1999

Curvature measures for generalized linear models

Bernard A. Ellem
University of Wollongong

CURVATURE MEASURES FOR GENERALIZED LINEAR MODELS

A thesis submitted in fulfilment of the

requirements for the award of the degree

DOCTOR of PHILOSOPHY

from

UNIVERSITY of WOLLONGONG

by

BERNARD A. ELLEM, B.Sc, M.Ec *NE*

SCHOOL of **MATHEMATICS**

and

APPLIED STATISTICS

1999

Declaration

In accordance with the regulations of the University of Wollongong, I hereby state that the work described here is my original work, except where due references are made, and has not been submitted for a degree in any university or institution.

Bernard A. Ellem

For *Cherie*.

Exegi monumentum aere perennius
regalique situ pyramidum altius ...

Q. Horatii Flacci
Carminum, Liber III, Carmen XXX

Acknowledgments

This research has been made possible by the interest, involvement and generosity of others.

I would like to thank my wife Cherie for her constant and unflagging support during the extended period of study which often meant I was absent at inconvenient times. This work would have been impossible without her full support.

My supervisor, Professor David Griffiths, deserves special mention due to his patient guidance from the initial program of reading through to the final write–up where his critical review was invaluable. His support throughout the various stages of the research was very much appreciated, and his suggestions at all stages have proved to be very worthwhile.

I commend the University of Wollongong on having the vision to provide the mechanism for part–time study at the doctoral level at a time when it was not always available elsewhere.

The Staff of the School of Mathematics and Applied Statistics at the University of Wollongong are to be congratulated for their continued encouragement of my endeavours at all stages of the program.

By providing an environment for research and the means to continue my study program, Charles Sturt University through the Faculty of Science and Agriculture have aided the completion of this doctoral thesis.

I would also like to record the support given by NSW Agriculture (Biometrics Section) during the course of this investigation.

Finally, my thanks go to Karen and Peter Hiscocks for their generosity over the long stretch of time that this study occupied.

Contents

List of Figures

List of Tables

Abstract

First addressed by Beale (1960), the use of curvature measures of nonlinearity in nonlinear regression has been elucidated most comprehensively by Bates and Watts (1980). They used differential geometric results that exploit features of the Euclidean space imposed by the Normality assumption. The partitioning of these measures into intrinsic effects (due to the model) and parameter effects (due to the form or parameterization of the model) allows a proper assessment of model departures from linearity. Indeed, the term 'linear' has become synonymous with a lack of *both* of these effects, since the commonly designated 'linear model' with Normal disturbance does not contain either effect. These curvature measures are used to unravel the effects of model reformulation on convergence of fitting procedures, and on the appropriateness of confidence regions based on the linearization assumption. For model criticism using residual analysis, the presence of intrinsic curvature in a nonlinear regression model can distort the visual assessment procedures borrowed from linear modelling, since the fundamental basis of these procedures can be undermined when the model is nonlinear.

When the disturbances are non–Normal, the consequent geometry is no longer Euclidean, necessitating a different approach, as outlined by Amari (1982a). The required approach generalizes the Euclidean inner product to a metric, and the ordinary derivative to an α–connection. The concept of these α–connections is fundamental to a proper understanding of the role of differential geometry to the investigation of estimator behaviour in the case of non–Normal errors. These connections provide the general method for comparing nearby points in the parameter

space, for general classes of error distributions. In these cases, such a comparison is complicated by the difficulty of the existence of different bases for the neighbouring tangent spaces derived from the likelihood. The exception or special case is the linear model with Normal errors, where no such difficulty arises.

Casting the generalization as being from Normal to non–Normal errors, the extension can be considered to cause an 'unbundling' of the statistical properties of estimators, which in the case of Normal errors can be enjoyed simultaneously by the same estimator. In the general non–Normal case, such behaviour can no longer be guaranteed, implying that all properties may need to be considered separately, since, in the general case, specific properties of the estimator are associated with particular values of α.

This thesis outlines the fundamentals of the generalization of curvature measures to models of exponential type, in particular curved exponential families for which generalized linear models are an important subclass. This approach is used to generate insights into the properties of generalized linear models, with particular reference to the *canonical* link function as the non–Normal generalization of a linear model with Normal errors.

Indeed, the underlying 'theme' of this study is the investigation of the generalization of 'linearity' for the Normal error linear model to the non–Normal error nonlinear model. The potential simultaneity of estimator properties for the Normal distribution does not carry over to the generalization from the Normal to the non–Normal, since now each property has to be investigated separately, for each particular value of α.

As shown in Chapter 2, this individual treatment involves the statistical interpretation of each α–connection to demonstrate how key values of α are associated with estimator properties such as unbiasedness, stability of variance, lack of skewness, 'normal' likelihood and sufficiency. In terms of data analysis, all of these investigations need to be performed on the regression coefficients rather than on the fitted value (expectation parameter) scale. This requires the use of curved exponential families involving an imbedding of the regression coefficients in the

original expectation space.

One of the properties of Normal error linear models is estimator sufficiency, which for generalized linear models implies a canonical link function. The associated α–connection is the exponential or Efron connection. This connection could be considered as the springboard for the generalization of Normal error linear models to non–Normal error nonlinear models, since for generalized linear models it mimics the special case of Normal errors, by the conditions under which it vanishes. The investigation of this connection and its special relationship with generalized linear models has generated in Chapter 2 a test of adequacy for canonical link functions, based on the skewness of the regression coefficients.

The generalization of curvature follows a similar path to the α–connections, being a function of them in terms of the expectation parameters. In line with the decomposition demonstrated by Bates and Watts (1980) for Normal errors, generalized α–curvature decomposes into intrinsic and parameter–effects curvature;now, each particular α–curvature is associated with individual properties of the model, depending on the value of α. The other main change from the curvature measures of Bates and Watts is that, in the general case, a contribution to curvature is made from the error distribution as well as from the model and its parameterization. A major new result in Chapter 3 has been the proof of the invariance of intrinsic α–curvature in the general case, using a coordinate based system. A consequence of examining the generalization has been to define in Chapter 3 a class of models, *generalized nonlinear models*, having a non–Normal error distribution and a general nonlinear response function. The relationship of this class with classes of known models such as generalized linear models again raises the question of what is meant by 'nonlinearity' in general. Several related derivations such as the invariance of parameter–effects curvature in generalized linear models, and results involving exponential curvature, generalized linear models and generalized nonlinear models verify expected behaviour and highlight the generalizations that are possible.

The generalized curvature measures are shown in Chapter 4 to be related to

quantities of statistical interest such as the bias and covariance of estimators for curved exponential families, mirroring the known situation for nonlinear regression. For generalized linear models, alternative link functions to the canonical can be chosen on the basis of properties such as variance stabilization, 'normal' likelihood and lack of skewness. As expected, these links have been shown in Chapter 4 to be associated with specific α–connections. A table is presented of those link functions that produce the required properties on the expected value scale for each error distribution in a generalized linear model.

The special relationship between curvature measures, nonlinear regression and generalized linear models is further demonstrated in Chapter 5 by the use of a new method for nonlinear regression based on a second order approximant to the non-linear function by means of a special generalized linear model. As expected, such an approximation follows the true function more closely than linearization;this is demonstrated empirically from calculations of leverage, parameter estimates and corresponding interval estimation. All these effects are predicted from considerations based on curvature measures, both intrinsic and parameter–effects.

The effect of replication on curvature is known empirically and theoretically in the case of nonlinear regression. In Chapter 5 it is shown that replication has two implications for the effects of curvature in a generalized nonlinear model. Firstly, the central limit theorem produces convergence to the Normal distribution, so that the error contribution to general α–curvature becomes zero asymptotically. The effect of replication on the model contribution is less clear, since the general limiting case is nonlinear regression if only the error component of α–curvature is considered. Locally, the generalized nonlinear model will be well approximated by a linear model. Secondly, under some conditions, a generalized nonlinear model will converge locally to a generalized linear model with canonical link. However, when the error component and the model component are considered, the overall effect of intense replication will be to produce locally a linear model with Normal errors.

Chapter 1

Introduction

1.1 Background

Curvature measures were proposed by Beale (1960) to assess the departure of a nonlinear regression model from its assumed linear approximation in the neighbourhood of the least squares estimate. The motivation of this analysis was the evaluation of the validity of linearization–based confidence regions for model parameters in the nonlinear regression model. This method of curvature measurement was formalised by the differential geometric approach of Bates and Watts (1980) which refined the measures in such a way that two different types of effects were clearly identified;

1. intrinsic curvature, ie., curvature peculiar to the model and which is unchanged by the particular parameterization of the model, and

2. parameter–effects curvature, ie., curvature that is dependent on the form of parameterization of the nonlinear model.

The impact of these two measures on proper construction of confidence regions for nonlinear regression models has been extensively researched (Bates and Watts, 1988). In particular, the use of likelihood–based confidence regions with those based on the linear approximation has been employed in the construction of

practical measures to assess the effect of nonlinearity on these confidence regions. These measures(t plots, traces and pair sketches) are based on *profile* likelihood where all parameters except those being considered are estimated.

Important properties, such as bias and correlation, of estimators in nonlinear regression have been shown to be related to these curvature measures. Thus, transformations which reduce bias and the absolute value of correlation can be found, in agreement with general results from earlier workers such as Box (1971), Bartlett (1953b) and Clarke (1980).

1.2 Rationale for the study

Several discussants to the paper of Bates and Watts (1980) raised from differing viewpoints the question of a non–Normal error distribution. Ross (1980a) was interested in the question of parameter transformation for general non–Normal errors while Reid (1980) was concerned about the general exponential family of models in the context of the measure of statistical curvature defined by Efron (1975). McCullagh (1980) queried the extension of the measures to estimation for error distributions from the exponential family, in particular generalized linear models. As pointed out by the authors, the assumption of a Normal disturbance was crucial to the approach, since this assumption implied a Euclidean metric, enabling the results of classical differential Euclidean geometry to be exploited. The generalization to non–Normal distributions requires a Riemannian metric, and the concept of an *affine connection.* As observed by Kass (1984), the extension of the approach of Bates and Watts to generalized linear models (GLMs) requires the use of a *family* of effects related to the one parameter α–connections of Amari (1982a), with key values of the parameter α being associated with special features of the estimator. A principal function of this thesis is to investigate the suggested generalization to GLMs with the purpose of using generalized curvature measures to examine the statistical behaviour of estimators in GLMs and associated models, especially those related to the exponential family.

A brief overview of Amari's α–connections is given in Seber and Wild (1989, pp159–165), with derivations that show the relationship between Amari's theory and the definition of statistical curvature given by Efron (1975), as well as results due to Kass (1984). The main thrust of their presentation is to demonstrate that the general theory of Amari reduces to the curvature measures of Bates and Watts for the case of nonlinear regression.

