, however, is with respect to the new basis, so the simplified algorithm and non-simplified algorithm

have a different result for $\hat{\beta}$. Thus we can use the simplified algorithm to find $\hat{\beta}$ for a general case through the following steps.

1. Orthonormalize the design matrix X
2. Obtain the fitted value $\hat{\mu}$ using the simplified algorithm
3. Find the estimate $\hat{\beta}$ using weighted least squares method with the design matrix X and the weight estimated by $\hat{\mu}$.

The maximum likelihood estimate $\hat{\mu}$ of the mean vector is a fixed point of $h = g^{-1}A^{-1}P_S AgA^{-1}P_T A$. In the next result we describe conditions under which a fixed point exists. We give conditions which ensure that the the fixed point is unique and that the algorithm converges to the fixed point.

Theorem 4.2. *Let* $h = g^{-1}A^{-1}P_S AgA^{-1}P_T A$. *We have*

i) If the domain of h *is homeomorphic to a sphere, then* h *has a fixed point.*

ii) If h *is contractive (so* $\|h(x) - h(y)\| \leq r\|x - y\|$ *for some* $0 < r < 1$*) then*

a) h *has a unique fixed point* z,

b) $\lim_k h^k(x) = z$ *for any* x *in the domain of* h, *and*

c) $z = \hat{\mu}$.

Proof. Statement i) is an almost immediate consequence of Brouwer's Fixed Point Theorem (Simmons, p.338). Statement ii) a) follows from the lemma of Schauder's Fixed Point Theorem (Simmons, p.338), while ii) b) follows from the proof of the lemma. Since $\hat{\mu}$ is a fixed point of h and the fixed point is unique, so ii) c) follows. □

4.4 A detailed example

To illustrate the new algorithm for fitting generalized linear models we consider the fitting process for the logistic regression example described in Chapter 3. The artificial data set, with three observations, is shown in Table 4.1.

y_i (Response)	n_i (Total)	x_i (Covariate)	$\log\left(\frac{y_i}{n_i}/(1-\frac{y_i}{n_i})\right)$
2	23	1	-2.3514
20	45	2	-0.2231
8	12	3	0.6931

Table 4.1: The artificial logistic regression data set, with three observations.

We fit the logistic regression model

$$\log\left(\frac{\mu_i}{n_i}/(1-\frac{\mu_i}{n_i})\right) = \beta_1 + \beta_2 x_{i2} \qquad \text{for} \qquad i = 1, 2, 3$$

to the data in Table 4.1 using the new algorithm. The first iteration of the algorithm is shown using eight steps in Table 4.2. Specifically,

Step 1 Take the initial estimate of μ to be the observation vector y.

Step 2 Consider the point y with respect to the new basis, Ay.

Step 3 Project the point Ay orthogonally onto the sufficiency affine plane $\mathcal{T}$ to give point $A\mu_T$.

Step 4 Consider $A\mu_T$ with respect to the standard basis, so giving μ_T (here μ_T=y).

Step 5 Take the logistic transformation of μ_T to obtain the point $g(\mu_T)$.

Step 6 Project the point $g(\mu_T)$ non-orthogonally onto the transformed model space $g(\mathcal{M}_{\mathbf{R}})$ using weighted least squares to obtain the point $Ag(\mu_M)$.

Step 7 Consider $Ag(\mu_M)$ with respect to the standard basis, giving $g(\mu_M)$.

Step 8 Take a back logistic transformation of $g(\mu_M)$ to obtain μ_M.

In the next iteration, the algorithm starts with the estimate μ_M of μ.

Iteration	Step	Item	Coordinates
1	1	y	(21.0000, 10.0000, 8.0000)
	2	Ay	(26.0000, -6.5000, 1.5000)
	3	$A\mu_T$	(26.0000, -6.5000, 1.5000)
	4	μ_T	(21.0000, 10.0000, 8.0000)
	5	$g(\mu_T)$	(2.3514, -1.2528, 0.6931)
	6	$Ag(\mu_M)$	(0.7319, -0.4957, 0.0000)
	7	$g(\mu_M)$	(0.2362, -0.2595, -0.7553)
	8	μ_M	(12.8518, 19.5964, 3.8360)

Table 4.2: The table shows all steps in the first iteration of the fitting process, for logistic regression with three observations, using the new algorithm.

The whole fitting procedure is shown in Table 4.3, including coordinates for the point $A\mu_T$ in the sufficiency affine plane $\mathcal{T}$ and the point $Ag(\mu_M)$ in the transformed model space $g(\mathcal{M}_{\mathrm{R}})$ (with respect to the new basis) in each iteration. From Table 4.3 we see that the set of points $A\mu_T$ has the same coordinates on the basis of the sufficiency space, while the set of points $Ag(\mu_M)$ has zero coordinate on the basis of auxiliary space. The locus of the fitting process is demonstrated in the untransformed context in Figure 4.3.

4.5 Link between the two algorithms

In the new algorithm, we use non-orthogonal projection (that is, weighted least squares) onto the log-transformed model space $g(\mathcal{M}_{\mathrm{R}})$, which is the same as the

Iteration	$A\mu_T$	$Ag(\mu_M)$
1	(26.0000, -6.5000, 1.5000)	(0.7319, -0.4957, 0)
2	(26.0000, -6.5000, -9.1876)	(1.7754, -0.9839, 0)
3	(26.0000, -6.5000, -8.9939)	(1.7757, -0.9840, 0)
4	(26.0000, -6.5000, -8.9933)	(1.7757, -0.9840, 0)
5	(26.0000, -6.5000, -8.9933)	(1.7757, -0.9840, 0)

Table 4.3: The table shows the coordinates, with respect to the new basis, of points $A\mu_T$ in the sufficiency affine plane and points $Ag(\mu_M)$ in the transformed model space $g(\mathcal{M}_{\mathbf{R}})$.

scoring method. For this reason, there is a natural connection between the two algorithms, summarized in the following theorem.

Theorem 4.3. *For a generalized linear model with canonical link function and orthonormal design matrix, the linearization of the new algorithm is the same as the scoring method.*

Proof. When the design matrix X is orthonormal, the new algorithm has the form

$$\mu_{k+1} = g^{-1}(X(X^T W_k X)^{-1} X^T W_k z_k) \tag{4.3}$$

$$\text{with} \quad z_k = g\left[\mu_k + XX^T(y - \mu_k)\right] \tag{4.4}$$

where $\mu_k = g^{-1}(X\beta_k)$.

Approximate the right-hand-side of (4.4) with its first order Taylor series for g to give

$$z_k \approx X\beta_k + {W_k}^{-1} XX^T(y - \mu_k)$$

Substitute $z_k = X\beta_k + {W_k}^{-1} XX^T(y - \mu_k)$ and $\mu_{k+1} = g^{-1}(X\beta_{k+1})$ into (4.3), the algorithm becomes

$$\beta_{k+1} = \beta_k + (X^T W_k X)^{-1} X^T(y - \mu_k)$$

which is the scoring method. Thus, the result holds. □

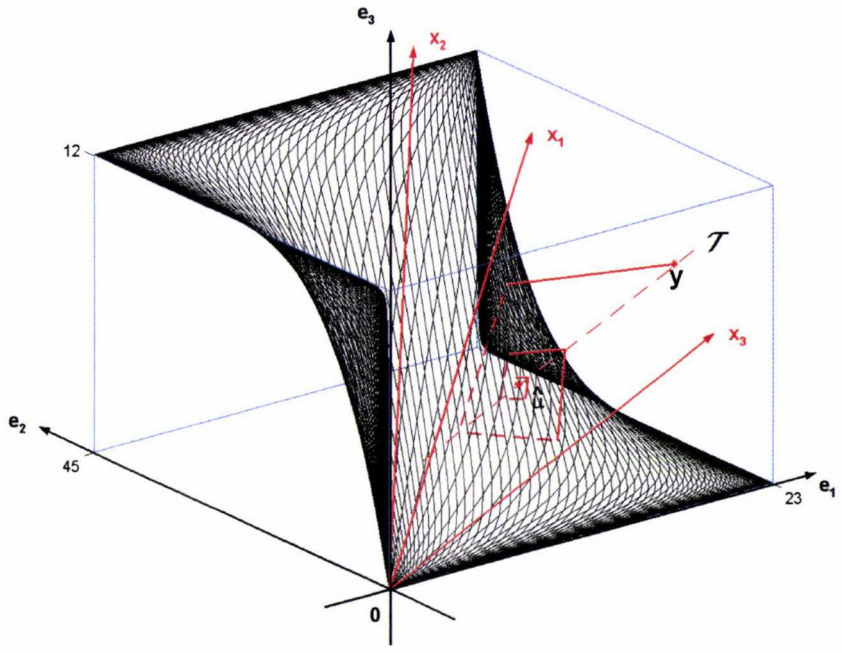


Figure 4.3: The figure shows the locus of critical points induced in fitting the logistic regression with three observations in the untransformed context. There two sets of points on the locus, one on the model space $\mathcal{M}_{\mathbf{R}}$ and the other on the sufficiency affine plane $\mathcal{T}$. Note that for showing clear effects we use the orthogonal projection instead of the non-orthogonal projection in the fitting process.

4.6 Numerical comparison of the two algorithms

To compare the two algorithms, we show some numerical results in Table 4.4. This table presents the number of iterations needed for convergence for a variety of models using the built-in function 'glmfit' in Matlab, the new algorithm (see Appendix A) and the 'Genmod' procedure in SAS. For all three methods convergence is achieved when the value of the norm of the difference between successive estimates of the parameter β is less than 10^{-6}. In the 'Models' column of the table, x_i indicates a covariate, f_i represents a factor and $f_i f_j$ denotes the interaction between factors f_i and f_j. In the 'Link' column the associated link function is specified for each model.

Table 4.4 shows that there is not much difference among three methods for simple models, but for complex models (Model 5 and 12) the new algorithm seems to take more iterations to converge than do other methods.

	Models	Link	Matlab	Geometric	SAS
1	$y/N = x$	Logit	6	6	5
2	$y/N = x$	Logit	5	6	5
3	$y/N = f_1 + f_2$	Logit	5	5	4
4	$y/N = x$	Logit	6	7	5
5	$y = x_1 + x_2 + f_1 + f_2 + f_1 f_2$	Logit	7	12	8
6	$y = f + x_1 + x_2$	Logit	6	5	4
7	$y/N = x_1 + x_2 + x_3$	Logit	7	8	6
8	$y = f_1 + f_2 + f_3$	Log	7	5	6
9	$y = x$	Log	7	6	6
10	$y = x$	Log	6	5	5
11	$y = f_1 + f_2$	Log	7	6	6
13	$y = x + f_1 + f_2$	Log	5	5	4
12	$y = f_1 + f_2 + f_3 + f_1 f_3 + f_2 f_3$	Log	7	14	5
14	$y = x + f + xf$	Reciprocal	8	6	6
15	$y = f_1 + f_2 + f_3$	Reciprocal	7	6	6

Table 4.4: This table shows the number of iterations needed for convergence for various models using three methods: the built-in function 'glmfit' in Matlab, the new algorithm and the 'Genmod' procedure in SAS.

To compare the scoring method and the new algorithm in more detail, in Table 4.5 we show the flops, an approximate number of floating point operations, needed for convergence for various models using Matlab functions `Sglmfit` and `Nglmfit` (see Appendix A). It seems that the performance of the new algorithm is worse than the scoring method.

The data sources for the models displayed in Table 4.4 and Table 4.5 are given in Appendix B. For example, Model 5 has the form

$$\begin{aligned} \texttt{Pain}(y) \;=\; & \texttt{Age}(x_1) + \texttt{Duration}(x_2) + \texttt{Treatment}(f_1) + \texttt{Sex}(f_2) \\ & + \texttt{Treatment*Sex}(f_1 f_2) \end{aligned}$$

	Models	Link	Scoring method	Geometric
1	$y/N = x$	Logit	7649	10145
2	$y/N = x$	Logit	7219	10137
3	$y = f_1 + f_2$	Logit	3100	3820
4	$y/N = x$	Logit	1325800	1471684
5	$y = x_1 + x_2 + f_1 + f_2 + f_1 f_2$	Logit	1831736	2729886
6	$y = f + x_1 + x_2$	Logit	1813484	2011354
7	$y = x_1 + x_2 + x_3$	Logit	3876210	4414085
8	$y = f_1 + f_2 + f_3$	Log	28524611	30326143
9	$y = x$	Log	1286880	1405419
10	$y = x$	Log	7259	8814
11	$y = f_1 + f_2$	Log	30389	36664
12	$y = f_1 + f_2 + f_3 + f_1 f_3 + f_2 f_3$	Log	21243	45302
13	$y = x + f_1 + f_2$	Log	87249	110289
14	$y = x + f + xf$	Reciprocal	293834	351018
15	$y = f_1 + f_2 + f_3$	Reciprocal	16589470	19400529

Table 4.5: This table shows the number of flops (an approximate number of floating point operations) needed for convergence for various models using two methods: the scoring method and the new algorithm.

where `Pain` is a binary response variable. The data comes from "SAS Institute Inc. (1999). *SAS OnlineDoc.* Example 39.3".

4.7 Conclusions

In this chapter a new algorithm for fitting generalized linear models with canonical link is constructed using the alternative geometry. This algorithm depends on sufficiency rather than the likelihood function, and uses two projections alternately, orthogonal projection onto the sufficiency affine plane and non-orthogonal projection onto the transformed model space. In the process, we match sufficient statistics and the model space iteratively until convergence. A linearization of the new algorithm yields the scoring method.

Chapter 5

The geometry of conditional independence statements

5.1 Introduction

In Chapter 2 the geometry of loglinear models for contingency tables has been discussed using two distinct approaches, the first contributed by Fienberg and the second by Haberman. A joint probability for a contingency table is decomposed additively with respect to the new basis before log-transformation (in Fienberg geometry) and after log-transformation (in Haberman geometry). This leads us to ignore the special features of loglinear models relating to their interpretation in terms of conditional independence statements. In this chapter we describe a geometric setting for a subset of loglinear models, those which are the intersection of a finite set of conditional independence (CI) statements. This is of interest in itself and also provides a framework for understanding the workings of iterative proportional fitting approaches.

For a given set of categorical variables $S = \{X_1, X_2, \ldots, X_m\}$ with cell index

$i = (i_1, i_2, \ldots, i_m)$, the symbolic form of the saturated loglinear model is

$$\log \pi_i = \sum_{A \subseteq S} \lambda^A_{i_A} \tag{5.1}$$

where $\lambda^A_{i_A}$ is the interaction effect among variables in A and depends on i only through i_A, the sub-tuple of i corresponding to A. Conventionally we write $\lambda^A_{i_A} = \mu$ when $A = \varnothing$. To achieve identifiability the model has the constraints that the sum of the parameters $\lambda^A_{i_A}$ for any index in i_A equals zero.

In practice, however, attention is usually restricted to hierarchical loglinear models. Such models have the property that whenever a particular λ-term is constrained to zero then all higher λ-terms containing the same set of superscripts are also set to zero, that is, if $\lambda^A_{i_A} = 0$ then $\lambda^{\blacksquare}_{i_{\blacksquare}} = 0$ whenever $A \subseteq D$. For instance, the no three-way interaction model

$$\log \pi_{ijk} = \mu + \lambda^{X_1}_i + \lambda^{X_2}_j + \lambda^{X_3}_k + \lambda^{X_1X_2}_{ij} + \lambda^{X_1X_3}_{ik} + \lambda^{X_2X_3}_{jk}$$

with model symbol $(1, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3)$ is a hierarchical loglinear model, but the model

$$\log \pi_{ijk} = \mu + \lambda^{X_3}_k + \lambda^{X_1X_2}_{ij}$$

with model symbol $(1, X_3, X_1X_2)$ is not a hierarchical loglinear model because $\lambda^{X_2}_j = 0$ while $\lambda^{X_1X_2}_{ij}$ is still present in the model. Clearly, a hierarchical loglinear model is specified in terms of the highest interaction terms which do not nest with each other in the model symbol. The collection of these interaction terms is called the generating class for such a loglinear model. For the no three-way interaction model the generating class is (X_1X_2, X_1X_3, X_2X_3). Thus a hierarchical model can be symbolized by its generating class. For more detail about hierarchical loglinear models we refer the reader to Bishop, Fienberg and Holland (1975).