1.3 Role of Curvature Measures in Nonlinear Regression

Even before the advent of the digital computer, the underlying nature of some regression problems had forced researchers to fit nonlinear models to data. Early attempts at solving this problem by using an iterative procedure based on linearization of the nonlinear function sometimes ran into difficulties of non–convergence and failed initializations. Various techniques can be used to find reasonable starting values (Draper and Smith, 1981). These include grid search, exact solution using minimal data and the transformation method in some cases. Some of these methods are crude fitting procedures (Sadler, 1975) which are capable of refinement. Modifications of the iterative method for fitting include damping of the Gauss–Newton method as per Hartley (1961), procedures based on the Newton–Raphson method and the Levenberg–Marquardt compromise (Sadler, 1975). All these methods can be interpreted in parameter space via the true residual sum of squares surface. The working of these adaptations of search procedures can be demonstrated in parameter space by viewing the behaviour of the approximate residual sum of squares surface (Sadler, 1975), since these adaptations employ a quadratic approximation to the true residual sum of squares surface, a linear approximation to the nonlinear function or a procedure employing both approximations. An alternative method of interpretation is to use the *sample space* (Draper and Smith, 1981), which facilitates the presentation of the linearization and hence

demonstrates the adequacy (or inadequacy) of the linear approximation to the nonlinear function, thus showing the reasons for algorithm failure. This approach formed the basis of the expectation surface geometry used by Bates and Watts (1980) to analyse in detail the underlying structure of nonlinear regression models. The consequent decomposition of nonlinearity into intrinsic and parameter–effects curvature enabled a new insight into the interpretation of nonlinear regression models. Low intrinsic curvature, or minimal bending of the solution locus, is a precondition for 'close to linear' behaviour of a nonlinear regression model (Ratkowsky, 1983), since a nonlinear regression model that also effectively exhibits a regular grid of parameter values along the solution locus mimics the behaviour of a linear model. This means that nonlinear regression models of this class can be treated as linear for the purposes of parameter interpretation via the desired inferential procedure of hypothesis testing or interval estimation (Bates and Watts, 1988). Empirical evidence (Bates and Watts, 1980) suggests that parameter–effects curvature tends to be the dominant effect in many cases, allowing for the possibility of inducing 'close to linear' behaviour via suitable reparameterization for a low intrinsic curvature model (Bates and Watts, 1981). In addition, a reduction in parameter bias can be effected by reparameterization (Bates and Watts, 1980). In practice, the procedures for discovering parameterizations with low parameter–effects curvature are most often empirical, such as those due to Ross (1980b), as further detailed in Bates and Watts (1981). The choice is between global or automatic methods that annihilate parameter–effects curvature and transformations to expected values [or 'stable ordinates', Ross(1980b)] that can induce low parameter–effects curvature. The disadvantages of the automatic method are twofold. The new parameters given by the automated procedure may not be easy to interpret in terms of the original problem, and they may not exist. The corresponding advantages of the expected value method are that the new parameters can be interpreted directly in terms of the original response and that they can require less computing subject to mild constraints. These constraints can imply equally spaced predictors for some models. Of secondary importance is the impact of parameter–effects curvature on

convergence of fitting algorithms. Lower parameter–effects curvature can mean fewer iterations for convergence to the optimum. Given the current developments in computing hardware and statistical software, this is less of a problem than inference for parameters. While a low parameter–effects parameterization offers advantages for inference statements on model parameters, interpretability of parameters and their relevance in the original context of the problem are probably of more concern to the user. Backtransformation to the original parameters is possible however, as is approximation to the covariance structure of the parameters on the original scale. Thus, model fitting and interpretation can quite satisfactorily use separate parameterizations. The effects of curvatures on residuals from nonlinear regression models bears some explanation. Parameter–effects curvature is not relevant to discussions about residual analysis, since the fitted value is independent of the form of parameterization and is a function solely of the model itself. Only intrinsic curvature affects the residual procedures borrowed from linear modelling. For a linear model, the residuals and fitted values are approximately uncorrelated, and the residuals are centred on zero. For a nonlinear model with non–trivial intrinsic curvature, the situation changes. The expected values of the residuals are no longer zero, and the residuals and fitted values are no longer uncorrelated, resulting in a negative expected slope in the plot of residuals against fitted values. This problem of using procedures designed for linear models on nonlinear models carries over even to modified residuals that have been proposed for linear models, although the projected residuals of Cook and Tsai (1985) appear free from such problems. For an overall discussion see Seber and Wild (1989, pp178-179).

The above discussion on residuals can be formalised using the concept of *leverage* or *potential* (Weisberg, 1985, p111). For a linear model, the *hat* or *projection* matrix $\boldsymbol{H}$ is defined by

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}$$

so that

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

The term *leverage* (or *potential*) is applied to the diagonal terms, since as $h_{ii} \to 1$, then $\widehat{y}_i \to y_i$. In short, the effect of the ith case will be large if h_{ii} is large, subject to the nature of the y_i; hence the terminology.

In the nonlinear case, the projection matrix can be similarly defined; however the operator $\boldsymbol{H}$ is now a function of the derivative of the nonlinear function with respect to the parameters, as described in Seber and Wild (1989, 4.41, p140, 4.145, p174 and 4.146, p175). This explains the descriptive results given in the above discussion involving intrinsic curvature, since intrinsic curvature is defined in terms of such derivatives.

The concept of *influence* involves the effect of a particular data point on model behaviour, and is typically assessed by systematic deletion of key data points and assessment of the corresponding regression diagnostics. Data points (cases) whose deletion cause major changes in the resulting diagnostics are called *influential.* Examples of such measures abound, eg., Cook's distance measures (Cook, 1977). If a case is deleted then the model is changed, and so the intrinsic curvature changes as well as parameter effects. For this reason, influence will not be given the attention of other model diagnostics such as leverage.

1.4 Curvature Measures

An empirical overview of the interpretation of the curvature measures of Bates and Watts (1980) follows. The model describing the response Y given predictors $\boldsymbol{X}$ is assumed to be

$$Y = f(\boldsymbol{X}; \boldsymbol{\theta}) + \varepsilon, \;\; \varepsilon \sim N(0, \sigma^2).$$

The solution locus in the space of expectations is defined by the set of parametric equations

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = f(\boldsymbol{X}; \boldsymbol{\theta}).$$

Nonlinearity is exhibited in this space of expectations $E(Y_i|\boldsymbol{\theta}), \;\; i = 1, \ldots n$, by

- curvature of the solution locus (intrinsic curvature), and

- non–uniformity of the coordinate system along the solution locus (parameter–effects curvature).

Operationally, these effects manifest themselves respectively in

- how well the tangent plane approximates the solution locus locally, and
- how well a uniform coordinate system on the tangent plane approximates the coordinate system on the solution locus.

(The solution locus is also called the expectation surface.)

A simple one–parameter example from Ratkowsky (1983) will be used to describe these two nonlinearity effects. The data for this problem are shown in Table 1.1.

Example 1

Observation Number	X	Y
1	2.0	2.5
2	3.0	10.0

Table 1.1: Illustrative Data Set

Two competing models are shown, one linear the other non–linear. The problem is designed so that the non–linear model better fits the data, as shown in Figure 1.1, where the curve is 'closer' to the data than the line. In expectation space, the *single* point that represents the data is closer to the solution locus for the non–linear (Figure 1.3) than for the linear model (Figure 1.2). The corresponding residual sums of squares (deviances) are 2.93 and 12.02, respectively.

In order to properly gauge these two effects in nonlinear models, their behaviour for linear models will be described first.

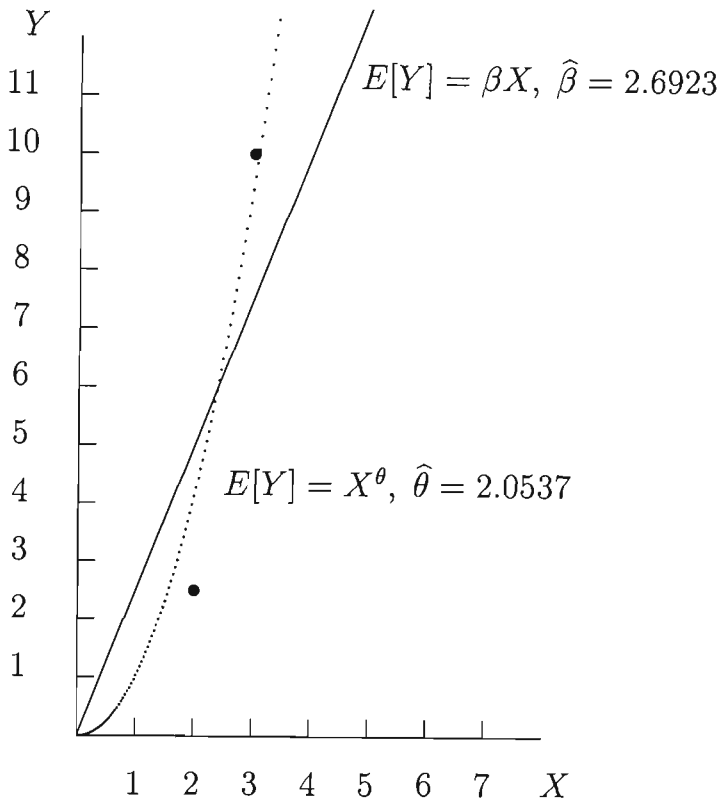

Figure 1.1: Ratkowsky Problem.

1.4.1 Linear Model

For linear models, both effects disappear, since

- $\boldsymbol{\mathcal{F}}(\theta) = \boldsymbol{X}\theta$, which is a line in the space of expectations, and
- the solution locus and the tangent plane coincide, so that equal increments of θ along the solution locus correspond to a uniform spacing thereon. The parametric equation describing the solution locus is equivalent to the parametric equation describing the tangent.

The linear model $E(Y) = X\beta$ can be demonstrated on the data from Example 1 (Figure 1.1) with θ replacing β. The solution locus for the linear model is shown

in Figure 1.2

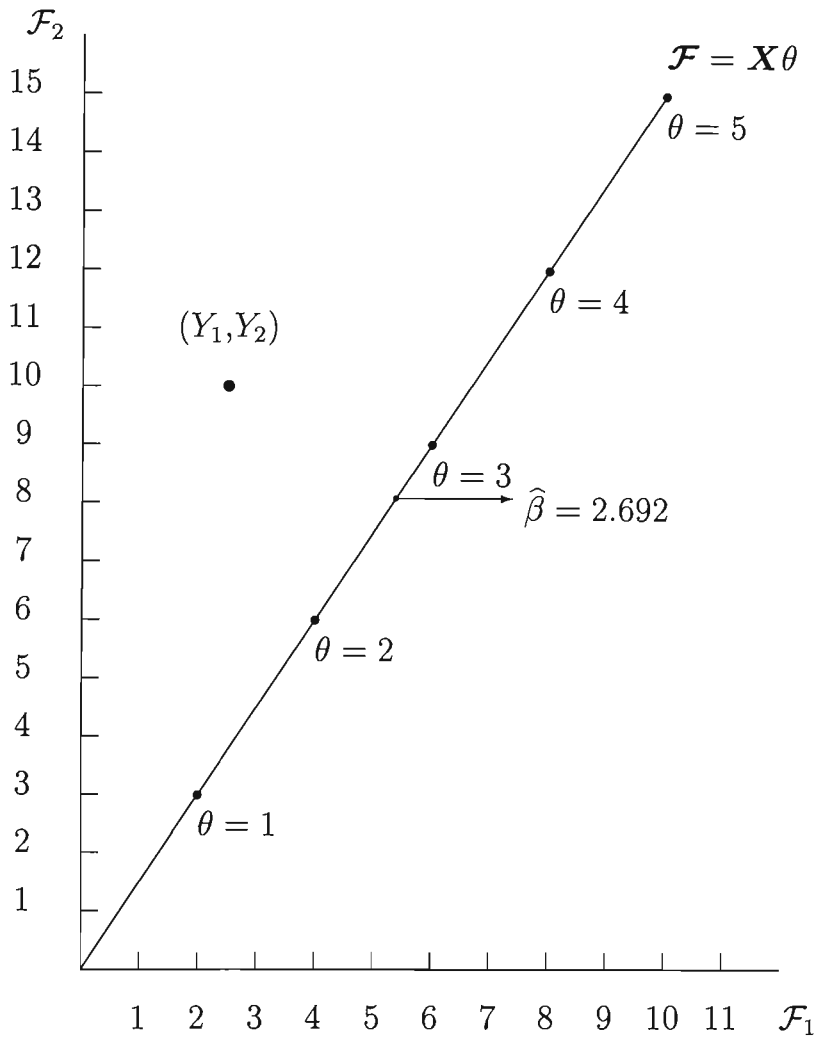

Figure 1.2: Solution Locus : Linear Model.

The parametric equation of the solution locus is :

$$(\mathcal{F}_1, \mathcal{F}_2) = (X_1\theta, X_2\theta).$$

For this model,

$$E(Y) = X\theta = X\beta$$

and the intrinsic curvature on the solution locus is zero, as shown in Figure 1.2, where there is no 'bending' of the solution locus. The spacings along the solu-

tion locus between points corresponding to equal spaced values of the parameter are identical, as shown in Figure 1.2. This indicates a constant velocity or rate of increase with respect to θ along the solution locus, ie., no acceleration or parameter–effects curvature.

1.4.2 Non–linear Model

For the nonlinear model

$$Y = f(X;\theta) + \varepsilon$$

the expectation relation is

$$\mu = f(X;\theta)$$

and so a curve will be generated[1].

The vector formed by the data and least squares solution will be perpendicular to the tangent on the curve, as shown in Figure 1.3. The following observations can be made :

- The solution locus is now a curve, in contrast to the line for the linear model.
- In both cases the vector formed by the data and the least squares solution point is perpendicular to the tangent to the expectation surface.
- Equal increments in θ correspond to equal step lengths along the expectation surface for the linear model $E(Y) = X\theta$, but not for the nonlinear model $E(Y) = X^{\theta}$.

Two parameterizations of the nonlinear model are shown in Figure 1.3.

For the chosen non–linear model, intrinsic curvature (as determined by the radius of curvature) appears slight,[2] but is precisely the same for both parameterizations of the model, ie., X^{θ} and $X^{\ln\phi}$. That is, the 'bending' of the solution locus is the same for these two forms of the same model. In contrast, parameter–effects

[1]Note that *expectation surface = solution locus.*

[2]Actually, the *inverse* of the radius of curvature. In the linear case, $r = \infty \Rightarrow$ intrinsic curvature = 0.

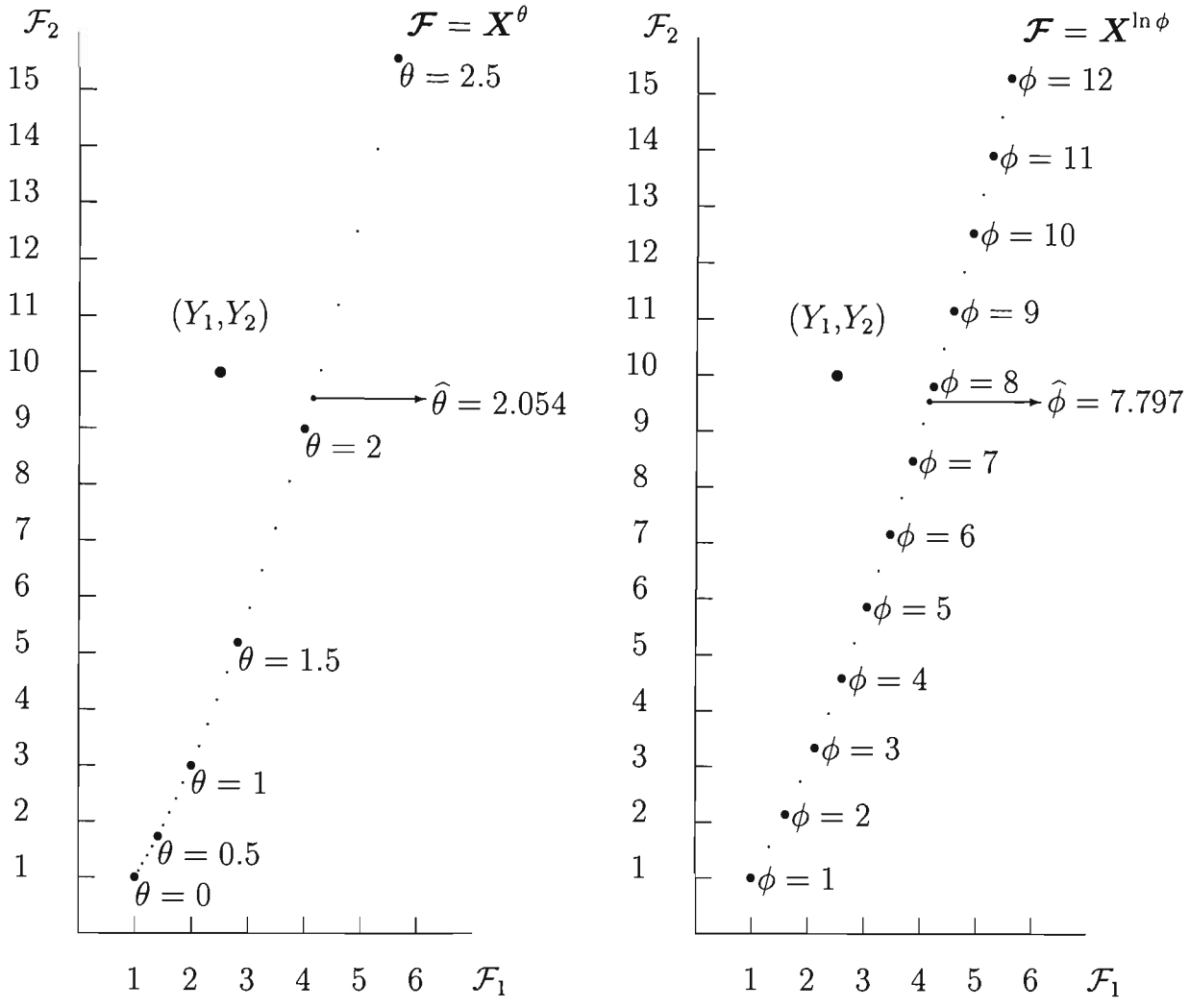

Figure 1.3: Solution Locus : Non–linear Model.

curvature (acceleration or change in speed along the solution locus), is large for X^{θ} compared to that for $X^{\ln \phi}$. Empirical evidence can be given to corroborate the visual preference for the second form of the model. Using the *linearization* procedure (Draper and Smith, 1981) from the same starting value for both forms, ie., $\theta = 0 (\phi = 1)$, convergence to the least squares minimum took 9 iterations for the first form(θ) but 3 iterations for the second(ϕ).

The utility of model reparameterization is contingent on the intrinsic curvature being slight, but fortunately empirical evidence suggests that this is often the state

of affairs, as shown by Bates and Watts (1980).

These non–linear and linear models demonstrate the following :

(a) Intrinsic curvature can only be changed by changing the model[3], as in this example from X^{θ} to $X\theta$.

(b) Parameter–effects curvature is conditional on the choice of model, eg., from a linear (zero) to a nonlinear model (non–zero). The main determinant for the size of parameter–effects curvature is often the choice of parameterization. A judicious choice of parameterization, here, $X^{\ln \phi}$, instead of X^{θ}, produces a locally uniform coordinate system on the solution locus. In general, such a judicious choice can render the assumption of a uniform coordinate system at least approximately true. If a parameterization can be determined with a low parameter–effects curvature, this means any confidence regions based on linearization will be close to likelihood based confidence regions. In short, the model under that parameterization will behave as a linear model if intrinsic curvature is slight.

The usefulness of this analysis is to separate 'departure from linearity' into :

- specific model dependent effects (intrinsic curvature), and
- model representation effects (parameter–effects curvature).

The latter can be manipulated by the particular formulation that a practitioner chooses for a specific model, whereas the former is fixed by the choice of model and design points. The results of such manipulation may not always be as predicted or expected.

1.4.3 Higher Dimensions

In the spirit of the expository example (Example 1, Figure 1.1) from Ratkowsky (1983), two additional examples are described to demonstrate the interpretation of the expectation surface in higher dimensions.

[3]The model is changed even if the X variable is changed, eg., to $\ln X$.

Example 2

The problem shown in Table 1.2 is taken from Draper and Smith (1981, page 517), and involves fitting the model $Y = e^{-\theta t} + \varepsilon$, to the given data.

t	Y
1	0.8
4	0.45
16	0.04

Table 1.2: Problem A : Draper and Smith (1981)

The expectation surface (solution locus) can be drawn without knowing the observed values for response variable (Y), but the predictor (X), ie., the design points, are needed. The following parametric equations define the expectation surface

$$t = 1 : F_1 = e^{-\theta}$$
$$t = 4 : F_2 = e^{-4\theta}$$
$$t = 16 : F_3 = e^{-16\theta}$$

so points on the solution locus are given by $(e^{-\theta}, e^{-4\theta}, e^{-16\theta})$. This space curve is shown in Figure 1.4, with the box symbol(□) indicating the data in expectation space.

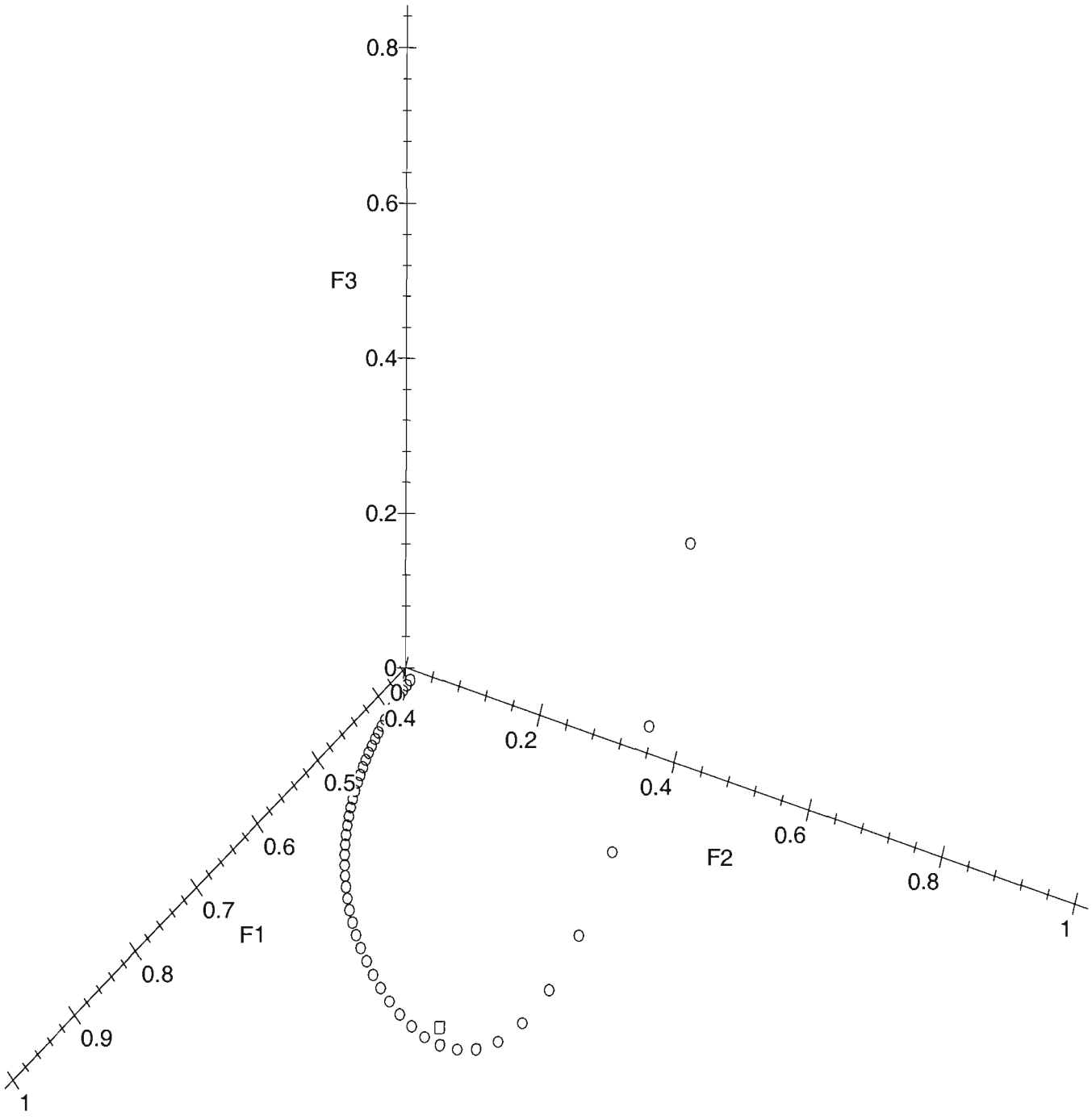

Figure 1.4: Solution Locus : View 1, Example 2.