On the other hand, any loglinear model has a graphical representation (Darroch, Lauritzen and Speed, 1980). An independence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for a loglinear model consists of a finite set $\mathcal{V}$ of vertices, each vertex representing a variable in the model, and a finite set $\mathcal{E}$ of edges, each edge connecting vertices appearing as (or embedded in) a term in the model symbol. Thus there are as many vertices as dimensions of the contingency table and an edge represents a partial association between the corresponding two variables. For example, the independence graphs of the model $(1, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3)$ and $(1, X_3, X_1X_2)$ are shown in Figure 5.1 (1) and (2) respectively. Note that we call a graph complete if there is an edge

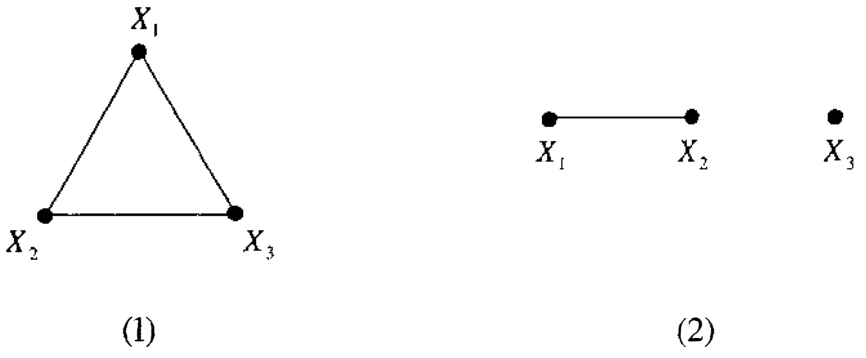


Figure 5.1: (1) The independence graph for the model $(1, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3)$. (2) The independence graph for the model $(1, X_3, X_1X_2)$.

between every pair of vertices. For instance, the graph in Figure 5.1 (1) is complete. Any subset $\mathcal{U} \subseteq \mathcal{V}$ induces a subgraph of $\mathcal{G}$, denoted $\mathcal{G}_{\mathcal{U}} = (\mathcal{U}, \mathcal{F})$, whose edge set $\mathcal{F}$ consists of those edges in $\mathcal{E}$ for which both endpoints are in $\mathcal{U}$. A subset $\mathcal{U} \subseteq \mathcal{V}$ is called complete if it induces a complete subgraph. A subset $\mathcal{U} \subseteq \mathcal{V}$ is called a clique if it is maximally complete. In other words, $\mathcal{U}$ is complete, and if $\mathcal{U} \subset \mathcal{W}$, then $\mathcal{W}$ is not complete. For example, the cliques in the graph shown in Figure 5.2 are $\{X_1, X_2, X_3\}$ and $\{X_3, X_4\}$.

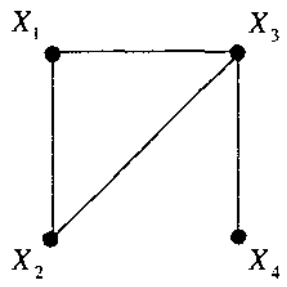


Figure 5.2: An incomplete graph.

Recently graph theory has been used to create a new type of statistical model, a graphical model, and we find that only some hierarchical loglinear models are graphical models, identified by the property that the generating class is directly given by the cliques of the independence graph of a loglinear model (Darroch, Lauritzen and Speed, 1980). For example, for the hierarchical loglinear models $(X_1X_2X_3)$ and (X_1X_2, X_1X_3, X_2X_3), the former model is a graphical loglinear model, but the latter one is not. These two models have the same independence graph as shown in Figure 5.1 (1), but only the first model's generating class is induced by the clique of its independence graph. Thus any hierarchical model has an independence graph representation, but not all hierarchical models are graphical models. The relationship among loglinear models, hierarchical models and graphical models is highlighted in Figure 5.3.

In this way we obtain a subset of hierarchical loglinear models, graphical loglinear models, the models discussed in this chapter. Darroch, Lauritzen and Speed (1980) showed that graphical loglinear models are always an intersection of a finite set of the following type of CI model, collectively denoted $CI(G)$. Such CI models are determined by a family of disjoint subsets $A_1, A_2, \ldots, A_t$ of S which are mutually

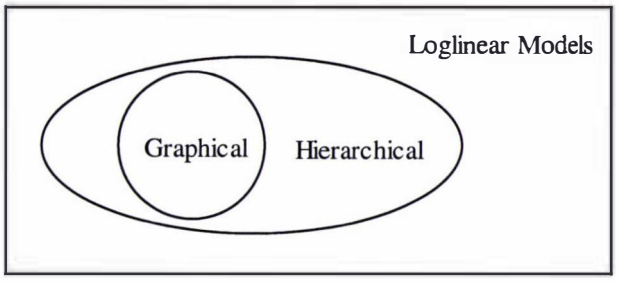


Figure 5.3: The relationship among loglinear models, hierarchical models and graphical models.

independent given all other variables in S, or more formally

$$A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t \mid C$$

where $C = (S \setminus \cup_{i=1}^{t} A_i)$.

Suppose that the disjoint subsets $A_1, A_2, \ldots, A_t, C$ of S have the number of combinations of levels of variables $m_1, m_2, \ldots, m_t, m$ respectively, then we choose in the sequel to consider A_k as a single categorical variable with m_k levels for $k = 1, 2, \ldots, t$ and C with m levels. The joint probability mass functions (or relative frequency tables) on the categorical variables $A_1, A_2, \ldots, A_t, C$ correspond to ordered n-tuples in $\mathbf{R}^n$ where $n = m_1 m_2 \ldots m_t m$. Since the sum of elements of these n-tuples is one, each of these n-tuples corresponds to a point in an $n-1$ dimensional simplex

$$S_{n-1} = \{(p_1, p_2, \ldots, p_n) \in \mathbf{R}^n \mid \sum_{i=1}^{n} p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i\} \subseteq \mathbf{R}^n$$

Note that the vertices of S_{n-1} are the unit vectors in $\mathbf{R}^n$. Thus a point in S_{n-1} can represent an ordered n-tuple with the sum of its elements equals to one, or a joint probability mass function, or a relative frequency table. These three representations are interchangeable throughout this chapter.

The model space of a graphical loglinear model with categorical variables $A_1, A_2,$

$\ldots, A_t, C$ is a subset within S_{n-1}. Our goal is to describe this model space through the geometry of CI statements and then to describe standard model fitting procedures within this framework. The motivation to explore the geometry of CI statements has come from the desire to better understand graphical models and their fitting. Graphical models, as we have pointed out, occur as finite sets of CI statements. This work builds on earlier papers of Fienberg (1968 and 1970), where the geometry of the independence model space for two independent categorical variables was explored. Based on the geometrical setting of a CI statement, the geometry underlying the iterative proportional fitting method (Deming and Stephan, 1940), a commonly used algorithm for finding the maximum likelihood estimate in loglinear models, will be constructed.

This will be accomplished as follows. In Section 5.2 we present the essential geometric tools required to describe the model space. Section 5.3 discusses the geometric setting for three kinds of distributions for a contingency table: the joint distribution, the marginal distribution and the conditional distribution. Section 5.4 first develops the geometric framework for unconditional independence models and then develops the geometry of general conditional independence models. Section 5.5 uses the geometric framework to illustrate with examples how the direct fitting method and the iterative proportional fitting method work for a conditional independence model. Finally, in Section 5.6, we summarize this chapter.

5.2 Technical preliminaries

In order to conveniently describe the geometric setting for $CI(G)$ statements, we now review key concepts and establish some notation (all notation is presented to be

consistent with later usage).

a) Affine independence

1. We say that P is an *affine combination* of P_i in $\mathbf{R}^N$, $i = 1, 2, \ldots, n$ when

$$P = \sum_{i=1}^{n} \alpha_i P_i \qquad \text{and} \qquad \sum_{i=1}^{n} \alpha_i = 1$$

for real numbers α_i. So all affine combinations of a set of vectors provide the line, plane etc. passing through them.

2. A set $S = \{P_1, P_2, \ldots, P_n\} \subseteq \mathbf{R}^N$ is *affinely dependent* provided at least one P_i for $i = 1, 2, \ldots, n$ is representable as an affine combination of the others. If a set S fails to be affinely dependent we call it *affinely independent.* So vectors are affinely independent if one of them does not lie in the smallest line, plane etc. passing through the other points.

b) Convexity

1. A set $A \subseteq \mathbf{R}^N$ is *convex* if for any $P_1, P_2 \in A$ and $0 \leq \alpha \leq 1$ we have

$$\alpha P_1 + (1 - \alpha) P_2 \in A$$

2. The *convex hull* of a subset A of $\mathbf{R}^N$ is the intersection of all the convex sets in $\mathbf{R}^N$ which contain A. It is denoted $\mathrm{co}(A)$.

c) Simplex

A general n *dimensional simplex* $S_n \subseteq \mathbf{R}^N$ $(n \leq N)$ is the convex hull of $n+1$ affinely independent points in $\mathbf{R}^N$. Thus a two dimensional simplex is a triangle, a three

dimensional simplex is a tetrahedron and so on.

d) Two ways of building sets

1. Corresponding point convex hull

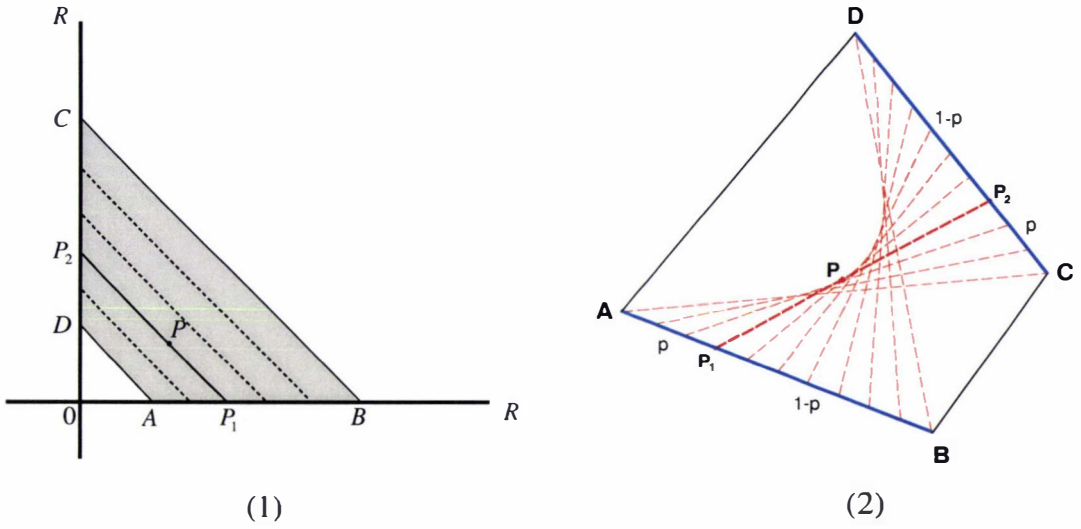
   We now define the *m-fold corresponding point convex hull* of a given compact set $K \subseteq \mathbf{R}^n$. First we embed K into $\mathbf{R}^{nm}$ m times in the following natural way,

$$\overbrace{K \times \{\mathbf{0}\} \times \ldots \times \{\mathbf{0}\}}^{m}$$
$$\{\mathbf{0}\} \times K \times \ldots \times \{\mathbf{0}\}$$
$$\vdots \qquad\qquad \vdots$$
$$\{\mathbf{0}\} \times \{\mathbf{0}\} \times \ldots \times K$$

   where $\mathbf{0}$ is the origin in $\mathbf{R}^n$. Consider the point in each of these m copies of K determined by $k \in K$ and denote the convex hull of these corresponding points by C_k. Then $\cup_{k \in K} C_k$ is the *m-fold corresponding point convex hull* of K, denoted $\mathrm{co}_c\{m, K\}$. The corresponding points, m copies of k, are denoted by $P_1, P_2, \ldots, P_m$ respectively in $\mathbf{R}^{nm}$.

   For example, a trapezium $ABCD$ is a corresponding point convex hull (see Figure 5.4 (1)), since it can be constructed as the corresponding point convex hull of two copies of $K = \{p : p \in [a, b]\}$, a one-dimensional simplex, in $\mathbf{R}^2$. On the other hand, a portion of a hyperbolic paraboloid within a tetrahedron $ABCD$ is a corresponding point convex hull, constructed as the corresponding point convex hull of two copies of $K = \{(p, 1-p) : p \in [0, 1])\}$, again a one-dimensional simplex, in $\mathbf{R}^4$ (see Figure 5.4 (2)). Each point (for example, the

point P) on these corresponding point convex hulls is uniquely determined by its corresponding points (the point P_1 and P_2).



(1) (2)

Figure 5.4: (1) The corresponding point convex hull of two copies of a line segment $K = \{p : p \in [a, b]\}$ in $\mathbf{R}^2$. Here we have $AB = CD = K$ and the points P_1 and P_2 with coordinates $(p, 0)$ and $(0, p)$ in $\mathbf{R}^2$ respectively. (2) The corresponding point convex hull of two copies of a line segment $K = \{(p, 1-p) : p \in [0, 1])\}$ in $\mathbf{R}^4$. Here we have $AB = CD = K$ and the points P_1 and P_2 with coordinates $(p, 1-p, 0, 0)$ and $(0, 0, p, 1-p)$ in $\mathbf{R}^4$ respectively.

2. Set convex hull

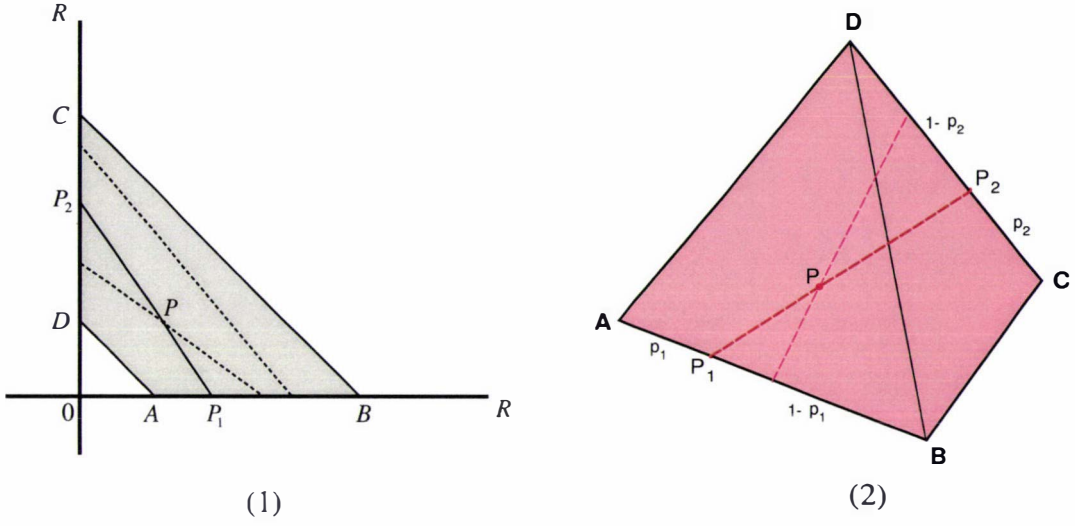
   The *set convex hull* of $K_1, K_2, \ldots, K_m \subseteq \mathbf{R}^N$ is

   $$\{\alpha_1 k_1 + \alpha_2 k_2 + \cdots + \alpha_m k_m \mid k_i \in K_i, \alpha_i \geq 0 \text{ for each } i \text{ and } \sum_{i=1}^{m} \alpha_i = 1\},$$

   denoted $\mathrm{co}_s\{K_1, K_2, \ldots, K_m\}$. In the special case where each K_i is a common compact set K in $\mathbf{R}^n$ naturally embedded, as described in (a), we write the set convex hull as $\mathrm{co}_s\{m, K\}$.

   For example, a face of a trapezium $ABCD$ is also a set convex hull of two copies, AB and CD, of $K = \{p : p \in [a, b]\}$ in $\mathbf{R}^2$ (see Figure 5.5 (1)). It occurs as the

union of all convex hulls of pairs of points, the points P_1 and P_2, one selected from each of a pair of line segments AB and CD. However, the set convex hull of two copies, AB and CD, of $K = \{(p, 1-p) : p \in [0,1])\}$ in $\mathbf{R}^4$ is the tetrahedron $ABCD$ (see Figure 5.5 (2)). Note that a point (for example, the point P) in the set convex hull can arise in more than one way.



(1) (2)

Figure 5.5: (1) The set convex hull of two copies of a line segment $K = \{p : p \in [a, b]\}$ in $\mathbf{R}^2$. Here we have $AB = CD = K$ and the points P_1 and P_2 with coordinates $(p_1, 0)$ and $(0, p_2)$ (where $p_1, p_2 \in [a, b]$) in $\mathbf{R}^2$ respectively. (2) The set convex hull of two copies of a line segment $K = \{(p, 1-p) : p \in [0,1])\}$ in $\mathbf{R}^4$. Here we have $AB = CD = K$ and the points P_1 and P_2 with coordinates $(p_1, 1-p_1, 0, 0)$ and $(0, 0, p_2, 1-p_2)$ (where $p_1, p_2 \in [0, 1]$) in $\mathbf{R}^4$ respectively.

5.3 Geometric setting for distributions

In this section we consider the geometry of three kinds of distribution for categorical variables $(A_1, A_2, \ldots, A_t, C)$: the joint distribution, the marginal distributions

and the conditional distributions. Here we first reveal the relationship among three distributions, and then discuss the geometry of distributions with fixed margin.

The relationship among three distributions

Recall that the number of levels of variables $A_1, A_2, \ldots, A_t$ and C are $m_1, m_2, \ldots, m_t$ and m respectively. Denote the distribution of A_k by P_{A_k} for $k = 1, 2, \ldots, t$ and C by P_C. The main geometric results about distributions for $(A_1, A_2, \ldots A_t, C)$ will be

1. The joint distribution of $(A_1, A_2, \ldots A_t, C)$, denoted by $P(A_1, \ldots, A_t, C)$, corresponds to a point P (named the global point) in a simplex S_{n-1} (named the global simplex) where $n = m_1 m_2 \ldots m_t m$.