Figure 1.4 shows the intrinsic curvature globally and locally. Overall (globally) the space curve bends markedly, whereas locally such curvature can be slight, eg, for low expected values. From Figure 1.5 it can be seen that the spacing of increments of θ along the solution locus is reasonably regular for low expected values, but not so for higher values. This shows that parameter–effects curvature can also vary from the local to the global. In practice, numerical measures are required to make a proper judgement of these curvatures. Ratkowksky (1983) gives computer code for such numerical measures.

Example 3

The model

$$E(Y|x) = \alpha x^{\beta}$$

is to be fitted to the data shown in Table 1.3 assuming Normal disturbances.

x	Y
1	1.0
2	2.8
3	5.2

Table 1.3: Two Parameter Example.

The following parametric equations define the expectation surface :

$$x = 1 : F_1 = \alpha$$
$$x = 2 : F_2 = \alpha 2^{\beta}$$
$$x = 3 : F_3 = \alpha 3^{\beta}$$

The solution locus is thus

$$(\alpha, \alpha 2^{\beta}, \alpha 3^{\beta}).$$

This surface is shown in Figure 1.6, with the data shown by the box symbol (□).

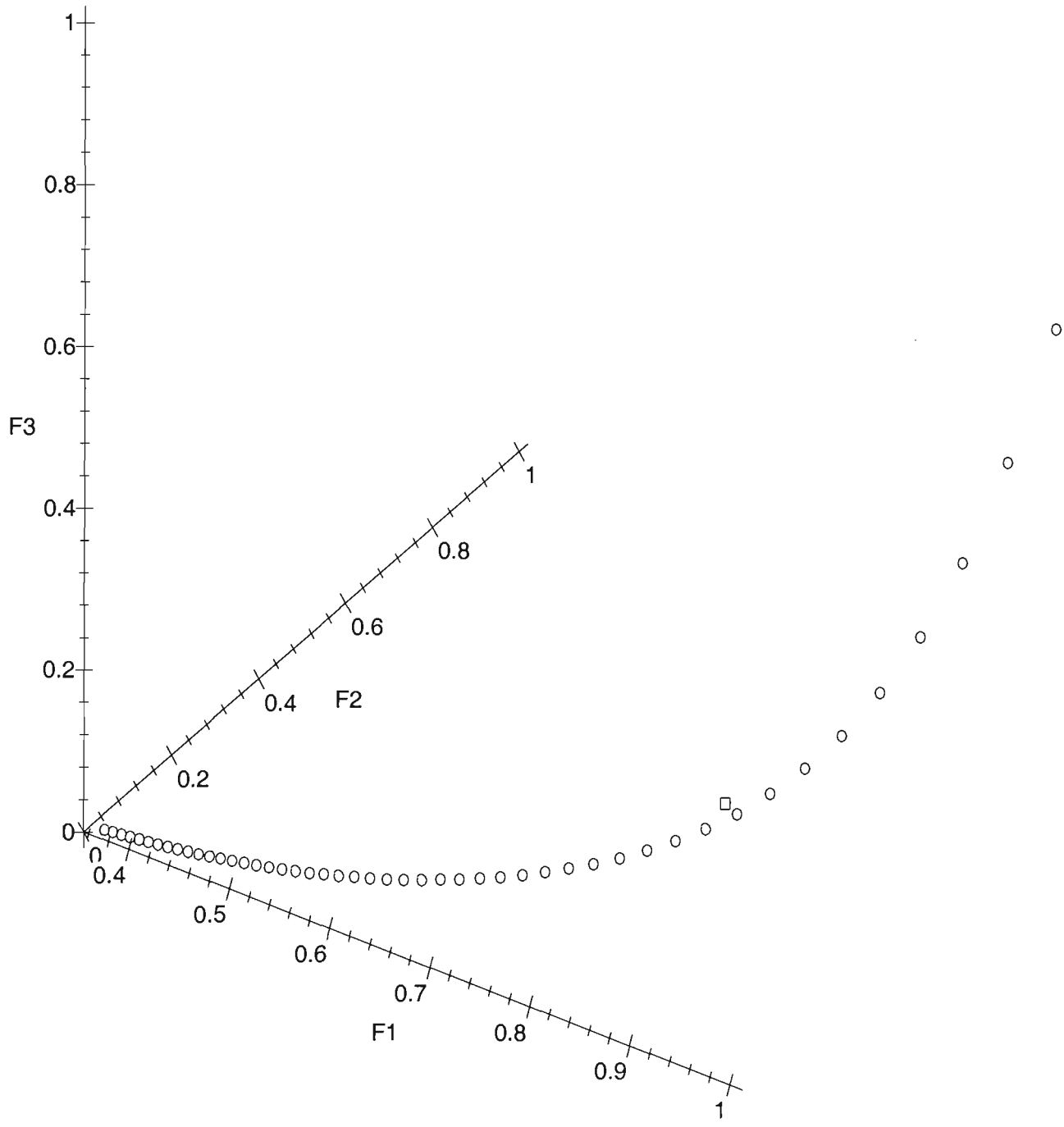

Figure 1.5: Solution Locus : View 2, Example 2.

Interpretation of the expectation surface can be made by considering level curves, ie, the paths traced out by holding α constant and β constant.

For β constant $(= c)$, the solution locus becomes

$$(\alpha, \alpha 2^c, \alpha 3^c)$$

which describe lines passing through the origin.

For α constant $(= d)$, the solution locus is defined by

$$(d, d2^\beta, d3^\beta)$$

constituting 'parallel' curves 'orthogonal' to the lines defined by holding β constant.

The two systems are shown in Figure 1.6, with the lines and curves clearly visible. The data are shown by the box symbol (□). This is effectively a one–parameter problem, with β being the nonlinear parameter. The nature of the expectation surface confirms that α is really a linear parameter with β being nonlinear. The sums of squares surface show the nonlinear parameter as being the most responsive, as is shown by the results of fitting the model via the statistical package GLIM (NAG, 1985). A GLM (generalized linear model) formulation of the model

$$E(Y) = \mu = \alpha X^\beta = e^{\ln \alpha} e^{\beta \ln X}$$

gives

$$\widehat{\ln \alpha} = -0.0142 \ (0.014135) \rightsquigarrow \widehat{\alpha} = 0.9859.$$

and

$$\widehat{\beta} = 1.513 \ (0.01425).$$

The values for α and β used to generate the data were respectively 1.0 and 1.5 and the quantities given in brackets are the estimated standard errors.

1.4.4 Practical Considerations

For ease of estimation and subsequent inference, both intrinsic and parameter–effects curvature need to be controlled, but experience (Bates and Watts, 1980,

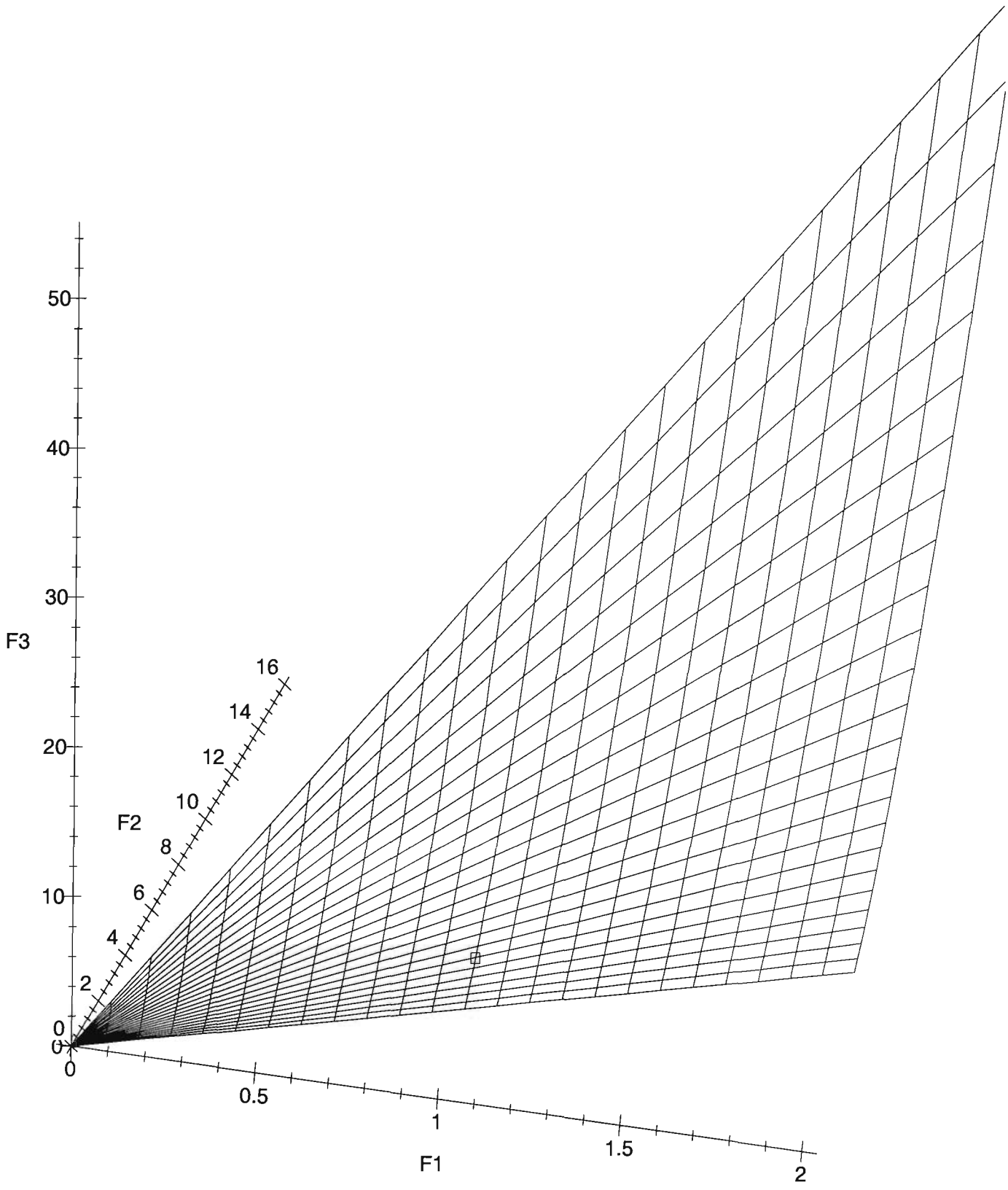

Figure 1.6: Solution Locus : Example 3

1988), (Ratkowsky, 1983), (Lowry and Morton, 1983) indicates that the latter problem appears to dominate[4], implying that a proper parameterization may alleviate difficulties. For estimation, choosing a parameterization that reduces parameter–effects curvature has the effect of reducing bias (Bates and Watts, 1980). It also has the potential to reduce the amount of calculation performed in the fitting process, by substantially reducing the number of iterations, as shown in Section 1.4.2. This is of less concern now than in 1980 due to the advances in computing hardware and software platforms, which have resulted in great increases in speed of computations. For inference, the operational consideration is that a parameterization with low parameter–effects will ensure the validity of the uniform coordinate assumption. This is needed for the use of the linear approximation method of constructing confidence regions. Profile likelihood methods (Bates and Watts, 1988) require only the planar assumption, and thus are operationally valid for low intrinsic curvature. Thus, the use of such profile methods would have appeal, since the user is then able to use the preferred (interpretable) parameterization, subject to the proviso of low intrinsic curvature. Such information on curvature is imbedded in the profiling methods (Bates and Watts, 1988).