2. The conditional distributions of $(A_1, A_2, \ldots, A_t)$ for given ith level of C, denoted $P(A_1, \ldots, A_t \mid C = c_i)$, correspond to a point P_i in an $r-1$ dimensional simplex $S_{r-1}^{(i)}$ (where $r = m_1 m_2 \ldots m_t$) for $i = 1, 2, \ldots, m$. Here P_i and $S_{r-1}^{(i)}$ are named the local point and the local simplex respectively.

3. The marginal distribution of C corresponds to a point P_C (named the marginal point) in a simplex S_{m-1} (named the marginal simplex). Any global point $P \in S_{n-1}$ can be mapped into a marginal point $P_C \in S_{m-1}$ by

$$M = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}_{m \times m}$$

with $\mathbf{0} = [0, 0, \ldots, 0]$, $\mathbf{1} = [1, 1, \ldots, 1] \in \mathbf{R}^r$.

4. The global point P can be represented as a convex combination of the local points $P_1, P_2, \ldots, P_m$ using as coefficients the marginal distribution of C. Specifically

$$P = c_1 P_1 + c_2 P_2 + \cdots + c_m P_m$$

where $P_C = \{c_1, c_2, \ldots, c_m\}$ with $c_i \geq 0$ and $\sum c_i = 1$ for $i = 1, 2 \ldots, m$.

We turn now to demonstrate these results.

A joint probability mass function for $(A_1, A_2 \ldots, A_t, C)$ corresponds an ordered n-tuple $(p_1, p_2, \ldots p_n)$ in $\mathbf{R}^n$ with $\sum_i p_i = 1$ and $p_i \in [0, 1]$ for all i. This n-tuple corresponds to a unique point P in the global simplex S_{n-1}, the convex hull of the unit vectors in $\mathbf{R}^n$. Regarding the marginal distribution of C, we rearrange the n-tuple $(p_1, p_2, \ldots p_n)$ as an $m \times r$ matrix, such that a row corresponds to the level of C, and a column corresponds to a combination of levels of variables in $(A_1, A_2, \ldots, A_t)$. Specifically, we consider

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mr} \end{bmatrix} \begin{matrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{matrix} \tag{5.2}$$

where p_{ij} is the joint probability of the ith level of C and the jth level of combinations of $(A_1, A_2, \ldots, A_t)$. Thus $\sum_{ij} p_{ij} = 1$ where $p_{ij} \geq 0$ for all i, j and $\sum_j p_{ij} = c_i$ for each i.

From the definition of conditional probability (George and Roger 1990, p.18), we know that

$$\begin{aligned} P(A_1, \ldots, A_t, C) &= c_1 P(A_1, \ldots, A_t \mid C = c_1) + c_2 P(A_1, \ldots, A_t \mid C = c_2) \\ &\quad + \cdots + c_m P(A_1, \ldots, A_t \mid C = c_m) \end{aligned} \tag{5.3}$$

where $P_C = \{c_1, c_2, \ldots, c_m\}$ with $c_i \geq 0$ and $\sum c_i = 1$ for $i = 1, 2 \ldots, m$.

Using (5.3), we can expand (5.2) as

$$\begin{bmatrix} p_{11} & \cdots & p_{1r} \\ p_{21} & \cdots & p_{2r} \\ \vdots & & \vdots \\ p_{m1} & \cdots & p_{mr} \end{bmatrix} = c_1 \begin{bmatrix} \frac{p_{11}}{c_1} & \cdots & \frac{p_{1r}}{c_1} \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 & \cdots & 0 \\ \frac{p_{21}}{c_2} & \cdots & \frac{p_{2r}}{c_2} \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} + \cdots + c_m \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ \frac{p_{m1}}{c_m} & \cdots & \frac{p_{mr}}{c_m} \end{bmatrix} \quad (5.4)$$

The left hand side of (5.4) is the joint probability mass function for $(A_1, A_2 \ldots, A_t, C)$, corresponding to a global point P in the global simplex S_{n-1} (see Figure 5.6). Here S_{n-1} has vertices

$$U_1^{(1)}, U_2^{(1)}, \ldots, U_r^{(1)}, U_1^{(2)}, U_2^{(2)}, \ldots, U_r^{(2)}, \ldots, U_1^{(m)}, U_2^{(m)}, \ldots, U_r^{(m)}$$

where

$$U_1^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \; U_2^{(1)} = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \; \ldots, U_r^{(m)} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

the unit vectors in $\mathbf{R}^n$. Thus P can be uniquely represented as a convex combination of unit vectors as

$$P = p_{11}U_1^{(1)} + \cdots + p_{1r}U_r^{(1)} + p_{21}U_1^{(2)} + \cdots + p_{2r}U_r^{(2)} + \cdots + p_{m1}U_1^{(m)} + \cdots + p_{mr}U_r^{(m)}$$

The ith matrix in the right hand side of (5.4) is the conditional distribution of $(A_1, A_2, \ldots, A_t)$ at the ith level of C; this corresponds to local point P_i in the local simplex $S_{r-1}^{(i)}$ with vertices $U_1^{(i)}, U_2^{(i)}, \ldots, U_r^{(i)}$ for $i = 1, 2, \ldots, m$. Then $S_{r-1}^{(i)}$ can be viewed as the ith boundary of the global simplex S_{n-1} (see Figure 5.6), and P_i as the convex combination of unit vectors $U_1^{(i)}, U_2^{(i)}, \ldots, U_r^{(i)}$ for all i given by

$$P_i = \frac{p_{i1}}{c_i}U_1^{(i)} + \frac{p_{i2}}{c_i}U_2^{(i)} + \ldots + \frac{p_{ir}}{c_i}U_r^{(i)}$$

Therefore, Equation (5.4) indicates that the global point P for the joint distribution of $A_1, A_2, \ldots, A_t, C$ is a convex combination of the local points $P_1, P_2, \ldots, P_m$, the conditional distributions of $(A_1, A_2, \ldots, A_t)$ for given level of C. The coefficients of this convex combination together form the marginal distribution P_C (see Figure 5.6). Specifically,

$$P = c_1 P_1 + c_2 P_2 + \cdots + c_m P_m$$

where $P_C = \{c_1, c_2, \ldots, c_m\}$, $P \in S_{n-1}$ and $P_i \in S_{r-1}^{(i)}$ for $i = 1, 2, \ldots, m$.

On the other hand, the sums of rows in the matrix (5.2) form the marginal distribution of C as $P_C = \{c_1, c_2, \ldots, c_m\}$ where $c_i = \sum_j p_{ij}$ for all i. Thus any marginal distribution of C corresponds to a point P_C in an $m-1$ dimensional simplex S_{m-1} such that

$$P_C = c_1 V_1 + c_2 V_2 \ldots + c_m V_m$$

where $V_1, V_2, \ldots, V_m$ are unit vectors in $\mathbf{R}^m$ (see Figure 5.6). Then the joint distribution P is related to the marginal distribution P_C by the mapping

$$MP^T = P_C^T$$

where $P \in S_{n-1}$, $P_C \in S_{m-1}$ and

$$M = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}_{m \times m}$$

with $\mathbf{0} = [0, 0, \ldots, 0]$, $\mathbf{1} = [1, 1, \ldots, 1] \in \mathbf{R}^r$. Note that this is a many-to-one mapping: for any point $P_C \in S_{m-1}$ its pre-image in S_{n-1} is all distributions with fixed margin P_C.

Now the relationship among the three kinds of distribution for categorical variables $(A_1, A_2, \ldots, A_t, C)$ can be summarized by Figure 5.6.

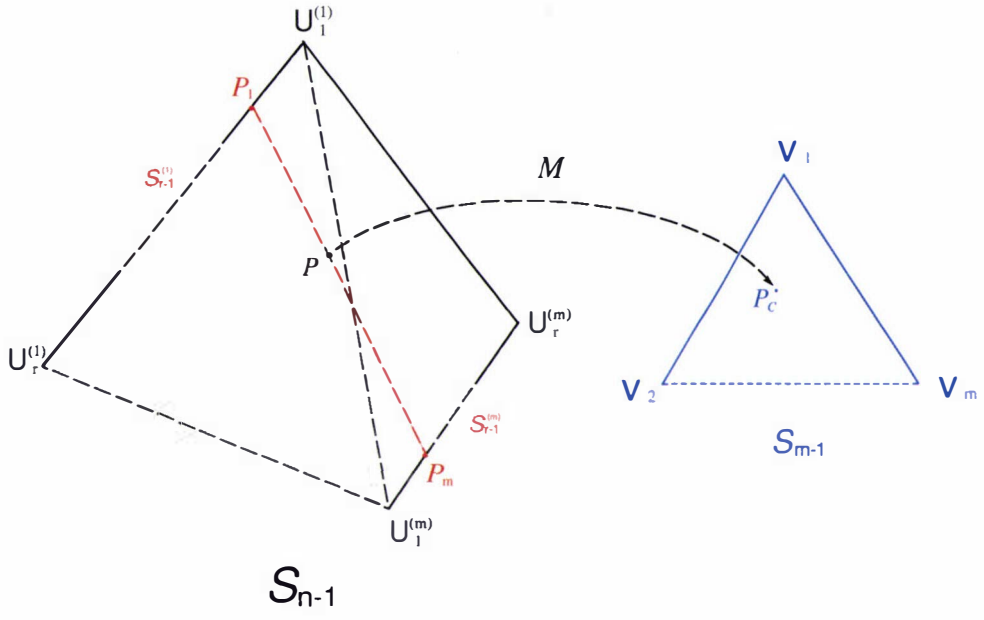


Figure 5.6: A joint distribution of $(A_1, A_2, \ldots, A_t, C)$ corresponds to a point P in the global simplex S_{n-1}. A conditional distribution of $(A_1, A_2, \ldots, A_t)$ at the ith level of C corresponds to a point P_i on the local simplex $S^{(i)}_{r-1}$, the ith boundary of the global simplex S_{n-1}, for $i = 1, 2, \ldots, m$. A marginal distribution of C is associated with a point P_C in an $m-1$ dimensional simplex S_{m-1}. The joint distribution P and the marginal distribution P_C are linked through the mapping M. Meanwhile, the point P is a convex combination of the local points $P_1, P_2, \ldots, P_m$ using as coefficients P_C.

Geometry of distributions with fixed margin

If we denote the ith row in the matrix M (see p.111) by s_i for $i = 1, 2, \ldots, m$, then the set of points $F \subset S_{n-1}$ with fixed C margin $P_C = \{c_1, c_2, \ldots, c_m\}$ is

$$F = \{P \in S_{n-1} \mid s_i . P = c_i \text{ for all } i\}$$

Geometrically, any point P with $s_i.P = c_i$ lies on an $n-1$ dimensional slice (hyperplane) which is perpendicular to the vector s_i in $\mathbf{R}^n$ and has a projection of length $c_i/\|s_i\|$ onto s_i. Thus F is an intersection of m of these hyperplanes, so has dimension $n-m$ in S_{n-1}, and is perpendicular to $s_1, s_2, \ldots, s_m$ simultaneously. In fact, from standard linear algebra results F is an affine transformation of the nullspace of the matrix M (Anton 1994, p.260).

Since $s_1, s_2, \ldots, s_m$ are linearly independent, so the $n-m$ dimensional hyperplane F is perpendicular to an m dimensional hyperplane $G = \text{span}\{s_1, s_2, \ldots, s_m\}$. The projection of any $P \in F$ onto the hyperplane G is

$$\begin{aligned} P' &= (P.\frac{s_1}{\|s_1\|})\frac{s_1}{\|s_1\|} + (P.\frac{s_2}{\|s_2\|})\frac{s_2}{\|s_2\|} + \cdots + (P.\frac{s_m}{\|s_m\|})\frac{s_m}{\|s_m\|} \\ &= c_1\frac{s_1}{r} + c_2\frac{s_2}{r} + \cdots + c_m\frac{s_m}{r} \qquad (\|s_1\| = \sqrt{r}) \end{aligned}$$

Since $s_1/r, s_2/r, \ldots, s_m/r$ are linearly independent, thus affinely independent, P' is in an $m-1$ dimensional simplex with vertices $s_1/r, s_2/r, \ldots, s_m/r$. This $m-1$ dimensional simplex is nested within the global simplex S_{n-1} and has the same centre as the global simplex S_{n-1} due to the facts that $s_i/r \in S_{n-1}$ for all i (since the sum of components of s_i/r is one) and

$$(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) = \frac{1}{m}\frac{s_1}{r} + \frac{1}{m}\frac{s_2}{r} + \cdots + \frac{1}{m}\frac{s_m}{r}$$

where $n = mr$. Thus we call this $m-1$ dimensional simplex a central simplex denoted by S^C_{m-1}. Now the marginal distributions of C correspond to points in the central simplex S^C_{m-1}, and for a given P_C the set F with fixed P_C is perpendicular to S^C_{m-1} at the point P'.

Similarly, the set of points in S_{n-1} with fixed A_k margin is found to be a hyperplane which is perpendicular to the central simplex $S^{A_k}_{m_k-1}$ for $k = 1, 2, \ldots, t$. Note that now

the n-tuple $(p_1, p_2, \ldots p_n)$ is arranged as a two-way array such that its rows correspond to the level of A_k, and columns to a combination of levels of variables in $S \setminus A_k$ where $S = \{A_1, A_2, \ldots, A_t, C\}$.

In summary, the distributions with fixed margin P_C correspond to points in the intersection between a hyperplane F (an affine transformation of the nullspace of matrix M) and the global simplex S_{n-1}. All marginal distributions of C associate with points in the central simplex S^C_{m-1} whose vertices are the normalized row vectors in M. The hyperplane F is perpendicular to the central simplex S^C_{m-1} at the point corresponding to P_C.

The above results are illustrated in the following example.

Example 1 Suppose an observed relative frequency table with binary variables X_1 and X_2 is

$$(p_{11},\ p_{12},\ p_{21},\ p_{22}) = (0.4,\ 0.3,\ 0.2,\ 0.1)$$

This joint distribution of X_1 and X_2 corresponds to a point P in the tetrahedron $ABCD$, the global simplex S_3 (see Figure 5.7). Consider the marginal distribution of X_1, $P_{X_1} = (0.7,\ 0.3)$, the 4-tuple $(0.4,\ 0.3,\ 0.2,\ 0.1)$ is laid out as

$$\begin{bmatrix} 0.4 & 0.3 \\ 0.2 & 0.1 \end{bmatrix} \begin{matrix} 0.7 \\ 0.3 \end{matrix} \tag{5.5}$$

where the first and second rows of the matrix (5.5) are the joint probabilities of (X_1, X_2) when $X_1 = 0$ and $X_1 = 1$ respectively.

The definition of conditional probability gives that

$$\begin{bmatrix} 0.4 & 0.3 \\ 0.2 & 0.1 \end{bmatrix} = 0.7 \begin{bmatrix} 0.57 & 0.43 \\ 0 & 0 \end{bmatrix} + 0.3 \begin{bmatrix} 0 & 0 \\ 0.67 & 0.33 \end{bmatrix}$$

where $(0.57, 0.43, 0, 0)$ is the conditional distribution of X_2 for given $X_1 = 0$; this corresponds to a local point P_1 in the local simplex $S_1^{(1)}$, the line segment AB, while

$(0, 0, 0.67, 0.33)$ is the conditional distribution of X_2 when $X_1 = 1$; this corresponds to a local point P_2 in the local simplex $S_1^{(2)}$, the line segment CD (see Figure 5.7). Thus we obtain $P = 0.7P_1 + 0.3P_2$.

Note that any conditional distribution of X_2 when $X_1 = 0$ corresponds to a point on the line segment AB, while any conditional distribution of X_2 when $X_1 = 1$ corresponds to a point on the line segment CD.

A given marginal distribution of X_1, $P_{X_1} = (0.7,\ 0.3)$, corresponds to a point P' in the central simplex

$$S_1^{X_1} = \{P = c_1(0.5, 0.5, 0, 0) + c_2(0, 0, 0.5, 0.5) \mid \sum c_i = 1 \text{ and } c_i > 0 \text{ for all } i\}$$

the line segment EF. Any point in the global simplex S_3, the tetrahedron $ABCD$, with fixed $P_{X_1} = (0.7, 0.3)$ is in the set

$$K = \{P \in S_3 \mid s_1.P = 0.7 \text{ and } s_2.P = 0.3\}$$

where $s_1 = [1, 1, 0, 0]^T$ and $s_2 = [0, 0, 1, 1]^T$. The set K is the plane which is perpendicular to the central simplex $S_1^{X_1}$ at the point $P' = 0.7(0.5, 0.5, 0, 0) + 0.3(0, 0, 0.5, 0.5)$ (see Figure 5.7).

Symmetrically, any point in the global simplex S_3, the tetrahedron $ABCD$, with fixed P_{X_2} is in a plane which is perpendicular to the central simplex

$$S_1^{X_2} = \{P = c_1'(0.5, 0, 0.5, 0) + c_2'(0, 0.5, 0, 0.5) \mid \sum c_i' = 1 \text{ and } c_i' > 0 \text{ for all } i\}$$

the line segment GH (see Figure 5.7).

Therefore, the direction for preserving both the X_1 and X_2 margins in the tetrahedron $ABCD$ is perpendicular to the central simplexes EF and GH simultaneously, namely the direction $IJ = (-0.5, 0.5, 0.5, -0.5)$ (see Figure 5.7).

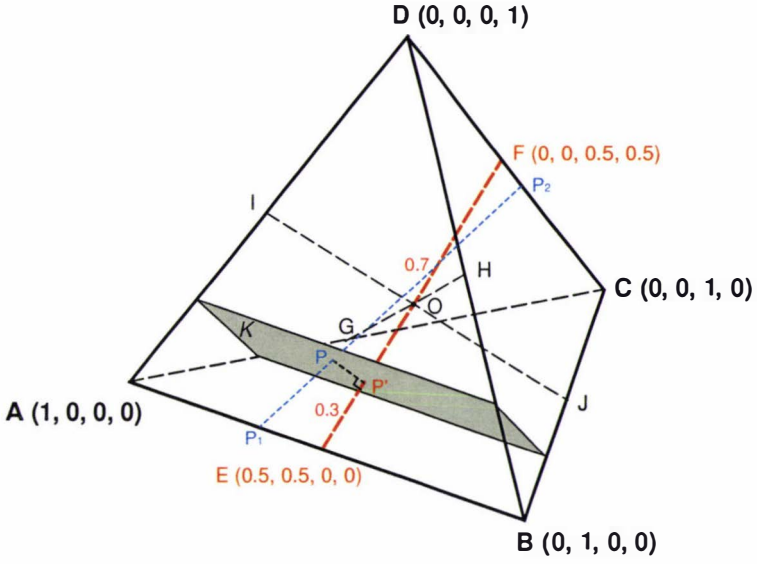


Figure 5.7: This figure shows the lines with constant margins for a relative frequency table involving binary variables X_1 and X_2. Any direction perpendicular to EF will preserve an X_1 margin. For example, points on the quadrilateral K have X_1 margin of $(0.7, 0.3)$. On the other hand, a direction for preserving an X_2 margin is perpendicular to GH. Thus along the direction IJ both X_1 and X_2 margins are preserved, where $I = (0.5, 0, 0, 0.5)$ and $J = (0, 0.5, 0.5, 0)$.

5.4 Geometric setting for CI models

Our aim now is to describe the subset of S_{n-1} providing the model space of

$$A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t \mid C$$

Fienberg (1968) describes the geometry of a two-way contingency table with independence between rows and columns, while Fienberg and Gilbert (1970) illustrate the associated results on a two by two contingency table. Here we start by building the model space for the unconditional independence statement $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$ in what we term the "local" simplex, and then extend this to describe the model space

for one CI(G) statement, $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t \mid C$, in the global simplex, formed as the convex hull of local simplexes.

Model space for an unconditional independence statement

Recall that the variables $A_1, A_2, \ldots, A_t$ have the number of levels $m_1, m_2, \ldots, m_t$ respectively, and P_{A_k} denotes the marginal distribution of A_k, an m_k-tuple for $k = 1, 2, \ldots, t$. We will show that the model space for $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$ is a recursively defined corresponding point convex hull in the simplex S_{r-1} where $r = m_1 m_2 \ldots m_t$. We begin by discussing the case where $t = 2$, then deal with the general situation, providing an illustrative example for each case.

Model space for $A_1 \perp\!\!\!\perp A_2$

The main result about the statement $A_1 \perp\!\!\!\perp A_2$ is:

> The model space of $A_1 \perp\!\!\!\perp A_2$ is the corresponding point convex hull of m_1 copies of S_{m_2-1} in $\mathbf{R}^{m_1 m_2}$, $\mathrm{co}_c\{m_1, S_{m_2-1}\}$, or the corresponding point convex hull of m_2 copies of S_{m_1-1} in $\mathbf{R}^{m_1 m_2}$, $\mathrm{co}_c\{m_2, S_{m_1-1}\}$.

The familiar joint probability mass function P for (A_1, A_2) can be represented as a 2-way array in which the ijth element of P is the joint probability associated with the ith level of A_1 and jth level of A_2. Letting $P_{A_1} = (\alpha_1, \alpha_2, \ldots, \alpha_{m_1})$ and $P_{A_2} = (\beta_1, \beta_2, \ldots, \beta_{m_2})$ where $\sum \alpha_i = 1$, $\alpha_i \in [0, 1]$ and $\sum \beta_j = 1$, $\beta_j \in [0, 1]$, then P is the $m_1 \times m_2$ matrix

$$\begin{bmatrix} \alpha_1\beta_1 & \alpha_1\beta_2 & \cdots & \alpha_1\beta_{m_2} \\ \alpha_2\beta_1 & \alpha_2\beta_2 & \cdots & \alpha_2\beta_{m_2} \\ \vdots & \vdots & & \vdots \\ \alpha_{m_1}\beta_1 & \alpha_{m_1}\beta_2 & \cdots & \alpha_{m_1}\beta_{m_2} \end{bmatrix}$$

Geometrically, we view the joint probability mass function P for (A_1, A_2) as an m_1m_2-tuple in the simplex $S_{m_1m_2-1} \subseteq \mathbf{R}^{m_1m_2}$. Conventionally, we write this m_1m_2-tuple lexicographically, laying out the rows in the above array sequentially as

$$\begin{aligned}&(\alpha_1\beta_1, \alpha_1\beta_2, \ldots, \alpha_1\beta_{m_2}, \alpha_2\beta_1, \alpha_2\beta_2, \ldots, \alpha_2\beta_{m_2}, \ldots\ldots, \alpha_{m_1}\beta_1, \alpha_{m_1}\beta_2, \ldots, \alpha_{m_1}\beta_{m_2}) \\ &= \alpha_1(P_{A_2}, \mathbf{0}, \ldots, \mathbf{0}) + \alpha_2(\mathbf{0}, P_{A_2}, \mathbf{0}, \ldots, \mathbf{0}) + \cdots + \alpha_{m_1}(\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, P_{A_2}) \qquad (5.6)\end{aligned}$$

where $\mathbf{0} = [0, 0, \ldots, 0] \in \mathbf{R}^{m_2}$.

Each possible value of the A_2 margin P_{A_2} corresponds to a point in a simplex S_{m_2-1}, a compact convex set. Expression (5.6) represents a point in the model space of $A_1 \perp\!\!\!\perp A_2$ as a convex combination of m_1 points with the same margin P_{A_2} in $\mathbf{R}^{m_1m_2}$. Hence the model space of $A_1 \perp\!\!\!\perp A_2$ is a corresponding point convex hull of m_1 copies of S_{m_2-1} in $\mathbf{R}^{m_1m_2}$, $\mathrm{co}_c\{m_1, S_{m_2-1}\}$. These m_1 corresponding points are linearly independent and thus affinely independent in $\mathbf{R}^{m_1m_2}$, so $\mathrm{co}_c\{m_1, S_{m_2-1}\}$ is a union of m_1-1 dimensional simplexes. Points in anyone of these simplexes correspond to joint probability mass functions with the same A_2 margin.

Symmetrically, the joint distribution can also be written as an (m_2m_1)-tuple by laying out the columns in the above array sequentially as

$$\begin{aligned}&(\alpha_1\beta_1, \alpha_2\beta_1, \ldots, \alpha_{m_1}\beta_1, \alpha_1\beta_2, \alpha_2\beta_2, \ldots, \alpha_{m_1}\beta_2, \ldots\ldots, \alpha_1\beta_{m_2}, \alpha_2\beta_{m_2}, \ldots, \alpha_{m_1}\beta_{m_2}) \\ &= (\beta_1(P_{A_1}, \mathbf{0}, \ldots, \mathbf{0}) + \beta_2(\mathbf{0}, P_{A_1}, \mathbf{0}, \ldots, \mathbf{0}) + \cdots + \beta_{m_2}(\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, P_{A_1})) \qquad (5.7)\end{aligned}$$

where $\mathbf{0} = [0, 0, \ldots, 0] \in \mathbf{R}^{m_1}$.

Expression (5.7) highlights that every point in the model space of $A_1 \perp\!\!\!\perp A_2$ is also a convex combination of m_2 points with the same margin P_{A_1} in $\mathbf{R}^{m_1 m_2}$, a corresponding point convex hull of m_2 copies of S_{m_1-1} in $\mathbf{R}^{m_1 m_2}$, $\mathrm{co}_c\{m_2, S_{m_1-1}\}$. Again $\mathrm{co}_c\{m_2, S_{m_1-1}\}$ is a union of $m_2 - 1$ dimensional simplexes, each simplex having fixed A_1 margin.

In summary, the model space for $A_1 \perp\!\!\!\perp A_2$ can be thought of geometrically as a corresponding point convex hull embedded in the simplex $S_{m_1 m_2 - 1}$. This corresponding point convex hull can be constructed using either of two families of simplexes. One family of simplexes is indexed by the A_1 margins, while the other family is indexed by the A_2 margins. Furthermore, since A_1 and A_2 are independent the joint probability mass function P is determined by the marginal distributions P_{A_1} and P_{A_2}, so we have that

1) Within each family the simplexes do not intersect, but between families each simplex will meet a simplex in another family at one point.

2) Any point on the independence model space meets precisely one simplex in each family.

We now illustrate the model space for $A_1 \perp\!\!\!\perp A_2$.

Example 2 When $m_1 = m_2 = 2$, the model space of $A_1 \perp\!\!\!\perp A_2$ is a portion of a hyperbolic paraboloid within a tetrahedron (a 3-dimensional simplex S_3). It is a doubly ruled surface and is completely determined by either family of lines (Fienberg and Gilbert, 1970). These two families of lines correspond to the two ways of constructing the independence surface. For one of them we build the independence surface in the following steps:

Step 1 Determine the unrestricted model space to be the 3-dimensional simplex with unit vectors in $\mathbf{R}^4$ as vertices, a tetrahedron $ABCD$ (see Figure 5.8).

Step 2 Partition the vertices of the tetrahedron $ABCD$ into m_1 (two) lots of m_2 (two) vertices.

Step 3 Embed m_1 (two) copies of the simplex associated with the marginal distribution of A_2, line segments $AB = CD = \{(\beta, 1-\beta) : \beta \in [0,1]\}$, in the tetrahedron $ABCD$.

Step 4 Find the points $(\beta, 1-\beta, 0, 0)$ and $(0, 0, \beta, 1-\beta)$ on AB and CD respectively corresponding to $P_{A_2} = (\beta, 1-\beta)$.

Step 5 Obtain the independence surface as the union of simplexes (the long dashed lines in Figure 5.8) with varying A_1 margin, each simplex being the convex hull of points $(\beta, 1-\beta, 0, 0)$ and $(0, 0, \beta, 1-\beta)$. The model space is

$$\cup_{\beta\in[0,1]}\{\alpha(\beta, 1-\beta, 0, 0) + (1-\alpha)(0, 0, \beta, 1-\beta) \mid \alpha \in [0,1]\} = P_{A_1} \otimes P_{A_2}$$

where "$\otimes$" denotes a tensor product and $P_{A_1} = (\alpha, 1-\alpha)$. This has a $\mathrm{co}_c\{2, S_1\}$ structure.

Note that the independence surface also can be generated by another family of simplexes (the short dashed lines in Figure 5.8), each with varying A_2 margin, so that the model space is

$$\cup_{\alpha\in[0,1]}\{\beta(\alpha, 0, 1-\alpha, 0) + (1-\beta)(0, \alpha, 0, 1-\alpha) \mid \beta \in [0,1]\} = P_{A_2} \otimes P_{A_1}.$$

The above procedure is illustrated in Figure 5.8.

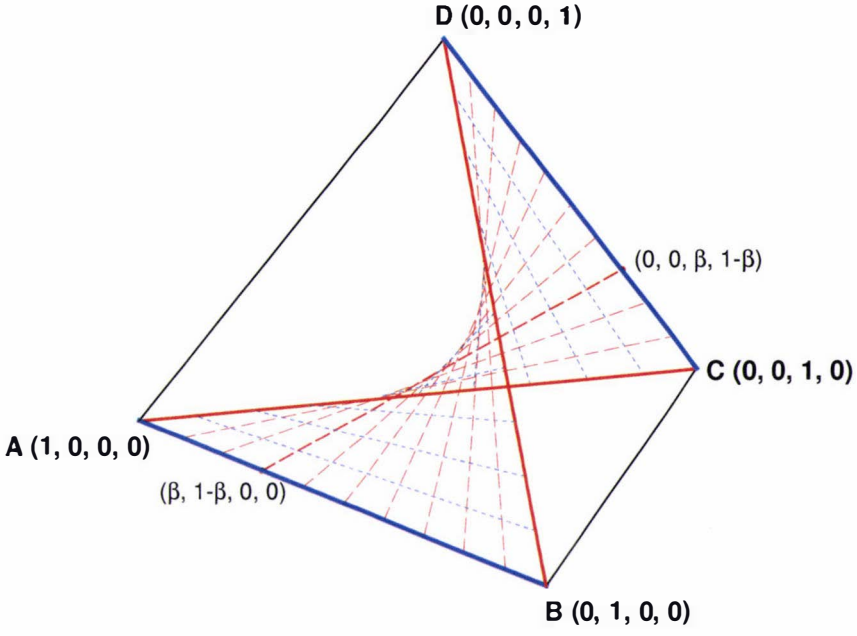


Figure 5.8: The model space for $A_1 \perp\!\!\!\perp A_2$ with $m_1 = m_2 = 2$ is part of a hyperbolic paraboloid within a tetrahedron $ABCD$. It can be generated by two families of simplexes, one with varying A_1 margin (the long dashed lines) and another with varying A_2 margin (the short dashed lines).

Example 3 When $m_1 = 3$ and $m_2 = 4$ the model space of $A_1 \perp\!\!\!\perp A_2$ is a corresponding point convex hull $\mathrm{co}_c\{3, S_3\}$ within an 11-dimensional simplex S_{11}. There are two ways of constructing the independence surface. For one of them we build the independence surface in the following steps:

Step 1 Determine the unrestricted model space to be the 11-dimensional simplex S_{11} with unit vectors in $\mathbf{R}^{12}$ as vertices.

Step 2 Partition the vertices of S_{11} into m_1 (three) lots of m_2 (four) vertices.

Step 3 Embed m_1 (three) copies of the simplex associated with the marginal distribution of A_2, tetrahedrons $ABCD, A'B'C'D', A''B''C''D''$, in the simplex S_{11}.

Step 4 Find the points $E = (b_1, b_2, b_3, b_4, 0, 0, 0, 0, 0, 0, 0, 0)$, $F = (0, 0, 0, 0, b_1, b_2, b_3, b_4, 0, 0, 0, 0)$ and $G = (0, 0, 0, 0, 0, 0, 0, 0, b_1, b_2, b_3, b_4)$ on $ABCD, A'B'C'D', A''B''C''D''$ respectively corresponding to $P_{A_2} = (b_1, b_2, b_3, b_4)$ where $\sum b_j = 1$ and $b_j \in [0, 1]$ for all j.

Step 5 Obtain the independence surface as the union of 2-dimensional simplexes (the triangle in Figure 5.9 (1)) with varying A_1 margin, each simplex being the convex hull of points E, F and G. The model space is

$$\cup_{(b_1,b_2,b_3,b_4)\in S_3}\{a_1E+a_2F+a_3G \mid \sum a_i = 1 \text{ and } a_i \in [0,1] \text{ for all } i\} = P_{A_1} \otimes P_{A_2}$$

where $P_{A_1} = (a_1, a_2, a_3)$. This has a $\mathrm{co}_c\{3, S_3\}$ structure.

Note that the independence surface also can be generated by another family of simplexes (the tetrahedrons in Figure 5.9 (2)), each with varying A_2 margin, so that the model space is

$$\cup_{(a_1,a_2,a_3)\in S_2}\{b_1A+b_2B+b_3C+b_4D \mid \sum b_j = 1 \text{ and } b_j \in [0,1] \text{ for all } j\} = P_{A_2} \otimes P_{A_1}$$

where

$$A = (a_1,\ 0,\ 0,\ 0,\ a_2,\ 0,\ 0,\ 0,\ a_3,\ 0,\ 0,\ 0)$$

$$B = (0,\ a_1,\ 0,\ 0,\ 0,\ a_2,\ 0,\ 0,\ 0,\ a_3,\ 0,\ 0)$$

$$C = (0,\ 0,\ a_1,\ 0,\ 0,\ 0,\ a_2,\ 0,\ 0,\ 0,\ a_3,\ 0)$$

$$D = (0,\ 0,\ 0,\ a_1,\ 0,\ 0,\ 0,\ a_2,\ 0,\ 0,\ 0,\ a_3).$$

The above procedures are illustrated in Figure 5.9.