The curvature measures used in practice represent a compromise between a maximum attained value and some averaged effect such as root mean square (RMS) curvature. All such curvatures are scaled, avoiding dependencies on the data and parameters. The averaged curvatures appear more attractive since the maximum measures tend to be pessimistic, being typically of the order of twice the magnitude of the average curvatures (Seber and Wild, 1989, p159). An additional reason for using RMS curvatures is that then magnitude can be gauged by comparison with the the desired critical point of the F distribution (Bates and Watts, 1988). However, parameter–effects curvature is a valuable tool for measuring the effects of reparameterization on bias and the adequacy of the linear approximation assumed by the inference procedures. Bates and Watts (1981) used curvature measures to determine appropriate transformations (or reparameterizations).

[4]This may indicate that only models with low intrinsic curvature have been chosen.

Since intrinsic curvature changes with the design as well as with the model, optimisation of the placement of design points as well as choice of the model can be considered, as in Bates and Watts (1981).

The rationale for curvature measures claimed by Bates and Watts (1980) is :

- 'the geometric approach to measuring nonlinearity is ... relatively simple and straightforward ...', and
- 'The concepts and methods of differential geometry ... make the study of nonlinearity as geometrically accessible and understandable as linear least squares.'[5]

A focus of this thesis is to extend this claim to the class of functions defined by generalized linear models, and to other models derived from the exponential family, along the lines of the suggestions of Kass (1984). This will be attempted by generalizing the parameter–effects curvature and intrinsic curvature for each geometry that corresponds to a particular estimator attribute associated with a value of α. As shown by Kass (1984, pp90–91), it is necessary to study the reduction of parameter–effects curvature not only for the exponential geometry but also for all the other geometries (values of α) that correspond to the other key properties of estimator behaviour. This expansion is required because all of the properties can be satisfied simultaneously by a single transformation for the Normal distribution, since all the α–connections coincide. For other error distributions, the α–connections are distinct and so the estimator properties cannot be satisfied by a single transformation. Thus each key value of α has to be considered in turn.

1.5 Generalized Linear Models

Consider a random sample $Y_1, \ldots, Y_n$ from a population with pdf $f(y;\theta)$. The generalized linear models (GLMs) defined by McCullagh and Nelder (1989), constitute

[5]Bates and Watts (1980, p14–p15).

a class defined by

$$E(Y_i) = \mu_i$$

for which the ith contribution to the log–likelihood can be written as

$$\ell_i \stackrel{\text{def}}{=} \ln f(y_i;\theta_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi).$$

This class of models was first expounded by Nelder and Wedderburn (1972), and is related to the class of curved exponential families defined by Efron (1975).

This thesis investigates the usefulness of differential geometric measures to models having a general error distribution. The family of models studied are of exponential type, with particular emphasis on generalised linear models, following the suggestions of Kass (1984). The rationale for such curvature measures stems from the work of Bates and Watts (1980), as applied to the case of nonlinear regression. The work of Efron (1975) on statistical curvature and the geometry of exponential families (Efron, 1978) also provides background for the thrust of this work, albeit in the one dimensional case.

The class of generalised linear models can be considered to be a subset of the class of models for which

$$E(Y_i) = \mu_i(\boldsymbol{\theta}).$$

So, some of the results obtained for GLMs may extend to this more general class of models. Certainly, it can be shown that the IRLS[6] algorithm (Green, 1984) for the more general problem, is equivalent to the GLIM algorithm for generalised linear models. From the user's point of view, however, the relaxing of the requirement for starting values is a non–trivial difference between the two approaches.

1.5.1 Leverage

The concept of leverage has been introduced for linear models in Section 1.3. For GLMs, the hat matrix and its corresponding leverage terms generalize because the

[6]IRLS = Iteratively Reweighted Least Squares.

GLIM algorithm uses a weighted regression on the working variate z, defined by

$$z = \eta + (Y - \mu)\left(\frac{d\eta}{d\mu}\right).$$

Since there is a 1–to–1 mapping between the linear predictor η and the fitted value μ, these quantities can be converted back to the original data scale. In the general form of the $\boldsymbol{H}$ matrix, $\boldsymbol{X}$ is replaced by $\boldsymbol{W}^{1/2}\boldsymbol{X}$, where $\boldsymbol{W}$ is the matrix of weights. The resulting hat matrix is

$$\boldsymbol{H_g} = \boldsymbol{W}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{W}^{1/2}$$

and the weights are defined by

$$W = V^{-1}\left(\frac{d\mu}{d\eta}\right)^2.$$

The function V is the GLM variance function in terms of the mean μ;

$$V = V(\mu), \text{ where } V = b''(\theta), \text{ and } \mu = b'(\theta).$$

The model is no longer purely linear, since the $\boldsymbol{H_g}$ operator is now a function of the derivatives of the fitted values (via W), as in the case of nonlinear regression.

Using the notation of section 5.1.1, in the nonlinear regression model

$$\boldsymbol{Y} = f(\boldsymbol{X};\boldsymbol{\theta}) + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(0,\sigma^2)$$

the nonlinear function $f(\boldsymbol{X};\boldsymbol{\theta})$ is replaced by

$$f_0 + \frac{\partial f}{\partial \boldsymbol{\theta}_0}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

to give the linear model

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

defined by

$$\boldsymbol{y} = \boldsymbol{Y} - f_0,$$

$$\boldsymbol{Z} = \frac{\partial f}{\partial \boldsymbol{\theta}_0}$$

and

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0.$$

The hat matrix for this linearized model is now

$$\boldsymbol{H}_l = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$$

which gives

$$\widehat{\boldsymbol{y}} = \boldsymbol{H}_l \boldsymbol{y}.$$

From the definition of $\boldsymbol{Z}$, it can be seen that the matrix $\boldsymbol{H}_l$ is a function of the derivative of the fitted values, as for the GLM. See McCullagh and Nelder, (1983, Ed. 1, p210) and (1989, Ed. 2, p397 and p405).

In terms of the data $\boldsymbol{Y}$ and the fitted values $\widehat{\boldsymbol{\mu}}$, the following holds[7] for a GLM,

$$\boldsymbol{V}^{-1/2}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) \approx \boldsymbol{H}_g \boldsymbol{V}^{-1/2}\left(\boldsymbol{Y} - \boldsymbol{\mu}\right),$$

where $\boldsymbol{V} = \text{diag}\left(V(\mu_i)\right)$, so that $\boldsymbol{H}_g$ measures the impact in standardized units of changes in the data on the fitted values. Defining $\boldsymbol{Y}_s$ and $\widehat{\boldsymbol{Y}_s}$ by

$$\boldsymbol{Y}_s = \boldsymbol{V}^{-1/2}\left(\boldsymbol{Y} - \boldsymbol{\mu}\right)$$

and

$$\widehat{\boldsymbol{Y}_s} = \boldsymbol{V}^{-1/2}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) = \boldsymbol{V}^{-1/2}\left(\widehat{\boldsymbol{Y}} - \boldsymbol{\mu}\right)$$

gives

$$\widehat{\boldsymbol{Y}_s} = \boldsymbol{H}_g \boldsymbol{Y}_s. \tag{1.1}$$

In *raw* terms

$$\widehat{\boldsymbol{Y}} = \boldsymbol{V}^{1/2}\boldsymbol{H}_g\boldsymbol{V}^{-1/2}\boldsymbol{Y} \stackrel{\text{def}}{=} \boldsymbol{\mathcal{H}}\boldsymbol{Y} \tag{1.2}$$

where $\boldsymbol{\mathcal{H}}$ is asymmetric, in general. Both Equation (1.1) and Equation (1.2) can be derived from the leverage equation expressed in terms of the weighted working variate, ie,

$$\boldsymbol{W}^{1/2}\widehat{\boldsymbol{z}} = \boldsymbol{H}_g \boldsymbol{W}^{1/2}\boldsymbol{z}$$

[7] McCullagh and Nelder (1989, 2nd Ed., p397).

as demonstrated in Appendix A.1.

For the case of a *linear* model with Normal errors,

$\eta = \mu$ and $\boldsymbol{V}$ = Diag(constant) = constant $\times \boldsymbol{I}$, which means that $\boldsymbol{W}$ = Diag(constant). Thus $\boldsymbol{H_g}$ reduces to

$$\boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top} = \boldsymbol{H}$$

as defined for linear models with Normal errors in Section 1.3. Since $\boldsymbol{V}$ = Diag(constant), then the standardized form [Equation (1.1)]

$$\boldsymbol{V}^{-1/2}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) = \boldsymbol{H}\boldsymbol{V}^{-1/2}\left(\boldsymbol{Y} - \boldsymbol{\mu}\right)$$

gives

$$\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} = \boldsymbol{H}\left(\boldsymbol{Y} - \boldsymbol{\mu}\right).$$

Since it can be shown that $\boldsymbol{\mu} = \boldsymbol{H}\boldsymbol{\mu}$, then

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

In raw terms, Equation (1.2) becomes

$$\widehat{\boldsymbol{Y}} = \boldsymbol{V}^{1/2}\boldsymbol{H}\boldsymbol{V}^{-1/2}\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

since $\boldsymbol{V}$ = Diag(constant).

Thus both the standardized [Equation (1.1)] and raw form [Equation (1.2)] of the leverage for a GLM reduce to the same leverage form for linear models with Normal errors.

1.6 Exponential Families

The class of GLMs is a special class of curved exponential families, having natural parameters that can be related to linear functions of the parameters of interest ($\boldsymbol{\beta}$), via

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} = h(\boldsymbol{X}\boldsymbol{\beta}),$$

using GLIM notation. In order to develop results for GLMs it is necessary to establish theory using the general formulation of an exponential family, viz,

$$\ln f(y_i;\theta_i) = y_i\theta_i - \Psi_i(\theta_i) + c_i(y_i)$$

following notation due to Amari (1982a). For the special case of independently distributed $Y_1,\ldots,Y_n$ we have

$$\ln f(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{i=1}^{n} y_i\theta_i - \Psi(\boldsymbol{\theta}) + c(\boldsymbol{y}).$$

If the Einstein tensorial summation convention is used, then

$$y_i\theta_i \stackrel{\text{def}}{=} \sum_{i=1}^{n} y_i\theta_i = y_1\theta_1 + \ldots + y_n\theta_n.$$

Most of the theoretical results in this thesis are derived in terms of such a general exponential family, and the ultimate goal is to not only produce and condense the approach in terms of the GLM form, but also to determine any simplifications or special cases that apply for this special subclass of the exponential family.

1.7 Curved Exponential Families

For the general exponential family defined by

$$\ell(\boldsymbol{y};\boldsymbol{\theta}) = \theta_i y_i - \psi(\boldsymbol{\theta}) + c(\boldsymbol{y})$$

the natural parameter space defined by $\boldsymbol{\theta}$ is the generalization of the Cartesian coordinate system from Euclidean space.

From a practitioner's point of view, interest is usually in the subset of parameters (regression coefficients) which generate the space of expectations (fitted values) via a parsimonious model. This subset of parameters is generally related to the natural parameters by a nonlinear function, hence the phrase 'curved subsets of a larger parameter exponential family' or 'curved exponential family' (Efron, 1975, page 1192).

Several examples are given to indicate the subsets of parameters that can arise in practice (the dimension of the natural space is given by k, while the curved space dimension is given by p).

The imbedding of the regression coefficient(s) ($\boldsymbol{\beta}$) in the space of natural parameters ($\boldsymbol{\theta}$) is given by a (nonlinear) relation

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\beta}).$$

This relation is shown in the examples.

1. Autoregressive Model – AR(1) : $k = 2, p = 1$.

2. Poisson Regression Model : $k = n, p = 1$.

3. Nonlinear Regression model : $k = n, p = m$.

4. Generalized Linear Model : $k = n, p = m$.

In particular, note that the AR(1) model is of exponential type. Most of the later results of this thesis will apply to such exponential type models, of which GLMs are an important subset.

Example 1 : Autoregressive Model

The stationary AR(1) model, following Efron (1975, p1194)

$$X_t = \phi X_{t-1} + \varepsilon_t, \quad t = 1 \ldots T$$

is proposed for the time series $X_0 \ldots X_T$, with $X_0 = \varepsilon_0$.