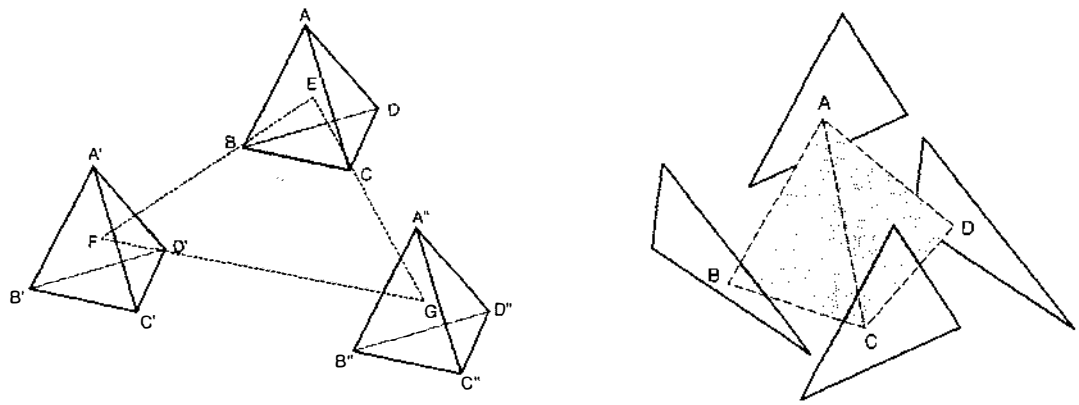


Figure 5.9: The model space for $A_1 \perp\!\!\!\perp A_2$ with $m_1 = 3$, $m_2 = 4$ can be constructed in two ways: (1) The union of 2-dimensional simplexes (triangles) with varying A_1 margin, or (2) The union of 3-dimensional simplexes (tetrahedrons) with varying A_2 margin.

Model Space for $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$

In this general case the main result about the model space of $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$ is:

> The model space of $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$ can be constructed as a corresponding point convex hull in $\mathbf{R}^n$ $(n = m_1 m_2 \ldots m_t)$ in t distinct ways. A typical form is
>
> $$\mathrm{co}_c\{m_1, \mathrm{co}_c\{m_2, \ldots, \mathrm{co}_c\{m_{t-1}, S_{m_t-1}\}\ldots\}\}.$$

To construct the model space of $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$, we start with the model space of $A_{t-1} \perp\!\!\!\perp A_t$, namely $\mathrm{co}_c\{m_{t-1}, S_{m_t-1}\}$, a compact set. The model space of $A_{t-2} \perp\!\!\!\perp (A_{t-1} \perp\!\!\!\perp A_t)$ is then the union of $m_{t-2}-1$ dimensional simplex $S_{m_{t-2}-1}$ whose vertices are the corresponding points in m_{t-2} copies of $\mathrm{co}_c\{m_{t-1}, S_{m_t-1}\}$ in $\mathbf{R}^n$. That is, the corresponding point convex hull of m_{t-2} copies of $\mathrm{co}_c\{m_{t-1}, S_{m_t-1}\}$, which

we write as $\mathrm{co}_c\{m_{t-2}, \mathrm{co}_c\{m_{t-1}, S_{m_t-1}\}\}$. Note that those corresponding points are determined by the joint margin of (A_{t-1}, A_t). Recursively we obtain that the model space of $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$ is the corresponding point convex hull

$$I = \mathrm{co}_c\{m_1, K\}$$

where $K = \mathrm{co}_c\{m_2, \ldots, \mathrm{co}_c\{m_{t-1}, S_{m_t-1}\} \ldots\}$, named a sub-corresponding point convex hull.

We know that the corresponding point convex hull I is a union of simplexes S_{m_1-1} whose vertices are the corresponding points on m_1 copies of the sub-corresponding point convex hull K. Each of these simplexes S_{m_1-1} has varying A_1 margin but keeps all other margins the same. Meanwhile we can permute the order of $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$ to obtain the same corresponding point convex hull in $\mathbf{R}^n$ in $t!$ distinct ways. This corresponding point convex hull however is only covered by each of t families of simplexes; a simplex in a family will have varying A_k margin but the same remaining margins. Specifically,

$$I = \mathrm{co}_c\{m_k, K\}$$

where K, a sub-corresponding point convex hull, is the independence model space of all the subsets in S except A_k for $k = 1, 2, \ldots, t$.

Since $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp A_t$ the joint probability mass function P_A for $(A_1, A_2, \ldots, A_t)$ is determined by the marginal distributions $P_{A_1}, P_{A_2}, \ldots, P_{A_t}$, so we have that

1) Within each family the simplexes do not intersect, but between families each simplex will meet all another family of simplexes in one point.

2) Any point on the independence model space meets precisely t simplexes with different margins.

We now illustrate the model space for $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp A_3$.

Example 4 When $m_1 = m_2 = m_3 = 2$, the model space of $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp A_3$ is a corresponding point convex hull in a simplex S_7. There are 3! ways of constructing the independence surface. For example, we can build the independence surface in the following steps:

Step 1 Determine the unrestricted model space, the simplex S_7 whose vertices are the unit vectors in $\mathbf{R}^8$.

Step 2 Partition the vertices of S_7 into two (m_1) lots of four $(m_2 m_3)$ vertices.

Step 3 Place a copy of the independence model space of $A_2 \perp\!\!\!\perp A_3$, $\mathrm{co}_c\{2, S_1\}$, in each of the two (m_1) partition simplexes $ABCD$ and $A'B'C'D'$ (see Figure 5.10). Here we have

$$\begin{aligned} \mathbf{c}\bullet_c\{2, S_1\} &= \cup_{\gamma\in[0,1]}\{\beta(\gamma, (1-\gamma), 0, 0, \gamma, (1-\gamma), 0, 0) \\ &\quad +(1-\beta)(0, 0, \gamma, (1-\gamma), 0, 0, \gamma, (1-\gamma)) \mid \beta \in [0,1]\} \end{aligned}$$

where $P_{A_2} = (\beta, 1-\beta)$ and $P_{A_3} = (\gamma, 1-\gamma)$.

Step 4 Find the corresponding points $(\beta\gamma, (1-\beta)\gamma, \beta(1-\gamma), (1-\beta)(1-\gamma), 0, 0, 0, 0)$, $(0, 0, 0, 0, \beta\gamma, (1-\beta)\gamma, \beta(1-\gamma), (1-\beta)(1-\gamma))$ for given P_{A_2} and P_{A_3} on the corresponding point convex hulls in $ABCD$ and $A'B'C'D'$ respectively.

Step 5 Form a corresponding point convex hull of these two (m_1) copies of $\mathrm{co}_c\{2, S_1\}$,

that is, produce

$$\begin{aligned} & \mathrm{co}_c\{2, \mathrm{co}_c\{2, S_1\}\} \\ = \; & \cup_{\beta,\gamma\in[0,1]}\{\alpha(\beta\gamma, \beta(1-\gamma), (1-\beta)\gamma, (1-\beta)(1-\gamma), 0, 0, 0, 0) \\ & +(1-\alpha)(0, 0, 0, 0, \beta\gamma, \beta(1-\gamma), (1-\beta)\gamma, (1-\beta)(1-\gamma)) \mid \alpha \in [0,1]\} \\ = \; & P_{A_1} \otimes P_{A_2} \otimes P_{A_3} \end{aligned}$$

where $P_{A_1} = (\alpha, 1-\alpha)$.

Here the model space $\mathrm{co}_c\{2, \mathrm{co}_c\{2, S_1\}\}$ is a union of one-dimensional simplexes (that is, line segments), and each of these line segments has varying A_1 margin but fixed A_2 and A_3 margins. The above procedure is illustrated in Figure 5.10.

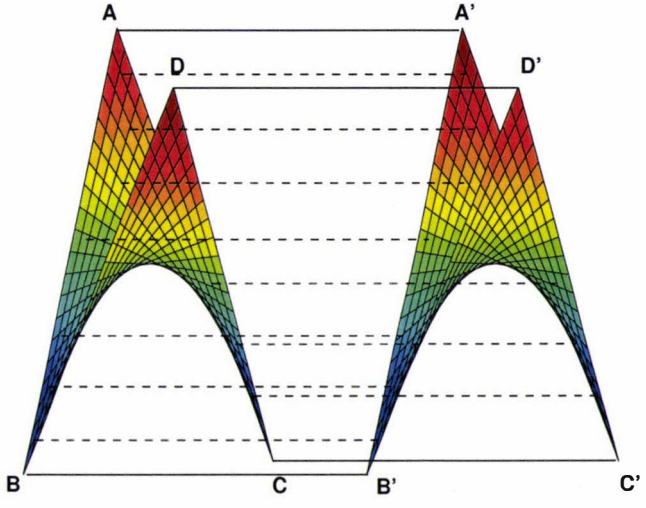


Figure 5.10: A 3-dimensional representation of the model space for $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp A_3$ with $m_1 = m_2 = m_3 = 2$. The corresponding point convex hull $\mathrm{co}_c\{2, \mathrm{co}_c\{2, S_1\}\}$, is constructed as the union of corresponding point convex hulls of two copies of the independence model space of A_2 and A_3, namely $\mathrm{co}_c\{2, S_1\}$, sketch as $ABCD$ and $A'B'C'D'$.

Model space for a conditional independence statement

Recall that the levels of variables $A_1, A_2, \ldots, A_t$ and C are $m_1, m_2, \ldots, m_t$ and m respectively, the conditional distribution of $(A_1, A_2, \ldots, A_t)$ for given ith level of C is denoted by P_i for $i = 1, 2, \ldots, m$ and the marginal distribution of C by $P_C = \{c_1, c_2, \ldots, c_m\}$. It is natural to reach conclusions about the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ by extending the results of the geometric setting for distributions (Section 5.3) and the unconditional independence statements. The main result is now described:

> The model space can be constructed geometrically as a set convex hull of m copies of I, the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$, in the global simplex S_{n-1} where $n = m_1 m_2 \cdots m_t m$, namely
>
> $$\text{co}_s\{m, I\} = \{c_1 P_1 + c_2 P_2 + \cdots + c_m P_m \mid P_i \in I_i,\ c_i \geq 0 \text{ and } \sum c_i = 1 \text{ for all } i\}$$
>
> where I_i is the ith copy of I in the local simplex $S_{r-1}^{(i)}$ for $i = 1, 2, \ldots, m$.

From Section 5.3 we know that the global point P for the joint distribution of $A_1, A_2, \ldots, A_t, C$ is a convex combination of the local points $P_1, P_2, \ldots, P_m$, the conditional distributions of $(A_1, A_2, \ldots, A_t)$ for given level of C, using the marginal distribution P_C as coefficients. Specifically,

$$P = c_1 P_1 + c_2 P_2 + \cdots + c_m P_m$$

where $P_C = \{c_1, c_2, \ldots, c_m\}$, $P \in S_{n-1}$ and $P_i \in S_{r-1}^{(i)}$ for $i = 1, 2, \ldots, m$.

Since $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$, for given level of C the distributions of $(A_1, A_2, \ldots, A_t)$ is an unconditional independence statement. The local point P_i, then, is on the model space I_i for $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$ within the local simplex $S_{r-1}^{(i)}$ for $i = 1, 2, \ldots, m$ (see

Figure 5.11). Thus, the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ can be regarded geometrically as a set convex hull in the global simplex S_{n-1} determined by

$$\mathrm{co}_s\{m, I\} = \{P = c_1P_1 + c_2P_2 + \cdots + c_mP_m \mid P_i \in I_i,\ c_i \geq 0 \text{ and } \sum c_i = 1 \text{ for all } i\}$$

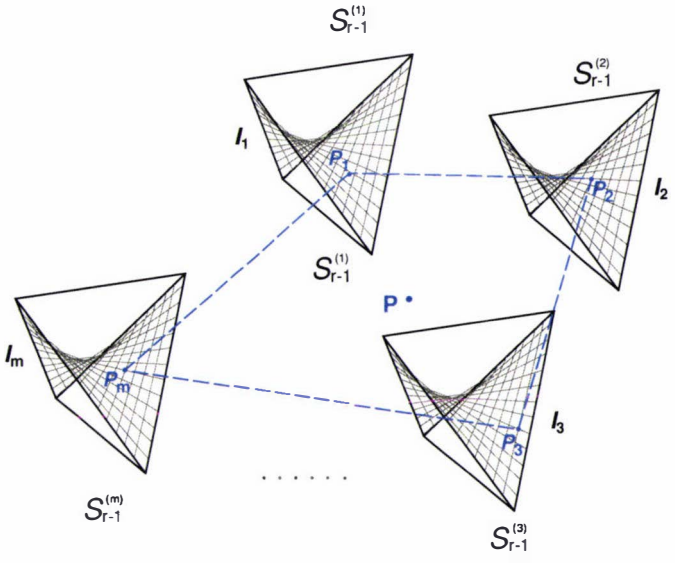


Figure 5.11: The global point P on the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ is a convex combination of the local points $P_1, P_2, \ldots, P_m$ where P_i is located on the ith copy of I, the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$, within the local simplex S_{r-1} for all i.

We now illustrate the model space for $A_1 \perp\!\!\!\perp A_2 \mid C$.

Example 5: When $m_1 = m_2 = 2$ and $m = 4$, the model space of $A_1 \perp\!\!\!\perp A_2 \mid C$ is described as follows

Step 1 Determine the unrestricted model space, the global simplex $S_{15} \subseteq \mathbf{R}^{16}$.

Step 2 Partition the vertices of S_{15} into four (m) lots of four (m_1m_2) vertices.

Step 3 Place four (m) copies of the independence model space of $A_1 \perp\!\!\!\perp A_2$, $I = \text{co}_c\{2, S_1\}$, in each of the partition simplexes, I_1, I_2, I_3, I_4. Here we have

$$I_1 = \{\text{co}_c\{2, S_1\}, \mathbf{0}, \mathbf{0}, \mathbf{0}\} \qquad I_2 = \{\mathbf{0}, \text{co}_c\{2, S_1\}, \mathbf{0}, \mathbf{0}\}$$

$$I_3 = \{\mathbf{0}, \mathbf{0}, \text{co}_c\{2, S_1\}, \mathbf{0}\} \qquad I_4 = \{\mathbf{0}, \mathbf{0}, \mathbf{0}, \text{co}_c\{2, S_1\}\}$$

where $\mathbf{0} = [0,\ 0,\ 0,\ 0]$.

Step 4 Take the four arbitrary points from four copies of I respectively and obtain the convex combination of these four points as $c_1P_1 + c_2P_2 + c_3P_3 + c_4P_4$, where $P_c = (c_1, c_2, c_3, c_4)$.

Step 5 Let P_i vary on I_i for all i to form the model space of $A_1 \perp\!\!\!\perp A_2 \mid C$

$$\text{co}_s\{4, I\} = \{c_1P_1 + c_2P_2 + c_3P_3 + c_4P_4 \mid P_i \in I_i \text{ for all } i\}$$

5.5 The MLE of a distribution satisfying $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$

In Section 5.4 we described the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ as a set convex hull $\text{co}_s\{m, I\}$. Here we will consider the MLE of the parameters of this model for an observed relative frequency table P using geometry. Birch (1963) shown that the MLE of the parameters of the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ is the unique point on the model space which matches the observed relative frequency table P with minimal sufficient statistics. For the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ the minimal sufficient statistics are the entries in the marginal tables $P_{A_1C}, P_{A_2C}, \ldots, P_{A_tC}$ (Bishop 1995, p.83) where P_{A_kC} is the joint marginal table of A_k and C for all k. We have two ways

to obtain the MLE of the parameters of the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$, one from a closed formula (Bishop, 1995) named the direct fitting method and another from the iterative proportional fitting method. The later method, formulated by Deming and Stephan in 1940, is a commonly used algorithm for MLE in loglinear models. This section discusses the geometric fitting of a CI statement using the two methods. Here we start with the fitting of an unconditional independence model.

Geometry of fitting of an unconditional independence model

A table of $(A_1, A_2, \ldots, A_t)$ is a t-way array indexed by $l = (l_1, l_2 \ldots, l_t)$ where l_k is an index of levels of A_k for $k = 1, 2, \ldots, t$. For the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$ with observed table $\{p_l\}$, the MLE table $\{\hat{p}_l\}$ is directly found by the formula

$$\hat{p}_l = p_{l_1+} p_{l_2+} \cdots p_{l_t+}$$

where '+' denotes summation over the remaining indices in l.

Let the observed table $\{p_l\}$ and its MLE table $\{\hat{p}_l\}$ correspond to points P and $\hat{P}$ in the simplex S_{r-1} (where $r = m_1 m_2 \ldots m_t$) respectively. From Birch (1963) $\hat{P}$ is on the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$ and has the same A_k margin as P for $k = 1, 2 \ldots, t$. Thus from Section 5.3, geometrically $\hat{P}$ is obtained by moving P to the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$ along the direction perpendicular to central simplexes $S_{m-1}^{A_1}, S_{m-1}^{A_2}, \ldots, S_{m-1}^{A_t}$ simultaneously.

Alternatively, $\hat{P}$ also can be found by the iterative proportional fitting method. The procedure is shown below.

Initial step

Set $k = 0$. Take as an initial estimate the table $\{\hat{p}_l^{(0)}\}$ which corresponds to the

point

$$\hat{P}^{(0)} = (\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r}) \in S_{r-1}$$

where $r = m_1 m_2 \ldots m_t$.

Iterative step

Step 1. Scale $\{\hat{p}_l^{(k)}\}$ to have margin $\{p_{l_1+}\}$ giving

$$\hat{p}_l^{(k+1,1)} = \hat{p}_l^{(k)} \frac{p_{l_1+}}{\hat{p}_{l_1+}^{(k)}}$$

where the first superscript of $\hat{p}_l^{(k+1,1)}$ refers to the iteration number, and the second to the step number within iterations. Now let $\{\hat{p}_l^{(k+1,1)}\}$ correspond to a point $\hat{P}^{(1)} \in S_{r-1}$.

Step 2. Scale $\{\hat{p}_l^{(k+1,1)}\}$ to have margin $\{p_{l_2+}\}$ giving

$$\hat{p}_l^{(k+1,2)} = \hat{p}_l^{(k+1,1)} \frac{p_{l_2+}}{\hat{p}_{l_2+}^{(k+1,1)}}$$

and let $\{\hat{p}_l^{(k+1,t-1)}\}$ correspond to a point $\hat{P}^{(t-1)} \in S_{r-1}$.

$\vdots$ $\vdots$

Step t. Scale $\{\hat{p}_l^{(k+1,t-1)}\}$ to have margin $\{p_{l_t+}\}$ giving

$$\hat{p}_l^{(k+1,t)} = \hat{p}_l^{(k+1,t-1)} \frac{p_{l_t+}}{\hat{p}_{l_t+}^{(k+1,t-1)}}$$

let $\{\hat{p}_l^{(k+1,t)}\}$ correspond to a point $\hat{P}^{(t)} \in S_{r-1}$ and $\{\hat{p}_l^{(k+1)}\} = \{\hat{p}_l^{(k+1,t)}\}$.

Stopping rule Stop when the fitted margin of A_k and the observed margin of A_k are sufficiently close for $k = 1, 2, \ldots t$. Otherwise, increment k and return to the iterative step.

Note that the MLE of the parameters of model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$ can be obtained from a closed formula, the algorithm will converge at the first iteration (Haberman 1974, p.197).

Recall that any point in the model space $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$ meets precisely t simplexes with the kth simplex has varying margin of A_k while all other margins remain constant for all k. From Section 5.4, geometrically, $\hat{P}$ is obtained by starting with the central point $\hat{P}^{(0)}$ of the simplex S_{n-1}. The point $\hat{P}^{(0)}$ is in the model space $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$, so the point $\hat{P}^{(1)}$ is found by moving $\hat{P}^{(0)}$ on the simplex with varying A_1 margin to meet the hyperplane with fixed margin P_{A_1} in S_{r-1}. Note that $\hat{P}^{(1)}$ has the same margin of A_1 as P, but the remaining margins as $\hat{P}^{(0)}$. Again $\hat{P}^{(1)}$ is in the model space $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$, then the point $\hat{P}^{(2)}$ is found by moving $\hat{P}^{(1)}$ on the simplex with varying A_2 margin to meet the hyperplane with fixed margin P_{A_2} in S_{r-1}. Now $\hat{P}^{(2)}$ has the same margins of A_1 and A_2 as P, but the remaining margins as $\hat{P}^{(0)}$. Following this pattern, we find points $\hat{P}^{(3)}, \hat{P}^{(4)}, \ldots, \hat{P}^{(t)}$, where finally the point $\hat{P}^{(t)}$ is on the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$ and matches the data point P in the minimal sufficient statistics, the marginal tables $P_{A_3}, P_{A_4}, \ldots, P_{A_t}$. Thus $\hat{P}^{(t)}$ is the fitted point $\hat{P}$.

Now we illustrate the fitting of the independence model of a 2×2 table using two methods just described.

Example 6 Consider an observed table P with binary variables X_1 and X_2 corresponding to a 4-tuple in lexicographical order X_1, X_2 of

$$(p_{11},\ p_{12},\ p_{21},\ p_{22}) = (0.2,\ 0.2,\ 0.1,\ 0.5)$$

Suppose $\hat{P} = (\hat{p}_{11},\ \hat{p}_{12},\ \hat{p}_{21},\ \hat{p}_{22})$ is the MLE of P for the model $X_1 \perp\!\!\!\perp X_2$. Then $\hat{P}$ can be found by the following two methods.

1. The maximum likelihood estimate $\hat{P}$ can be directly calculated by the formula $\hat{p}_{ij} = p_{i+}p_{+j}$, where $p_{i+} = p_{i1} + p_{i2}$ and $p_{+j} = p_{1j} + p_{2j}$ for all i, j. Then we have

$$\hat{P} = (\hat{p}_{11},\ \hat{p}_{12},\ \hat{p}_{21},\ \hat{p}_{22}) = (0.12,\ 0.28,\ 0.18,\ 0.42)$$

The fitted point $\hat{P}$ is a point on the model space of $X_1 \perp\!\!\!\perp X_2$ and has the same X_1 and X_2 margins as the observed point P. Thus geometrically (see Figure 5.12), the fitted point $\hat{P}$ can be found by sliding the observed point P down to the model space $\mathrm{co_c}\{2,\ S_1\}$, part of hyperbolic paraboloid, along the direction of IJ which preserves the X_1 and X_2 margins in the tetrahedron $ABCD$ (see Example 1).

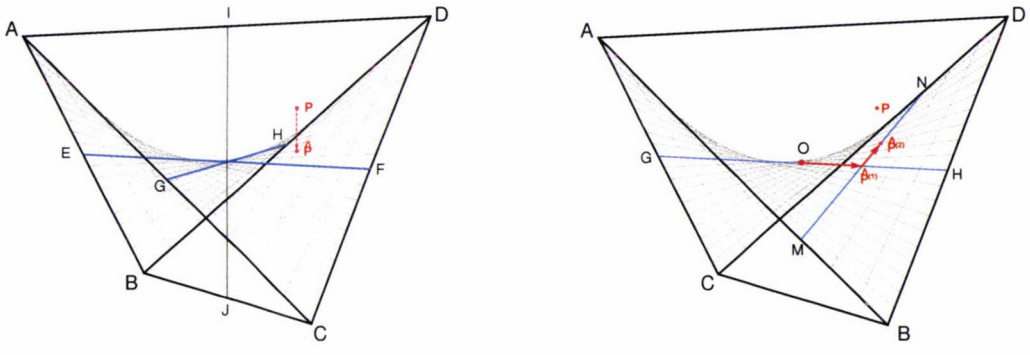


Figure 5.12: The left panel shows that for a given data point P, the MLE $\hat{P}$ for model $X_1 \perp\!\!\!\perp X_2$ is found by sliding P down to the model space of $X_1 \perp\!\!\!\perp X_2$ along the direction IJ, the direction for preserving X_1 and X_2 margins. Alternatively, the right panel shows that $\hat{P}$ can be found by starting with the central point O, then moving along GH (the simplex with varying X_1 margin) to point $\hat{P}^{(1)}$ with margin P_{X_1}, then shifting $\hat{P}^{(1)}$ along MN (the simplex with varying X_1 margin) to point $\hat{P}^{(2)}$ with margins P_{X_1} and P_{X_2}. Thus we find $\hat{P} = \hat{P}^{(2)}$.

2. On the other hand, the fitted point $\hat{P}$ can be found by the following procedure:

   1) Start with an initial estimate $\hat{P}^{(0)} = (0.25,\ 0.25,\ 0.25,\ 0.25)$.

2) Scaling $\hat{P}^{(0)}$ to match the X_1 margin of P giving $\hat{P}^{(1,1)} = (0.2,\ 0.2,\ 0.3,\ 0.3)$ using

$$\hat{p}_{ij}^{(1,1)} = \hat{p}_{ij}^{(0)} \frac{p_{i+}}{\hat{p}_{i+}^{(0)}} \text{ for all } i,\ j$$

3) Scaling $\hat{P}^{(1,1)}$ to match X_2 margin of P we obtain $\hat{P}^{(1,2)} = (0.12,\ 0.28,\ 0.18,\ 0.42)$ using

$$\hat{p}_{ij}^{(1,2)} = \hat{p}_{ij}^{(1,1)} \frac{p_{+j}}{\hat{p}_{+j}^{(1,1)}} \text{ for all } i,\ j$$

The point $\hat{P}^{(1,2)}$ is on the model space $X_1 \perp\!\!\!\perp X_2$ and matches P in the X_1 and X_2 margins. Thus we obtain the fitted point $\hat{P} = \hat{P}^{(1,2)}$.

To interpret the above procedure geometrically (see Figure 5.12), we can describe it alternatively as follows:

1) Start from the central point O of the tetrahedron $ABCD$, the initial estimate of $\hat{P}$.

2) Move the point O along the simplex GH, whose points have varying X_1 margin (but fixed X_2 margin) to the point $\hat{P}^{(1)}$ which has X_1 margin of $(0.4,\ 0.6)$.

3) Move the point $\hat{P}^{(1)}$ along the simplex MN, whose points have varying X_2 margin (but fixed X_1 margin) to obtain the point $\hat{P}^{(2)}$ which has X_2 margin of $(0.3,\ 0.7)$.

Now the point $\hat{P}^{(2)}$ is on the model space $X_1 \perp\!\!\!\perp X_2$ and has the same X_1 and X_2 margins as the given point P. According to Birch's result (1963) $\hat{P}^{(2)}$ is the required fitted point $\hat{P}$.

Geometry of fitting of a conditional independent statement

As described in Section 5.3, an observed contingency table for the variable sets $A_1, A_2, \ldots, A_t, C$ can be viewed as a point P in the global simplex S_{n-1}. The global point P is a convex combination of the local points $P_1, P_2, \ldots, P_m$, using as coefficients the marginal distribution of C. Specifically

$$P = c_1 P_1 + c_2 P_2 + \cdots + c_m P_m$$

where $P_C = \{c_1, c_2, \ldots, c_m\}$ with $c_i \geq 0$ and $\sum c_i = 1$ for $i = 1, 2 \ldots, m$.

Let $\hat{P}$ be the MLE of P for the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$. Then the fitted point $\hat{P}$ is located on the set convex hull $\mathrm{co}_s\{m, I\}$ where I is the model space of $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$, and

$$\hat{P} = \hat{c}_1 \hat{P}_1 + \hat{c}_2 \hat{P}_2 + \ldots + \hat{c}_m \hat{P}_m \tag{5.8}$$

where $\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_m$ are the fitted local points located on m copies of I in S_{n-1}, and $\{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_m\}$, denoted by $\hat{P}_C$, is the C margin of $\hat{P}$.

Denote the joint distribution of A_k and C for P and $\hat{P}$ by P_{A_kC} and $\hat{P}_{A_kC}$ respectively for all k, and the conditional distribution of A_k for given level of C for P and $\hat{P}$ by $P_{A_k|C}$ and $\hat{P}_{A_k|C}$ respectively for all k. From Birch's results (1963) we know that the fitted point $\hat{P}$ on the set convex hull $\mathrm{co}_s\{m, I\}$ is uniquely determined by the observed point P with $\hat{P}_{A_kC} = P_{A_kC}$ which indicates $\hat{P}_C = P_C$ and $\hat{P}_{A_k|C} = P_{A_k|C}$. Thus Equation (5.8) becomes

$$\hat{P} = c_1 \hat{P}_1 + c_2 \hat{P}_2 + \ldots + c_m \hat{P}_m$$

Recall that P_i (or $\hat{P}_i$) represents the conditional distribution of $A_1, A_2, \ldots, A_t$ for given ith level of C for all i. From $\hat{P}_{A_k|C} = P_{A_k|C}$ for all k we have that the fitted

local point $\hat{P}_i$ has the same $A_1, A_2, \ldots, A_t$ margins as the given local point P_i in the local simplex $S_{r-1}^{(i)}$ for $i = 1, 2, \ldots, m$.

Therefore, to find the fitted point $\hat{P}$ we only need locate the fitted local points $\hat{P}_i$ for all i. Again a fitted local point $\hat{P}_i$ is uniquely determined by the given local point P_i in the model space I_i (the ith copy of I) in the local simplex $S_{r-1}^{(i)}$ for all i. Thus, the problem of fitting a conditional independence model reduces to fitting an independence model. The fitting process of the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ then will be

Step 1 Represent the data point P in the form $P = c_1 P_1 + c_2 P_2 + \ldots + c_m P_m$ where $P_1, P_2, \ldots, P_m$ are associated local points.

Step 2 Find the local fitted point $\hat{P}_i$ in the model space $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$ from the local point P_i in $S_{r-1}^{(i)}$ using the two methods discussed in the last section for $i = 1, 2, \ldots, m$.

Step 3 Obtain the MLE point $\hat{P}$ by the following convex combination

$$\hat{P} = c_1 \hat{P}_1 + c_2 \hat{P}_2 + \ldots + c_m \hat{P}_m$$

where $P_C = \{c_1, c_2, \ldots, c_m\}$.

The two methods for fitting the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ are now discussed separately. Note that we only demonstrate the two methods geometrically here, so the numerical results in Example 7 and 8 are quoted from the algebraic representation of the two methods.

The direct fitting method

Using the direct fitting method the MLE for the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ is obtained directly from a closed formula (Bishop et al., 1975). Here we describe the direct fitting method geometrically in three steps

Step 1 Start with the given point $P = c_1P_1 + c_2P_2 + \ldots + c_mP_m$ where $P_1, P_2, \ldots, P_m$ are associated local points.

Step 2 Move each local point P_i to obtain $\hat{P}_i$ on I_i, the copy of the corresponding point convex hull I. The direction $P_i\hat{P}_i$ is perpendicular to t central simplexes in $S_{r-1}^{(i)}$ for all i.

Step 3 Obtain the MLE point

$$\hat{P} = c_1\hat{P}_1 + c_2\hat{P}_2 + \ldots + c_m\hat{P}_m.$$

We illustrate this procedure by Figure 5.13 and Example 7.

Example 7: To a given data set involving binary variables X_1, X_2, X_3 and X_4, it corresponds a 16-tuple by lexicographical order X_3, X_4, X_1, X_2, which is

$$\begin{array}{cccccccc} (0.0167 & 0.0067 & 0.0317 & 0.1083 & 0.0317 & 0.0083 & 0.0183 & 0.0100 \\ 0.1300 & 0.0417 & 0.0983 & 0.0150 & 0.2200 & 0.0150 & 0.0867 & 0.1617) \end{array}$$

We find the MLE of $X_1 \perp\!\!\!\perp X_2 \mid X_3, X_4$ in the following steps.

Step 1 Represent the given point P as a convex combination of local points P_1, P_2, P_3, P_4, such that

$$P = 0.1634P_1 + 0.0683P_2 + 0.285P_3 + 0.4834P_4$$

with

$$P_1 = (p_1,\ \mathbf{0},\ \mathbf{0},\ \mathbf{0}), \qquad P_2 = (\mathbf{0},\ p_2,\ \mathbf{0},\ \mathbf{0}),$$

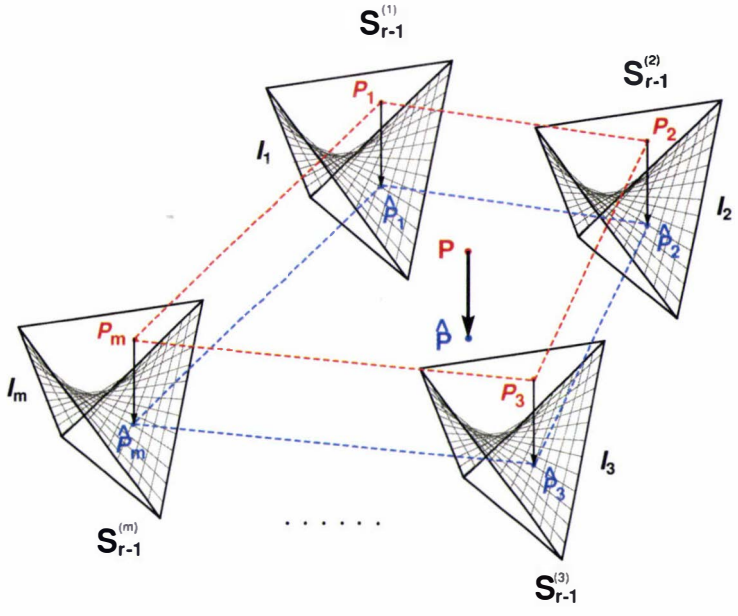


Figure 5.13: The graphic shows, for a given data point P, how to find the MLE for the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$. First we start with given local points $P_1, P_2, \ldots, P_m$, then drop each local points along the direction preserving $A_1, A_2, \ldots, A_t$ margins onto the model space $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t$, to obtain the MLE local points $\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_m$. The MLE point $\hat{P}$, then, is a convex combination of $\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_m$ using as coefficients the marginal distribution P_C.

$$P_3 = (\mathbf{0},\ \mathbf{0},\ p_3,\ \mathbf{0}), \qquad P_4 = (\mathbf{0},\ \mathbf{0},\ \mathbf{0},\ p_4)$$

where $\{0.1634,\ 0.0683,\ 0.285,\ 0.4834\}$ is the joint distribution of X_3 and X_4,

$$\begin{aligned} p_1 &= (0.1020,\ 0.0408,\ 0.1939,\ 0.6633), \\ p_2 &= (0.4634,\ 0.1220,\ 0.2683,\ 0.1463), \\ p_3 &= (0.4561,\ 0.1462,\ 0.3450,\ 0.0526), \\ p_4 &= (0.4552,\ 0.0310,\ 0.1793,\ 0.3345) \end{aligned}$$

and $\mathbf{0} = [0,\ 0,\ 0,\ 0]$.