It is assumed that $\varepsilon \sim N(0, 1)$ and that $-1 < \phi < 1$.

Using

$$X_t - \phi X_{t-1} \sim N(0, 1),$$

the likelihood when written in exponential form yields :

$$\theta_1 = -\frac{1 + \phi^2}{2}, \; \theta_2 = \phi, \quad (\beta = \phi)$$

with

$$y_1 = x_0^2 + \ldots + x_{T-1}^2, \quad y_2 = x_1 x_0 + \ldots x_T x_{T-1},$$

while the remaining terms are

$$\Psi(\boldsymbol{\theta}) = 0, \quad c(\boldsymbol{y}) = -\frac{1}{2}x_T^2 - \ln 2\pi \cdot (T+1)/2 \,.$$

Example 2 : Poisson Regression Model

This example is adapted from Efron (1975, p1193).

Independent Poisson variables $X_1, \ldots, X_n$ have means $\lambda_i = a + \tau b_i$ where a and the b_i are known parameters. The parameter τ is such that $a + \tau b_i > 0$ for $i = 1 \ldots n$. From the exponential form of the likelihood

$$\theta_i = \ln(a + \tau b_i), \quad y_i = x_i, \quad (\beta = \tau)$$

with

$$\Psi(\boldsymbol{\theta}) = \sum_i e^{\theta_i}, \quad c(\boldsymbol{y}) = \sum_i \ln y_i! \,.$$

Example 3 : Nonlinear Regression Model

The nonlinear regression model can be cast (Amari, 1990, p154) as

$$X_i = f(c_i, \boldsymbol{\beta}) + \varepsilon_i$$

where the response variable X_i, $i = 1 \ldots n$ are NIID and $\varepsilon_i \sim N(0,1)$ without loss of generality. The nonlinear function f contains predictors c_i (known control parameters) and unknown m–dimensional parameters $\boldsymbol{\beta}$, the regression coefficients. Casting the resulting likelihood into the exponential form gives

$$\theta_i = f(c_i, \boldsymbol{\beta}), \quad y_i = x_i$$

with

$$\Psi(\boldsymbol{\theta}) = \frac{1}{2}\sum_i \theta_i^2, \quad c(\boldsymbol{y}) = -\frac{1}{2}\sum_i y_i^2 \,.$$

Example 4 : Generalized Linear Model

Independent random variables $Y_1, \ldots, Y_n$ follow a distribution of exponential type described by a contribution to the log–likelihood by a single observation y_i of

$$\ell_i = \ln f(y_i; \theta_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).$$

Considering such models with a scale parameter of 1, eg, Poisson, Bernoulli or Exponential errors, gives $a(\phi) = 1$, and so

$$\ell_i = \ln f(y_i; \theta_i) = y_i\theta_i - b(\theta_i) + c(y_i, \phi).$$

The log likelihood for the whole sample (ℓ) is given by

$$\ell = \sum_i \ell_i$$

due to the assumption of independence (McCullagh and Nelder, 1989, p24). This gives

$$\theta_i = \theta_i, \quad y_i = y_i, \quad \Psi(\boldsymbol{\theta}) = \sum_i b(\theta_i), \quad c(\boldsymbol{y}) = \sum_i c(y_i; \phi).$$

Also, the imbedding of the regression coefficients $\boldsymbol{\beta}$ is denoted by

$$\theta_i = \mathsf{f}(\boldsymbol{X}_i{}^\top \boldsymbol{\beta})$$

where $X_{1i}, \ldots, X_{mi}$ are the m predictors at each Y_i. The function f results from the use of the relations

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}, \; g(\boldsymbol{\mu}) = \boldsymbol{\eta} \text{ and } E(Y_i) = \mu_i = b'(\theta_i)$$

as described in Appendix B.7.

1.8 Tensor Notation

The concept of a *tensor* is fundamental to a differential geometric approach to statistical distributions. Rather than attempt to define all the terms and outline all notation in a single section, these will be introduced as needed. Some explanation

of elementary tensor notation is outlined below with references to texts on the topic.

There are several accessible references on the topic of the tensor calculus and its application to differential geometry. A brief but concise overview of tensor analysis can be found in Spiegel (1990), while a fuller treatment of tensor calculus with applications to differential geometry is given in Kay (1988). The latter text also includes a coordinate free approach. Tensor algebra (and related topics) is neatly expounded in Spain (1960), using the classical coordinate system approach. Finally, developments such as subspaces of a Riemannian manifold (Lovelock and Rund, 1989, p267) that are closely allied to statistical developments, will be used extensively throughout later chapters.

1.8.1 Indexing

Following the approach of Bishop and Goldberg (1980, p20), 'the customary tensor indexing of coordinates' is used hereafter, whereby the index is a superscript not a power; hence the results of Section 1.6 would be written as

$$\ln f(\boldsymbol{y};\boldsymbol{\theta}) = y_i\theta^i - \Psi(\boldsymbol{\theta}) + c(\boldsymbol{y})$$

since $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^n)$ is used as a coordinate system.

Powers are shown via the use of brackets, viz, $(\theta^i)^2$ or as x_i^2, if this is unambiguous.

1.8.2 Summation Convention

The convention used is the *sum index*, whereby summation is implied if the index is repeated, irrespective of location as a superscript or subscript.

A *nonsum index* is shown using an upper case index (Bishop and Goldberg, 1980), eg.

$$a^A e_A \, .$$

In cases where the upper case index convention is not used, it is sometimes desirable to use a lower case 'nonsum' index as in

$$a^i e_i \quad (i \text{ not summed})$$

in which case the explicit parenthetic comment spells this out.

1.8.3 Tensor Laws

A general definition is given for a tensor followed by specific examples.

In particular, *transformation laws* will be of utmost importance.

Assume a transformation of coordinates from $\boldsymbol{\theta}$ to $\bar{\boldsymbol{\theta}}$, viz,

$$\bar{\theta}^i = \phi^i(\theta^1 \ldots \theta^n)$$

and

$$\theta^i = \psi^i(\bar{\theta}^1 \ldots \bar{\theta}^n).$$

General Tensors

A general mixed tensor T or (r, s) tensor is defined by the transformation equation[8]

$$\bar{T}^{v_1 \ldots v_r}_{w_1 \ldots w_s} = \frac{\partial \bar{\theta}^{v_1}}{\partial \theta^{h_1}} \cdots \frac{\partial \bar{\theta}^{v_r}}{\partial \theta^{h_r}} \frac{\partial \theta^{k_1}}{\partial \bar{\theta}^{w_1}} \cdots \frac{\partial \theta^{k_s}}{\partial \bar{\theta}^{w_s}} T^{h_1 \ldots h_r}_{k_1 \ldots k_s}. \tag{1.3}$$

An example of the transformation law for a (1,2) tensor is given in Amari (1982a, p364, equation 2.27).

Contravariant Tensors

A contravariant tensor A or $(1, 0)$ tensor is defined as satisfying

$$\bar{A}^i = \frac{\partial \bar{\theta}^i}{\partial \theta^j} A^j \,. \tag{1.4}$$

[8]Kay (1988, p29, equation 3.14).

An example of a contravariant tensor is the *tangent* $\frac{\partial \theta^i}{\partial u}$ to the curve defined by $\theta^i = \theta^i(u)$, since if

$$A^i = \frac{\partial \theta^i}{\partial u}$$

then, by the chain rule,

$$\bar{A}^i = \frac{\partial \bar{\theta}^i}{\partial u} = \frac{\partial \bar{\theta}^i}{\partial \theta^j}\frac{\partial \theta^j}{\partial u} = \frac{\partial \bar{\theta}^i}{\partial \theta^j}A^j$$

as required.

Covariant Tensors

A covariant tensor B or $(0,1)$ tensor is defined by

$$\bar{B}_i = \frac{\partial \theta^j}{\partial \bar{\theta}^i}B_j \,. \tag{1.5}$$

An example of a covariant tensor is the *gradient* $\frac{\partial f}{\partial \theta^i}$ since

$$B_i = \frac{\partial f}{\partial \theta^i}$$

hence

$$\bar{B}_i = \frac{\partial f}{\partial \bar{\theta}^i} = \frac{\partial f}{\partial \theta^j}\frac{\partial \theta^j}{\partial \bar{\theta}^i} = B_j\frac{\partial \theta^j}{\partial \bar{\theta}^i}$$

as required.

1.8.4 Coordinate Free Methods

In his preface, McCullagh (1987) gives a delightful account of the balance between the 'coordinate–free' approach in tensor analysis and the mundane computation using indices. While the appeal of coordinate–free methods is obvious in the context of claiming invariance once tensorial behaviour is established, this is not the approach used in this thesis. Indeed, the very problem of examining the effects of reparameterization suggests that the pedestrian approach of working from one coordinate system to another is mandatory for problems in applied statistics. The proofs of invariance that are established use this very method and establish

invariance by using scalar forms for the tensorial quantities of interest. Furthermore, since empirical evidence from the nonlinear regression model suggests that parameter–effects curvature is the dominant effect (Bates and Watts, 1988), the ultimate question for a particular model in general will be 'what is the best way to cast the model?'. Thus, the method of moving from one coordinate system to another will be preferred in practice.

1.9 The Generalization

This Section is an attempt to provide a non–technical account of the generalization of curvature measures from Normal to non–Normal errors, with particular emphasis on the special case of generalized linear models.

For the most part, the development of curvature measures for non–Normal error models mirrors the approach used by Bates and Watts (1988) for nonlinear regression. For example, total curvature is shown to decompose into intrinsic and parameter–effects curvature as in nonlinear regression. The fundamental difference is the concept of an α–connection which is vital to a study of estimator behaviour in general, since key values of α are tied to particular properties of estimators, as shown in Section 2.10 and Section 2.11. It is possible to describe the key properties of the estimators (ie, key values of the α) as being 'unbundled' in the move from Normal to non–Normal errors, since all of these properties (unbiasedness, minimum variance and zero skewness) are satisfied simultaneously in the Normal case (Hougaard, 1982). This simultaneity is due to all the α–connections being identical when errors are Normal (Kass, 1984). By contrast, in the non–Normal situation, not all properties may be guaranteed to be satisfied together in the same estimator (Hougaard, 1982), since the α–connections are distinct in the general non–Normal case, as shown by Kass (1984).

Subject to the above caveat about the significance of key values of α, other facets of curvature measures generalize from the Normal to the non–Normal case. The decomposition of total curvature into normal and tangential components pro-

duces intrinsic and parameter–effects curvatures respectively, as for nonlinear regression. Following the suggestions of Kass (1984), there are *families* of such curvatures each depending on the chosen value of α, each having a special interpretation for estimator properties. A new feature is that a non–Normal error distribution contributes a component to both intrinsic and parameter–effects curvatures, whereas for Normal errors no such contribution is made from the error distribution. The precise interpretation and use of the curvatures (for a given α) is a topic which is exploited in later developments when special cases are considered, such as the class of generalized nonlinear models (GNMs) defined in Section 3.6. The following list of results, proved later in the thesis, shows the application of these generalized curvatures to specific situations.

Section 3.3.2 intrinsic curvature is invariant (*in general*).

Section 4.2 parameter–effects curvature is invariant for a generalized linear model (GLM).

Section 4.3.2 the scalar form of exponential intrinsic curvature for a GLM is minimal when the link is canonical; this is a generalization of the situation for nonlinear regression, where the canonical link (identity) yields zero intrinsic curvature.

Section 4.5 a generalized nonlinear model (GNM) with zero exponential curvature is a GLM with canonical link.

Section 4.7 a zero information connection implies a variance stabilizing link in a GLM and conversely.