Step 2 Find the MLE local points $\hat{P}_i$ by moving P_i down to the model space of

$X_1 \perp\!\!\!\perp X_2$ along the direction for preserving X_1 and X_2 margins, we obtain

$$\hat{P}_1 = (\hat{p}_1,\ \mathbf{0},\ \mathbf{0},\ \mathbf{0}), \qquad \hat{P}_2 = (\mathbf{0},\ \hat{p}_2,\ \mathbf{0},\ \mathbf{0}),$$

$$\hat{P}_3 = (\mathbf{0},\ \mathbf{0},\ \hat{p}_3,\ \mathbf{0}), \qquad \hat{P}_4 = (\mathbf{0},\ \mathbf{0},\ \mathbf{0},\ \hat{p}_4)$$

where

$$\begin{aligned}
\hat{p}_1 &= (0.0423 \quad 0.1006 \quad 0.2536 \quad 0.6035),\\
\hat{p}_2 &= (0.4283 \quad 0.1570 \quad 0.3034 \quad 0.1112),\\
\hat{p}_3 &= (0.4826 \quad 0.1198 \quad 0.3186 \quad 0.0791),\\
\hat{p}_4 &= (0.3085 \quad 0.1777 \quad 0.3260 \quad 0.1878)
\end{aligned}$$

Step 3 Obtain the MLE point $\hat{P}$ by

$$\begin{aligned}
\hat{P} &= 0.1634\hat{P}_1 + 0.0683\hat{P}_2 + 0.285\hat{P}_3 + 0.4834\hat{P}_4\\
&= \begin{matrix}(0.0069 & 0.0164 & 0.0414 & 0.0986 & 0.0293 & 0.0107 & 0.0207 & 0.0076\\ 0.1375 & 0.0341 & 0.0908 & 0.0225 & 0.1491 & 0.0859 & 0.1576 & 0.0908).\end{matrix}
\end{aligned}$$

The iterative proportional fitting method

On the other hand, the MLE for the model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ can be found by the iterative proportional fitting method which is described geometrically in the following steps.

Step 1 Represent the data point P in the form $P = c_1P_1 + c_2P_2 + \ldots + c_mP_m$ where $P_1, P_2, \ldots, P_m$ are associated local points.

Step 2 Start with the local central points $\hat{P}_1^{(0)}, \hat{P}_2^{(0)}, \ldots, \hat{P}_m^{(0)}$.

Move the local point $\hat{P}_i^{(0)}$ on the central simplex with varying A_1 margin in I_i to find a point $\hat{P}_i^{(1)}$ which has the same A_1 margin as P_i for all i.

Move the local point $\hat{P}_i^{(1)}$ on the central simplex with varying A_2 margin in I_i to find a point $\hat{P}_i^{(2)}$ which has the same margins A_1, A_2 as P_i for all i.

$\vdots \qquad\qquad \vdots$

Move the local point $\hat{P}_i^{(t-1)}$ on the simplex with varying A_t margin in I_i to find a point $\hat{P}_i^{(t)}$ which has the same margins $A_1, A_2, \ldots, A_t$ as P_i for all i.

Step 3 Obtain the fitted point

$$\hat{P} = c_1 P_1^{(t)} + c_2 P_2^{(t)} + \ldots + c_m \hat{P}_m^{(t)}.$$

Again, the MLE of the parameters of model $A_1 \perp\!\!\!\perp A_2 \ldots \perp\!\!\!\perp A_t \mid C$ can be obtained from a closed formula, the algorithm will converge at the first iteration (Haberman 1974, p.197).

The procedure above is demonstrated in Figure 5.14 and Example 8.

Example 8: We will perform the iterative proportional fitting method to find the MLE of $X_1 \perp\!\!\!\perp X_2 \mid X_3, X_4$ for the data set given in Example 7.

Step 0 Start with the local central points

$$\hat{P}_1^{(0)} = (\hat{p}_1^{(0)},\ \mathbf{0},\ \mathbf{0},\ \mathbf{0}), \qquad \hat{P}_2^{(0)} = (\mathbf{0},\ \hat{p}_2^{(0)},\ \mathbf{0},\ \mathbf{0}),$$

$$\hat{P}_3^{(0)} = (\mathbf{0},\ \mathbf{0},\ \hat{p}_3^{(0)},\ \mathbf{0}), \qquad \hat{P}_4^{(0)} = (\mathbf{0},\ \mathbf{0},\ \mathbf{0},\ \hat{p}_4^{(0)})$$

where $\hat{p}_i^{(0)} = (0.2500\ 0.2500\ 0.2500\ 0.2500)$ for all i and $\mathbf{0} = [0,\ 0,\ 0,\ 0]$.

Step 1 Move each $\hat{P}_i^{(0)}$ on the associated copy of model space of $X_1 \perp\!\!\!\perp X_2$ along varying X_1 margin direction to find point $\hat{P}_i^{(1)}$ which has the same X_1 margin as the data local point P_i for all i, which are

$$\hat{P}_1^{(1)} = (\hat{p}_1^{(1)},\ \mathbf{0},\ \mathbf{0},\ \mathbf{0}), \qquad \hat{P}_2^{(1)} = (\mathbf{0},\ \hat{p}_2^{(1)},\ \mathbf{0},\ \mathbf{0}),$$

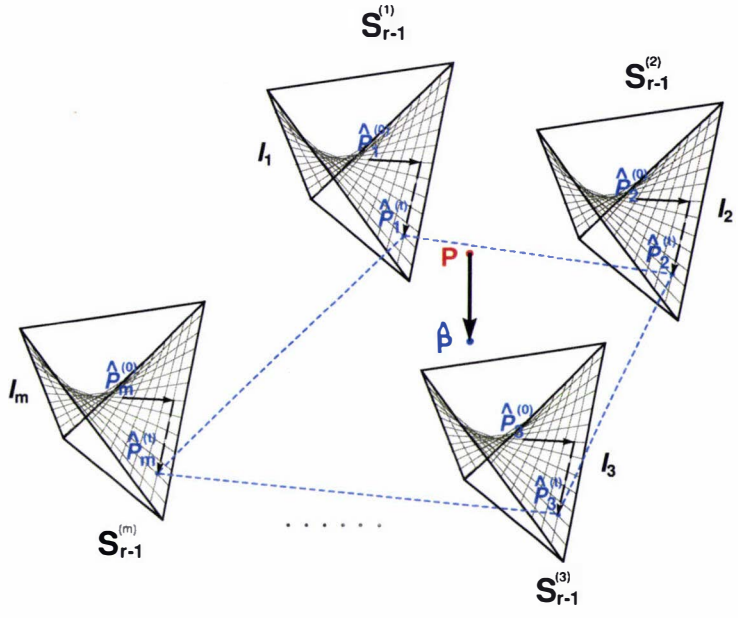


Figure 5.14: The working of the iterative proportional fitting method: we start with the local central point $\hat{P}_i^{(0)}$, then move $\hat{P}_i^{(0)}$ on the copy of independence model space I_i step by step to meet $\hat{P}_i^{(t)}$ for all i. In each step match the sufficient statistics P_{A_k} cummulatively for all k. The MLE point $\hat{P}$ is the convex combination of $\hat{P}_1^{(t)}, \hat{P}_2^{(t)}, \ldots, \hat{P}_m^{(t)}$.

$$\hat{P}_3^{(1)} = (0,\ 0,\ \hat{p}_3^{(1)},\ 0), \qquad \hat{P}_4^{(1)} = (0,\ 0,\ 0,\ \hat{p}_4^{(1)})$$

where

$$\begin{aligned}
\hat{p}_1^{(1)} &= (0.0716 \quad 0.0716 \quad 0.4284 \quad 0.4284),\\
\hat{p}_2^{(1)} &= (0.2928 \quad 0.2928 \quad 0.2072 \quad 0.2072),\\
\hat{p}_3^{(1)} &= (0.3012 \quad 0.3012 \quad 0.1988 \quad 0.1988),\\
\hat{p}_4^{(1)} &= (0.2431 \quad 0.2431 \quad 0.2569 \quad 0.2569)
\end{aligned}$$

Step 2 Move each $\hat{P}_i^{(1)}$ on the associated copy of model space of $X_1 \perp\!\!\!\perp X_2$ along varying X_2 margin direction to find point $\hat{P}_i^{(2)}$ who has the same X_2 margin as the data local point P_i for all i, Now $\hat{P}_i^{(2)}$ and P_i have the same X_1 and X_2

margins. which are

$$\hat{P}_1^{(2)} = (\hat{p}_1^{(2)},\ 0,\ 0,\ 0), \qquad \hat{P}_2^{(2)} = (0,\ \hat{p}_2^{(2)},\ 0,\ 0),$$

$$\hat{P}_3^{(2)} = (0,\ 0,\ \hat{p}_3^{(2)},\ 0), \qquad \hat{P}_4^{(2)} = (0,\ 0,\ 0,\ \hat{p}_4^{(2)})$$

where

$$\begin{aligned}
\hat{p}_1^{(2)} &= (0.0423 \quad 0.1006 \quad 0.2536 \quad 0.6035),\\
\hat{p}_2^{(2)} &= (0.4283 \quad 0.1570 \quad 0.3034 \quad 0.1112),\\
\hat{p}_3^{(2)} &= (0.4826 \quad 0.1198 \quad 0.3186 \quad 0.0791),\\
\hat{p}_4^{(2)} &= (0.3085 \quad 0.1777 \quad 0.3260 \quad 0.1878)
\end{aligned}$$

Step 3 Obtain the MLE point

$$\begin{aligned}
\hat{P} &= 0.1634\hat{P}_1^{(2)} + 0.0683\hat{P}_2^{(2)} + 0.285\hat{P}_3^{(2)} + 0.4834\hat{P}_4^{(2)}\\
&= \begin{array}{l}(0.0069 \quad 0.0164 \quad 0.0414 \quad 0.09860.0293 \quad 0.0107 \quad 0.0207 \quad 0.0076\\ \quad 0.1375 \quad 0.0341 \quad 0.0908 \quad 0.02250.1491 \quad 0.0859 \quad 0.1576 \quad 0.0908)\end{array}
\end{aligned}$$

where $\{0.1634,\ 0.0683,\ 0.285,\ 0.4834\}$ is the joint distribution of X_3 and X_4.

Both the direct fitting method and iterative proportional fitting method find the fitted point which matches the minimum sufficient statistics of the data point on the model space. The direct fitting method, however, moves the data point to the fitted point directly along the direction for preserving all minimum sufficient statistics simultaneously. On the other hand, the iterative proportional fitting method starts at the centre of the model space and then moves on the model space to match each minimum sufficient statistic cumulatively.

5.6 Conclusion

In this chapter we have considered the geometry of graphical loglinear models 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 corresponding point convex hull is a union of simplexes defined by the corresponding points, while the set convex hull is union of convex hulls whose vertices are on the associated corresponding point convex hulls. Meanwhile, we discussed the geometric framework for three kinds of distributions for categorical variables: the joint distribution, the marginal distribution and the conditional distribution. Based on the above geometric settings, we have illustrated the finding of the MLE for one CI statement geometrically using the direct fitting method and iterative proportional fitting method.

Chapter 6

Conclusions

6.1 Thesis Contribution

The geometry of generalized linear models has been discussed in this thesis. The main findings are now listed and briefly discussed.

1. A geometric framework for generalized linear models

An observation with n values is viewed as a vector in Euclidian space $\mathbf{R}^n$, This space then is partitioned into two orthogonal spaces, the sufficiency space $\mathcal{S}$ and the auxiliary space $\mathcal{A}$. Two mean sets are introduced in $\mathbf{R}^n$, related to generalized linear models, namely $\mathcal{M}_{\mathbf{R}}$, the untransformed model space and $g(\mathcal{M}_{\mathbf{R}})$, the link transformed model space. When a generalized linear model employs a canonical link there are two properties drawn out related to the maximum likelihood estimate of the parameters of the model. In the untransformed model space, the coefficients of the basis of the sufficiency space, the sufficient statistics, are preserved in the fitting process. In the link-transformed model space, the coefficients of the basis of the auxiliary space are zeroed in the fitting

process. Linear models and loglinear models become special cases of generalized linear models with identity and log link respectively. The following tables summarize the main results for linear models (Chapter 1), loglinear models (Chapter 2) and generalized linear models (Chapter 3).

Model	Linear models
Observation	$y \in \mathbf{R}^n$
$\mathcal{M}_{\mathbf{R}}$	$\mathbb{M} = \{\mu : \mu = X\beta,\ \beta \in \mathbf{R}^q\}$
$g(\mathcal{M}_{\mathbf{R}})$	$\mathbb{M} = \{\mu : \mu = X\beta,\ \beta \in \mathbf{R}^q\}$
$\mathbf{R}^n$	$\mathbf{R}^n = \mathbb{M} \oplus \mathbb{E}$ where $\mathbb{M} = \text{span}\{x_1, x_2, \ldots, x_q\}$ and $\mathbb{E} = \mathbb{M}^{\perp}$
MLE ($\hat{\mu}$) fitting properties	$X\hat{\mu} = Xy$ and $\hat{\mu} \in \mathbb{M}$

Model	Loglinear models
Observation	$p \in S_{n-1}$ where S_{n-1} is an $n-1$ dimensional simplex
$\mathcal{M}_{\mathbf{R}}$	$\{\pi : \pi_i \in [0,1]$ and $\sum \pi_i = 1\}$
$g(\mathcal{M}_{\mathbf{R}})$	$\{\log \pi : \log \pi = X\beta,\ \beta \in \mathbf{R}^q\}$
$\mathbf{R}^n$	$\mathbf{R}^n = \mathcal{S} \oplus \mathcal{A}$ where $\mathcal{S} = \text{span}\{x_1, x_2, \ldots, x_q\}$ and $\mathcal{A} = \mathcal{S}^{\perp}$
MLE ($\hat{\pi}$) fitting properties	$X\hat{\pi} = Xp$ and $\log(\hat{\pi}) \in \mathcal{S}$

Model	Generalized linear models
Observation	$y \in \mathbf{R}^n$
$\mathcal{M}_{\mathbf{R}}$	$\{\mu(f) : f \in \mathcal{M}\}$ where $\mathcal{M}$ is a q-dimensional set of density functions
$g(\mathcal{M}_{\mathbf{R}})$	$\{g(\mu) : g(\mu) = X\beta,\ \beta \in \mathbf{R}^q\}$
$\mathbf{R}^n$	$\mathbf{R}^n = \mathcal{S} \oplus \mathcal{A}$ where $\mathcal{S} = \text{span}\{x_1, x_2, \ldots, x_q\}$ and $\mathcal{A} = \mathcal{S}^{\perp}$
MLE ($\hat{\mu}$) fitting properties	$X\hat{\mu} = Xy$ and $g(\hat{\mu}) \in \mathcal{S}$

2. **A new algorithm for fitting generalized linear models with canonical link**

   This algorithm, discussed in Chapter 4, is based on the two properties of the maximum likelihood estimate of the parameters of the model. There are two projections performed alternately in the algorithm, orthogonal projection onto the sufficiency affine plane and non-orthogonal projection onto the transformed model space. In the process, we match the model space in the transformed world and sufficient statistics in the untransformed world iteratively until convergence. The new algorithm becomes the scoring method after linearization.

3. **A geometric description of the model space of a conditional independence statement**

   A geometric description of the model space for a conditional independence statement was constructed in Chapter 5, using the concepts of "corresponding point convex hull" and "set convex hull". In this geometric framework the fitting of one conditional independence statement was discussed using the direct fitting method and the iterative proportional fitting method.

6.2 Further Research Directions

Some further research directions suggested by the work of this thesis are now outlined as follows.

1. The geometry of generalized linear models developed here relies on the canonical link, so extension to generalized linear models with non-canonical link is needed.

2. In the thesis we only discussed the conditions under which a fixed point exist, when the fixed point is unique and how the algorithm converges to the fixed point. Thus it is required to prove that the new algorithm matches those conditions.

3. The comparison of the new algorithm and the scoring method was made using numerical results. To understand the difference between the two methods, further study is needed of the theoretical background. This depends heavily on the convergence proof for the new algorithm.

4. The model space of one conditional independence statement is described geometrically in the thesis. A graphical loglinear model, however, is an intersection of a finite set of conditional independence statements. A neat geometric description of the model space for an intersection of conditional independence statements is required for graphical loglinear models.

Appendix A

Matlab functions

In this appendix we show Matlab functions `Nglmfit` and `Sglmfit`, coded by the author, for fitting a generalized linear model with canonical link using the new algorithm and the scoring method respectively.

```
function [b,fit,iter,flop]=Nglmfit(x,y,distr)
```

`Nglmfit` fits a generalized linear model with canonical link using the new algorithm.

`[b,Fit,iter,flop]=Nglmfit(x,y,distr)` fits a generalized linear model using the design matrix `x`, response `y`, and distribution `distr`. The result `b` is a vector of coefficient estimates. The result `fit` is a vector of fitted value of the response `y`. The result `iter` is the number of iterations needed for convergence. The result `flop` is the approximate number of floating point operations when the algorithm converges. Acceptable values for `distr` are `'normal'`, `'binomial'`, `'poisson'`, `'gamma'`, and `'inverse gaussian'`. The distribution parameter is fit as a function of the `x` columns using the canonical link.