Each of the curvatures, intrinsic and parameter–effects, consists of model and disturbance effects. Asymptotically and under appropriate replication, any disturbance component (based on the mean) becomes Normal and so the generalized nonlinear model (GNM) collapses into nonlinear regression. This is effectively the model proposed by Wei (1994), where the skewness is bounded so that it vanishes

asymptotically, ensuring Normality in the limit, in line with the result of Kass (1984). Hence, the question of which α–connection to use for these models becomes irrelevant since the error distribution becomes Normal in the limit. This bounded model of Wei (1994) has been extended (Wei and Zhu, 1997) to a class of models called 'exponential family nonlinear regression models' which are similar to the generalized nonlinear models(GNMs) described in Section 3.6. A comprehensive exposition of the approach is given in Wei (1998), together with applications using the construction of confidence regions and regression diagnostics involving leverage and influence estimates.

With the partitioning of generalized intrinsic curvature into effectively *two* components, one due to the model and the other due to the error distribution, the influence of design points on the curvature measures, and hence estimator behaviour, can be investigated. The purpose of such research would be to produce designed experiments that induce low intrinsic curvature, enabling current 'close–to–linear' methodology (Ratkowsky, 1983) to be employed. This 'close–to–linear' label assumes, as a precondition to possible model parameterization, that intrinsic curvature is low. Thus, this condition of low intrinsic curvature should be a feature of design rather than an implicit assumption, even though experimental evidence suggests that many of the models employed in practice exhibit low intrinsic curvature.

Chapter 2

Differential Geometric Approach

2.1 Preliminaries

Several discussants [Ross (1908a), Reid (1980), Atkinson (1980), and McCullagh (1980)] to the paper by Bates and Watts (1980) raised the question of a non–Normal error distribution. In particular, McCullagh (1980) noted that nonlinear models were often associated with non–Normal errors. Under these conditions, the use of least squares as a criterion for model fitting can give too much weight to a few outlying observations, so the use of the correct error distribution can be crucial for proper estimation. Two of the discussants, Reid (1980) and McCullagh (1980), proposed the exponential family as a model for non–Normality, and the form of model proposed by Reid (1980) was

$$p(\boldsymbol{y};\boldsymbol{\theta}) = exp\{c(\boldsymbol{y}) + \theta^i y_i - \Psi(\boldsymbol{\theta})\}, \tag{2.1}$$

for the random variables $Y_1, \ldots, Y_n$ with each Y_i having probability density function (pdf) $p(y_i;\theta_i)$. The quantities in bold denote *vectors* so that

$$\boldsymbol{y} = (y_1, \ldots, y_n)^\top$$

and

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top.$$

The following developments are in terms of a general probability distribution function $p(\boldsymbol{y};\boldsymbol{\theta})$, but the ultimate goal is to use the general technique in examining the exponential family. In particular, Generalized Linear Models (GLMs) will be investigated in detail, leading to other allied models.

2.1.1 Likelihood

Given the random variable $\boldsymbol{y}$ and a set of parameters $\boldsymbol{\theta}$, the distribution of $\boldsymbol{y}$ can be specified by the pdf $p(\boldsymbol{y};\boldsymbol{\theta})$. The corresponding log–likelihood is specified as

$$\ell(\boldsymbol{y};\boldsymbol{\theta}) = \ln p(\boldsymbol{y};\boldsymbol{\theta}).$$

2.1.2 Regularity Conditions

Subsequent developments rely on the following regularity conditions

1. $p(\boldsymbol{y};\boldsymbol{\theta}) > 0$.

2. For fixed $\boldsymbol{\theta}$, the functions

$$\partial_i \ell \stackrel{\text{def}}{=} \frac{\partial \ell(\boldsymbol{y};\boldsymbol{\theta})}{\partial \theta^i} \quad (i = 1 \ldots n)$$

   are linearly independent, in that it is assumed that they form a set of basis vectors to span the tangent space.

3. The moments of $\partial_i \ell$ exist up to required orders.

4. Partial derivatives and integration can always be interchanged, eg.,

   $\partial_i \left(\int f\right) = \int \left(\partial_i f\right)$.

These conditions will be assumed without being stated throughout the derivations and working hereafter.

2.2 Tangent Spaces

Let S be the space defined by the parameters $\boldsymbol{\theta}$ as a coordinate system. The tangent space $T_{\boldsymbol{\theta}}$ is a vector space obtained by a local linearization of S around $\boldsymbol{\theta}$ composed of tangent vectors to the coordinate curves passing through $\boldsymbol{\theta}$. The space $T_{\boldsymbol{\theta}}$ is spanned by the functions $\boldsymbol{e}_i$ known as basis vectors, given by

$$\boldsymbol{e}_i(\boldsymbol{\theta}) = \partial_i \ell(\boldsymbol{y}; \boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\partial \ell}{\partial \theta^i}$$

as shown in Figure 2.1.

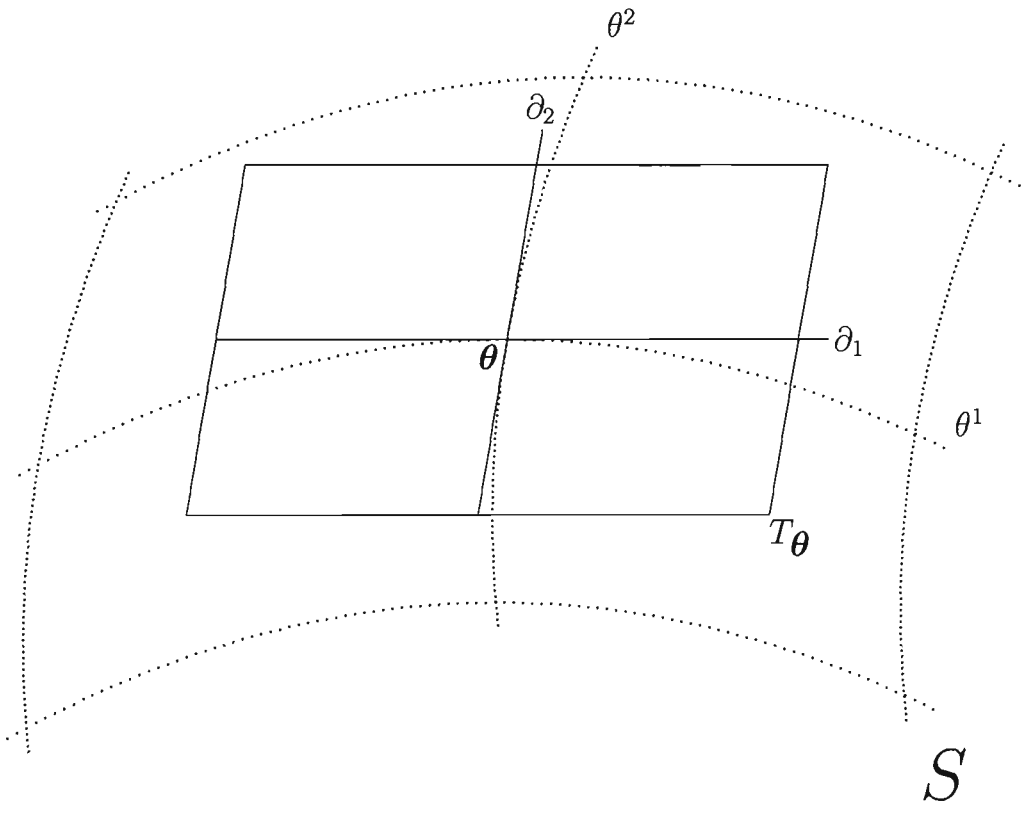

Figure 2.1: The tangent space T in parameter space S.

So any tangent vector $A \in T_{\boldsymbol{\theta}}$ is a linear combination of these basis vectors, $\boldsymbol{e}_i$, viz,

$$A = A^i \partial_i \ell = A^i \boldsymbol{e}_i(\boldsymbol{\theta})$$

as illustrated in Figure 2.2.

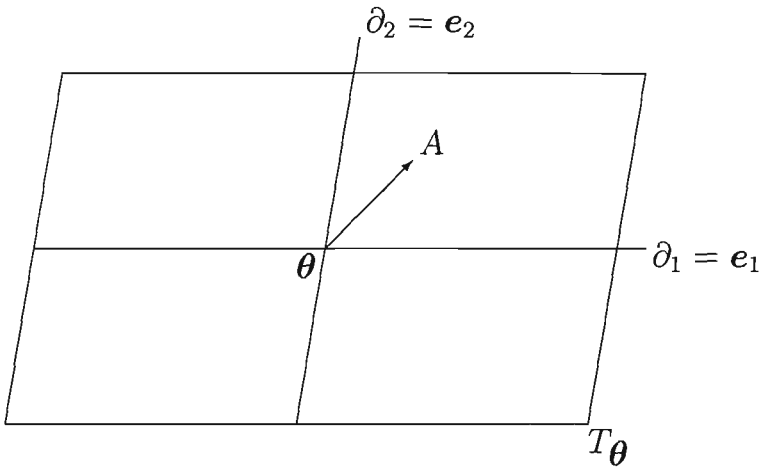

Figure 2.2: Basis vectors span the tangent space.

Consider a neighbouring point to $\boldsymbol{\theta}$, say $\boldsymbol{\theta}+d\boldsymbol{\theta}$. The two corresponding tangent spaces are shown in Figure 2.3, in two dimensions.

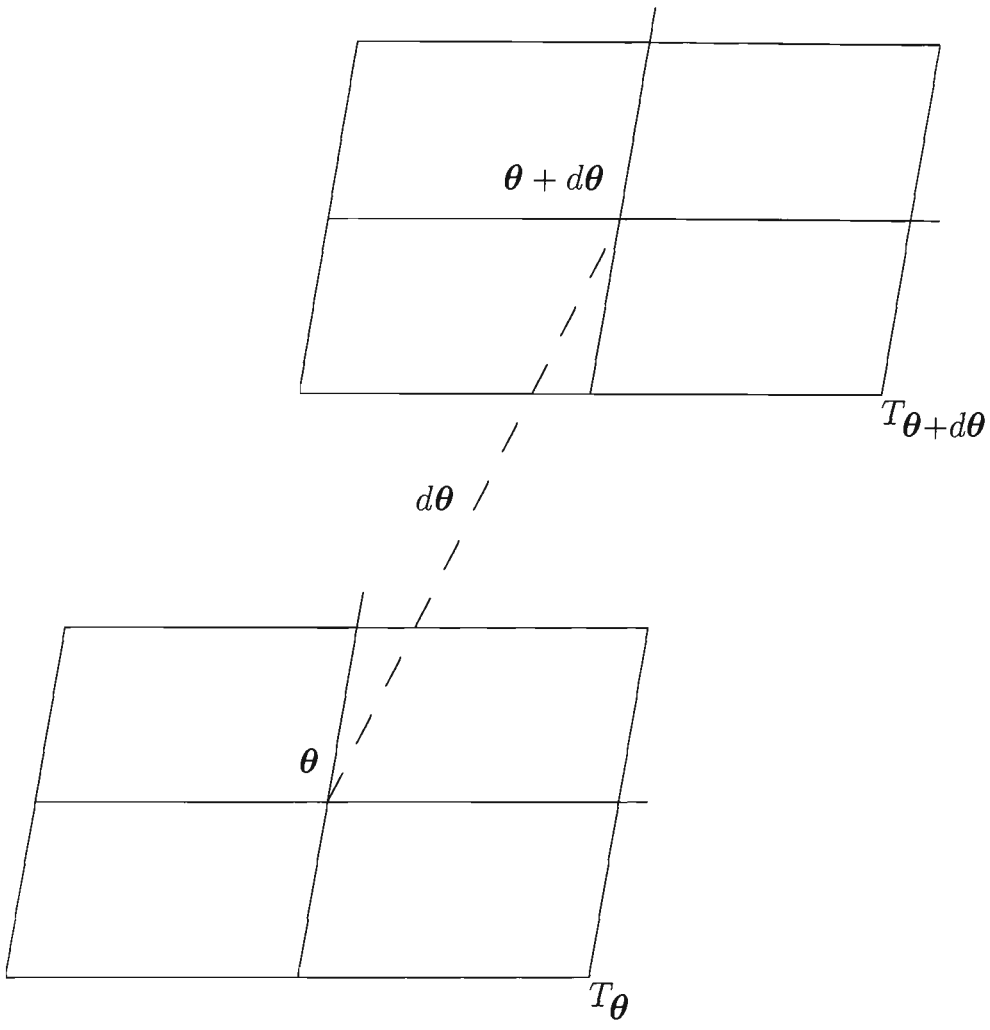

Figure 2.3: Neighbouring tangent spaces.