Example:

```
b = Nglmfit(x, [y N], 'binomial')
```

This example fits a logistic regression model for `y` on `x`. Each `y(i)` is the number of successes in `N(i)` trials.

```
flops(0);
if (nargin<2), error('At least two arguments are required'); end
if (nargin<3 | isempty(distr)), distr = 'normal'; end
xx=[ones(size(x,1),1) x];
[n p]=size(xx);
%Orthonomaliz the design matrix x
A=orth([xx eye(n)]);
x=A(:,1:p);

convcrit = 1e-6;
b = zeros(p,1);
b0 = b+1;
iter = 0;
iterlim = 200;
switch(distr)
      case 'normal'

            y0 = y;
           eta = y0;

      case 'binomial'

           if (size(y,2) =2), error('Y must have two columns.'); end
           N = y(:,2);
           y = y(:,1)./ N;
           y0 = (N.*y + 0.5)./ (N + 1);
           eta = log(y0./(1-y0));

      case 'poisson'

           y0 = y + 0.25;
           eta = log(y0);
```

```
      case 'gamma'

           if (any(y(:)<=0))
           delta = min(abs(y(y =0))) * .001;
           y0 = max(delta, y);
           else
           y0=y;
           end
           eta = 1./y0;

      case 'inverse gaussian'

           if (any(y(:)<=0))
           delta = min(abs(y(y =0))) * .001;
           y0 = max(delta, y);
           else
           y0=y;
           end
           eta = 1./(y0.^2);

      otherwise, error('Distribution name is invalid.');

end
while(1)

      iter = iter+1;
      % Compute parameter by using inverse link function
      switch(distr)

           case 'normal'

               t = eta;
               mu=(t+(x*x')*(y-t));

           case 'binomial'
```

```
        p = 1 ./ (1 + exp(-eta));
        t=(p.*N+(x*x')*(y.*N-p.*N));
        mu=max(eps,min(1-eps,t./N));
    case 'poisson'
        t = exp(eta);
        mu=(t+(x*x')*(y-t));
        mu = max(eps,mu);
    case 'gamma'
        t = 1./eta;
        mu=(t+(x*x')*(y-t));
        mu = max(eps,mu);
    case 'inverse gaussian'
        t = 1./(sqrt(eta));
        mu=(t+(x*x')*(y-t));
        mu = max(eps,mu);
end
% Compute adjusted dependent variable for least squares fit
switch(distr)
    case 'normal'
        z = mu;
    case 'binomial'
        z=log(mu ./ (1-mu));
    case 'poisson'
        z = log(mu);
    case 'gamma'
        z = 1./mu;
```

```
        case 'inverse gaussian'
            z = 1./(mu.^2);
    end
    % Check stopping conditions
    if (norm(b-b0) < convcrit), break; end
    if (iter>iterlim), warning('Iteration limit reached.'); break; end
    % Compute weight function as the inverse of the variance function
    switch(distr)
        case 'normal'
            w =ones(size(y,1),1);
        case 'binomial'
            w=N.*(mu .* (1-mu));
        case 'poisson'
            w =mu;
        case 'gamma'
            w =mu.^2;
        case 'inverse gaussian'
            w =mu.^3;
    end
    % Compute coefficient estimates for this iteration
    b0 = b;
    [b,R] = wfit(z , x, w);
    % Form current linear combination
    eta = x * b;
end
% Calculate the fitted value
switch(distr)
```

```
      case 'normal'

           fit = x*b;

      case 'binomial'

           fit = N ./ (1 + exp(-x*b));

      case 'poisson'

           fit = exp(x*b);

      case 'gamma'

           fit = 1./x*b;

      case 'inverse gaussian'

           fit = 1./sqrt(x*b);

end
% Compute coefficient estimates based on the unorthonomalized design matrix xx
b=wfit(z,xx,w);
flop=flops;


% Perform a weighted least squares fit function
[b,R]=wfit(y,x,w)
sw = sqrt(w);
[r c] = size(x);
yw = y .* sw;
xw = x .* sw(:,ones(1,c));
[Q,R]=qr(xw,0);
b = R\(Q'*yw);
```

```
function [b,fit,iter,flop]=Sglmfit(x,y,distr)
```

`Sglmfit` fits a generalized linear model with canonical link using the scoring method.

`[b,Fit,iter,flop]=Sglmfit(x,y,distr)` fits a generalized linear model using the design matrix `x`, response `y`, and distribution `distr`. The result `b` is a vector of coefficient estimates. The result `fit` is a vector of fitted value of the response `y`. The result `iter` is the number of iterations needed for convergence. The result `flop` is the approximate number of floating point operations when the algorithm converges. Acceptable values for `distr` are `'normal'`, `'binomial'`, `'poisson'`, `'gamma'`, and `'inverse gaussian'`. The distribution parameter is fit as a function of the `x` columns using the canonical link.

Example:

```
b = Sglmfit(x, [y N], 'binomial')
```

This example fits a logistic regression model for `y` on `x`. Each `y(i)` is the number of successes in `N(i)` trials.

```
flops(0);
if (nargin<2), error('At least two arguments are required'); end
if (nargin<3 | isempty(distr)), distr = 'normal'; end
xx=[ones(size(x,1),1) x];
[n p]=size(xx);
% Orthonormalize the design matrix x
A=orth([xx eye(n)]);
x=A(:,1:p);

convcrit = 1e-6;
b = zeros(p,1);
b0 = b+1;
iter = 0;
iterlim = 200;
switch(distr)
```

```
case 'normal'

     y0 = y;
    eta = y0;

case 'binomial'

    if (size(y,2) =2), error('Y must have two columns.'); end
    N = y(:,2);
    y = y(:,1)./ N;
    y0 = (N.*y + 0.5)./ (N + 1);
    eta = log(y0./(1-y0));

case 'poisson'

    y0 = y + 0.25;
    eta = log(y0);

case 'gamma'

    if (any(y(:)<=0))
    delta = min(abs(y(y =0))) * .001;
    y0 = max(delta, y);
    else
    y0=y;
    end
    eta = 1./y0;

case 'inverse gaussian'

    if (any(y(:)<=0))
    delta = min(abs(y(y =0))) * .001;
    y0 = max(delta, y);
    else
    y0=y;
```

```
            end
            eta = 1./(y0.^2);
      otherwise, error('Distribution name is invalid.');
end
while(1)
      iter = iter+1;
      % Compute parameter by using inverse link function
      switch(distr)
            case 'normal'
                mu=eta;
            case 'binomial'
                mu = 1 ./ (1 + exp(-eta));
                mu = max(eps, min(1-eps, mu));
            case 'poisson'
                mu = exp(eta);
                mu = max(0,mu);
            case 'gamma'
                mu = 1./eta;
                mu = max(0,mu);
            case 'inverse gaussian'
                 mu = 1./sqrt(eta);
                mu = max(0,mu);
      end
      % Compute adjusted dependent variable for least squares fit
      switch(distr)
```

```
    case 'normal'
        z = eta;
    case 'binomial'
        z = eta + (y - mu)./( mu .* (1-mu));
    case 'poisson'
        z = eta + (y - mu)./mu;
    case 'gamma'
        z = eta - (y - mu)./(mu.^2);
    case 'inverse gaussian'
        z = eta - 2*((y - mu)./(mu.^3));
end
% Check stopping conditions
if (norm(b-b0) < convcrit), break; end
if (iter>iterlim), warning('Iteration limit reached.'); break; end
% Compute weight function as the inverse of the variance function
switch(distr)
    case 'normal'
        w =ones(size(y,1),1);
    case 'binomial'
        w=N.*(mu .* (1-mu));
    case 'poisson'
        w =mu;
    case 'gamma'
        w =mu.^2;
    case 'inverse gaussian'
```

```
            w =mu.^3;
    end
    % Compute coefficient estimates for this iteration
    b0 = b;
    [b,R] = wfit(z , x, w);
    % Form current linear combination
    eta = x * b;
end
% Calculate the fitted value
switch(distr)
    case 'normal'
        fit = x*b;
    case 'binomial'
        fit = N ./ (1 + exp(-x*b));
    case 'poisson'
        fit = exp(x*b);
    case 'gamma'
        fit = 1./x*b;
    case 'inverse gaussian'
        fit = 1./sqrt(x*b);
end
% Compute coefficient estimates based on the unorthonomalized design matrix xx
b=wfit(z,xx,w);
flop=flops;
```

```
% Perform a weighted least squares fit function
[b,R]=wfit(y,x,w)
sw = sqrt(w);
[r c] = size(x);
yw = y .* sw;
xw = x .* sw(:,ones(1,c));
[Q,R]=qr(xw,0);
b = R\(Q'*yw);
```

Appendix B

Data sources

In this appendix we show the data sources for the models displayed in Table 4.4 and Table 4.5. Here x_i indicates a covariate, f_i represents a factor and $f_i f_j$ denotes the interaction between factors f_i and f_j for all i, j.

1 Model

$$\texttt{Killed/Total}(y/N) = \texttt{Dose}(x)$$

Data from "Dobson, A. (1990). *An Introduction to Generalized Linear Models.* p.109".

2 Model

$$\texttt{satell/cases}(y/N) = \texttt{width}(x)$$

Data from "Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* p.106".

3 Model

$$\texttt{yes/cases}(y/N) = \texttt{race}(f_1) + \texttt{azt}(f_2)$$

Data from "Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* p.119".

4 Model

$$\texttt{y/n}(y/N) = \texttt{width}(x)$$

Data from "Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* p.82".

5 Model

$$\begin{aligned}\texttt{Pain}(y) \;=\; & \texttt{Age}(x_1) + \texttt{Duration}(x_2) + \texttt{Treatment}(f_1) + \texttt{Sex}(f_2) \\ & + \texttt{Treatment*Sex}(f_1 f_2)\end{aligned}$$

Data from "SAS Institute Inc. (1999). *SAS OnlineDoc.* Example 39.3".

6 Model

$$\texttt{wheeze}(y) = \texttt{city}(f) + \texttt{age}(x_1) + \texttt{smoke}(x_2)$$

Data from "Ware, J.H., Dockery, Spiro A. III, Speizer, F.E., and Ferris, B.G., Jr. (1984). Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases* 129: 366–374".

7 Model

$$\texttt{Kyphosis/Total}(y/N) = \texttt{Age}(x_1) + \texttt{Number}(x_2) + \texttt{Start}(x_3)$$

Data from "Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* p.301".

8 Model

$$\texttt{days}(y) = \texttt{sex}(f_1) + \texttt{origin}(f_2) + \texttt{type}(f_3) + \texttt{grade}(f_4)$$

Data from "Der, G. and Everitt, B. S. (2002). *A Handbook of Statistical Analyses Using SAS* (2nd edition). p.118".

9 Model

$$\texttt{satell}(y) = \texttt{width}(x)$$

Data from "Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* p.82".

10 Model

$$\texttt{satell}(y) = \texttt{width}(x)$$

Data from "Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* p.90".

11 Model

$$\texttt{Count}(y) = \texttt{Gender}(f_1) + \texttt{Type}(f_2)$$

Data from "Dunn, P. K. (2000). *glmlab Using MATLAB for Analysing Generalised Linear Models.* p.26".

12 Model

$$\begin{aligned}\texttt{count}(y) \;=\; & \texttt{Alcohol}(f_1) + \texttt{Cigarette}(f_2) + \texttt{Marijuana}(f_3) \\ & + \texttt{Alcohol*Marijuana}(f_1 f_3) + \texttt{Cigarette*Marijuana}(f_2 f_3)\end{aligned}$$

Data from "Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* p.152".

13 Model

$$\texttt{count}(y) = \texttt{assoc}(x) + \texttt{premar}(f_1) + \texttt{birth}(f_2)$$

Data from "Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* p.181".

14 Model

$$\texttt{Time}(y) = \texttt{log(WBC)}(x) + \texttt{Age}(f) + \texttt{log(WBC)*Age}(xf)$$

Data from "Dunn, P. K. (2000). *glmlab Using MATLAB for Analysing Generalised Linear Models.* p.32".

15 Model

$$\texttt{Costs}(y) = \texttt{PolAge}(f_1) + \texttt{CarGroup}(f_2) + \texttt{VehicAge}(f_3)$$

Data from "McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd edition). p.298".

Bibliography

[1] Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley.

[2] Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* New York: John Wiley.

[3] Amari, S. (1982). Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika* 69: 1–17.

[4] Amari, S. (1990). *Differential-Geometric Methods in Statistics, Lecture Notes in Statistics 28.* Berlin: Springer-Verlag.

[5] Anton, H. (1994). *Elementary Linear Algebra* (7th edition). New York: John Wiley.

[6] Bartlett, M. S. (1933-34). The vector representation of a sample. *Proceedings of the Cambridge Philosophical Society* 30: 327–340.

[7] Barndorff-Nielsen, O. E. (1987). Differential geometry and statistics: some mathematical aspects. *Indian J. Math.* 29: 335–350.

[8] Birch, M. W. (1963). Maximun likelihood in three-way contingency tables. *Journal of the Royal Statistical Society* Ser. B, 25: 220–233.

[9] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

[10] Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experinenters: An Introduction to Design, Data Analysis and Model Building.* New York: John Wiley.

[11] Box, J. F. (1978). *R. A. Fisher, The Life of a Scientist.* New York: John Wiley.

[12] Bryant, P. (1984). Geometry, statistics, probability: Variations on a common theme. *The American Statistician* 38: 38–48.

[13] Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models* (3rd edition). New York: Springer-Verlag.

[14] Corsten, L. C. A. (1958). Vectors, a tool in statistical regression thory. *Mededelingen van de Landbouwhogeschool te Wageningen, Nederland* 58: 1–92.

[15] Cox, D. R. and Hinkley, D. (1974). *Theoretical Statistics.* London: Chapman and Hall.

[16] Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980). Markov fields and loglinear interaction models for contigency tables. *The Annals of Statistics* 8: 522–539.

[17] Dempster, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Reading, Mass: Addison-Wesley Pub. Co..

[18] Deming, W. E. and Stephan F. F. (1940). On a least square adjustment of a sampled frequency table when expected marginals are known. *Annals of Mathematical Statistics* 11: 427–444.

[19] Dobson, A. (1990). *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

[20] Draper, N. R. and Smith H. (1998). *Applied Regression Analysis* (3rd edition). New York: John Wiley.

[21] Durbin, J. and Kendall, M. G. (1951). The geometry of estimation. *Biometrika* 38: 150 – 158.

[22] Edwards, D. (1995). *Introduction to Graphical Modelling.* New York: Springer-Verlag.

[23] Efron, B. (1975). Defining the curvature of a statistical problem (with discussion). *Annals of Statistics* 3: 1189–1217.

[24] Efron, B. (1978). The geometry of exponential families. *Annals of Statistics* 6: 362–376.

[25] Fienberg, S. E. (1968). The geometry of an $r \times c$ contingency table. *Annals of Mathematical Statistics* 39: 1186–1190.

[26] Fienberg, S. E. (2000). Contingency tables and log-linear models: Basic results and new developments. *Journal of the American Statistical Association* 95: 643–647.

[27] Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a 2×2 contingency table. *Journal of the American Statistical Association* 65: 694–701.

[28] Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an idefinitely large population. *Biometrika* 10: 507–521.

[29] Fisher, R. A. (1925). Applications of "Student's" distribution. *Metron* 5: 90–104.

[30] Fisher, R. A. (1935). The case of zero survivors (Appendix to Bliss, C. I. (1935)). *Ann. Appl. Biol.* 22: 164–165.

[31] Haberman, S. J. (1973). Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Annals of Statistics* 1: 617–632.

[32] Haberman, S. J. (1974). *The Analysis of Frequency Data.* Chicago: University of Chicago Press.

[33] Herr, D. G. (1980). On the history of the use of geometry in the general linear model. *The American Statistician* 34: 43–47.

[34] Jeffreys, H. (1948). *Theory of Probability* (2nd edition). Clarendon Press: Oxford.

[35] Kass, R. E. and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference.* New York: John Wiley.

[36] Kruskal, W. H. (1961). The coordinate-free approach to Gauss-Markov estimation and its application to missing and extra observations. *4th Berkeley Symposium on Mathematical Statistics* 39: 70–75.

[37] Lauritzen, S. L. (1996). *Graphical Models.* Oxford: Clarendon Press.

[38] Margolis, M. S. (1979). Perpendicular projections and elementary statistics. *The American Statistician* 33: 131–135.

[39] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd edition). London: Chapman and Hall.

[40] Norton, J., Lawrence, G. and Wood, G. R. (1998). The australian public's perception of genetically-engineered foods. *Australasian Biotechnology* 8: 172–181.

[41] Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* 37:81–89.

[42] Roy, S. N. and Kastenbaum, M. A. (1956). On the hypothesis of no interaction in a multi-way contingency table. *Annals of Mathematical Statistics* 27: 749–757.

[43] Saville, D. J. and Wood, G. R. (1986). A method for teaching statistics using N-dimensional geometry. *The American Statistician* 40: 205–214.

[44] Saville, D. J. and Wood, G. R. (1991). *Statistical Methods: The Geometric Approach.* New York: Springer-Verlag.

[45] Watson, G. S. (1967). Linear least squares regression. *Annals of Mathematical Statistics* 38: 1679–1699.

[46] Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics* 38: 1092–1109.