Now

$$\ell(\boldsymbol{y}; \boldsymbol{\theta} + d\boldsymbol{\theta}) = \ell(\boldsymbol{y}; \boldsymbol{\theta}) + \partial_i \ell d\theta^i + \ldots$$

by Taylor's theorem. This expansion can be recast as

$$\partial_i \ell d\theta^i \approx \ell(\boldsymbol{y}; \boldsymbol{\theta} + d\boldsymbol{\theta}) - \ell(\boldsymbol{y}; \boldsymbol{\theta}) = \boldsymbol{e}_i d\theta^i$$

(Amari, 1982a, p359), showing that the linear combination is described by $A^i = d\theta^i$, using the vector addition shown in Figure 2.4.

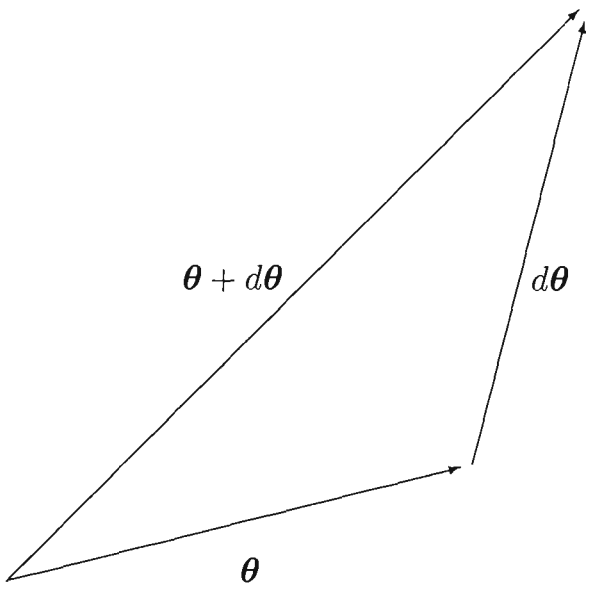

Figure 2.4: Vector addition for neighbouring parameter spaces.

2.3 Inner Product

Following the development in Barndorff–Nielsen, Cox and Reid (1986), a measure of the distance between the distributions at $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$ [due to Jeffreys (1961)] produces[1]

$$ds^2 = g_{ij} d\theta^i d\theta^j$$

where

$$\boldsymbol{e}_i \cdot \boldsymbol{e}_j = E_\theta(\partial_i \ell \partial_j \ell) \stackrel{\text{def}}{=} g_{ij}$$

using the basis vectors defined in the previous Section. Thus, g_{ij} is defined as an inner product and forms the *metric tensor* having the statistical interpretation

[1] A full derivation is given in Appendix B.1.

of corresponding to the Fisher information matrix. An alternative form for the tensor describing the information matrix is

$$g_{ij} = -E_\theta[\partial_i \partial_j \ell]$$

following Amari (1990, p29, 2.10). This form is derived in Appendix B.2. Both forms of the metric will be used extensively.

2.4 Metric Tensor

If P is a point on an n–dimensional surface specified by parameters $(\theta^1, \ldots, \theta^n)$, then the squared distance between P and the nearby point $P + dP$ specified by $(\theta^1 + d\theta^1, \ldots, \theta^n + d\theta^n)$, is given by

$$g_{ij} d\theta^i d\theta^j \tag{2.2}$$

using the Einstein summation convention. Note that superscripts are indexes, not powers.

(i) The terms g_{ij} $(g_{ij} = g_{ji})$ form the *metric tensor*, and

(ii) if the quadratic form given by (2.2) is positive definite, the surface forms a Riemannian manifold. A practical example of such a manifold is the surface of a sphere, which while being imbedded in Euclidean 3D space, forms a 2D Riemann manifold (Barndorff–Nielsen, Cox and Reid, 1986, pp83–84).

This metric tensor is the Fisher information matrix, when the distance function is measuring the infinitesimal distance between distributions.[2] In terms of the exponential family previously described we now have

$$g_{ij} = -E[\partial_i \partial_j \ell] = \partial_i \partial_j \Psi(\boldsymbol{\theta}) \tag{2.3}$$

where $\partial_i \ell = \dfrac{\partial \ell}{\partial \theta^i}$ and $\ell = \ln p(\boldsymbol{y}; \boldsymbol{\theta})$. [3] A collection of results involving the metric tensor is given in Appendix B.3. Before presenting the extensions required to

[2] Barndorff–Nielsen, Cox and Reid (1986, pp86–87).

[3] Amari, (1982a, p359).

handle general exponential families, five simple examples are presented, with the aim of demonstrating the role of the metric tensor in statistical problems.

2.4.1 Example 1, Normal distribution with known variance

Take $Y_1, \ldots, Y_n$ as a random sample where $Y_i \sim N(\mu_i, 1)$, with corresponding log–likelihood

$$\ell = \log p = -\frac{1}{2}\{y_i y_i + \mu^i \mu^i - 2y_i \mu^i\} - \frac{n}{2}\ln 2\pi = y_i \mu^i - \frac{1}{2}\mu^i \mu^i - \frac{n}{2}\ln 2\pi - \frac{1}{2}y^i y^i$$

giving

$$\theta^i = \mu^i, \text{ and } \Psi(\boldsymbol{\theta}) = \frac{1}{2}\theta^i \theta^i + \frac{n}{2}\ln 2\pi, \;\; c(\boldsymbol{y}) = -\frac{1}{2}y_i y_i,$$

and

$$g_{ij} = \partial_i \partial_j \Psi = \delta_{ij},$$

i.e., the Euclidean metric. So, the squared distance between $\theta^1, \ldots, \theta^n$ and $\theta^1 + d\theta^1, \ldots, \theta^n + d\theta^n$ is

$$ds^2 = (d\theta^1)^2 + \ldots + (d\theta^n)^2.$$

This demonstrates the correspondence between Normal errors and the Euclidean metric, when the parameter of interest is the mean of a Normal distribution with known variance.

2.4.2 Example 2, Normal distribution with known mean

Let $Z_1, \ldots, Z_n$ be a random sample where $Z_i \sim N(0, \sigma_i^2)$ giving

$$\ell = \log p = y_i \theta^i + \frac{1}{2}\mathcal{I}^i \ln(-2\theta^i) + \frac{n}{2}\ln 2\pi$$

with[4]

$$y_i = (z_i)^2, \;\; \theta^i = -\tfrac{1}{2(\sigma_i)^2}, \; \Psi(\boldsymbol{\theta}) = -\frac{1}{2}\mathcal{I}^i \ln(-2\theta^i) + \frac{n}{2}\ln 2\pi, \;\; c(\boldsymbol{y}) = 0.$$

[4] $\mathcal{I}$ is a unit scalar.

Amari (1982a) puts the constant term with $\Psi(\boldsymbol{\theta})$; it could equally as well go with $c(\boldsymbol{y})$. This produces the metric tensor as

$$g_{ij} = \frac{\delta_{ij}}{2(\theta^i)^2} = 2(\sigma_i)^4 \delta_{ij},$$

a not unexpected result, being proportional to the variance of the sample variance. So now the distance function is

$$2(\sigma_1)^4 (d\theta^1)^2 + \ldots + 2(\sigma_n)^4 (d\theta^n)^2$$

ie., a non–Euclidean metric. For a non–statistical example of a simple non–Euclidean metric, see Barndorff–Nielsen, Cox and Reid (1986, p84).

2.4.3 Example 3, Normal distribution

Take $X_1, \ldots, X_n$ as a random sample where $X_i \sim N(\mu_i, \sigma_i^2)$. As the observations are independently distributed, the log–likelihood is the sum of the individual terms. Thus the analysis can be undertaken for a single observation (McCullagh and Nelder, 1983, p17, p20 and p32). Thus the corresponding log–likelihood (contribution) for a single observation is

$$\ell = \frac{1}{2}(x-\mu)^2/\sigma^2 - \ln\sigma - \frac{1}{2}\ln 2\pi \quad , \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}.$$

This gives

$$\partial_1 \ell = -\frac{(x-\mu)}{\sigma^2}(-1) = \frac{x-\mu}{\sigma^2}$$

and

$$\partial_2 \ell = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}$$

with

$$\partial_1\partial_1 \ell = -\frac{1}{\sigma^2} \text{ and } \partial_2\partial_2 \ell = -\frac{3(x-\mu)^2}{\sigma^4} - \frac{(-1)}{\sigma^2}$$

and

$$\partial_2\partial_1 \ell = -\frac{2(x-\mu)}{\sigma^3} = \partial_1\partial_2 \ell.$$

The contribution to the metric tensor from an individual observation is

$$-E\begin{bmatrix} \partial_1\partial_1\ell & \partial_1\partial_2\ell \\ \partial_2\partial_1\ell & \partial_2\partial_2\ell \end{bmatrix} = E\begin{bmatrix} 1/\sigma^2 & 2(x-\mu)/\sigma^3 \\ 2(x-\mu)/\sigma^3 & 3(x-\mu)^2/\sigma^4 - 1/\sigma^2 \end{bmatrix} = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix}$$

See Murray and Rice (1993, p17) and Amari (1990, p29, Example 2.3) for corresponding derivations. For independent observations $x_1, \ldots, x_n$, the parameters $\boldsymbol{\theta}$ become

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \vdots \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \sigma_1 \\ \mu_2 \\ \sigma_2 \\ \vdots \end{pmatrix}$$

which makes the metric tensor

$$g_{ij}(\boldsymbol{\theta}) = -E(\partial_i\partial_j\ell)$$

a diagonal matrix with the entries grouped in pairs, viz,

$$Diag[1/\sigma_1^2, 2/\sigma_1^2, \ldots, 1/\sigma_n^2, 2/\sigma_n^2].$$

Since $g_{ij} \neq \delta_{ij}$, the geometry is not Euclidean[5]. If the observations are $NIID$, then $\mu_i = \mu$ and $\sigma_i = \sigma$ and the observations become replicates, yielding

$$\ell(x_1, \ldots, x_n) = n\ell(\bar{x})$$

in a similar vein to the development in Amari (1982a, p372, 4.17). The distribution can be cast as a member of the exponential family, as defined by Equation (2.1). Using the likelihood contribution for a single observation

$$\ell = -(x^2 + \mu^2 - 2x\mu)/2\sigma^2 - \ln\sigma - \ln(2\pi)/2 = (\frac{-1}{2\sigma^2})x^2 + x(\frac{\mu}{\sigma^2}) - \frac{\mu^2}{2\sigma^2} - \ldots$$

gives

$$y_1 = x^2,\ y_2 = x,\ \ \theta^1 = -1/2\sigma^2,\ \theta^2 = \mu/\sigma^2,$$

[5]The full dimension of g_{ij} is in fact $2n \times 2n$.

and

$$\Psi(\boldsymbol{\theta}) = -\ln(-2\theta^1)/2 + (\theta^2)^2/4\theta^1, \;\; c(\boldsymbol{y}) = \ln(2\pi)/2$$

in agreement with Examples 1 and 2. As before, the full n independent observations will produce $2n$ parameters for $\boldsymbol{\theta}$.

2.4.4 Example 4, Multinomial distribution

The following discussion is a synthesis of Amari (1990, p12, p24, 2.2 and p31, 2.4), with corrections and changes in notation where considered necessary. The description of the multinomial distribution is cast in the particular form to highlight that it is a mixture distribution (Amari, 1990, p40). The pdf and likelihood are quoted for a single observation.

Let Y be a random variable taking integer values $\{1, 2, 3, \ldots, n+1\}$ with π_i being the probability that y is equal to i, and

$$\sum_{i=1}^{n+1} \pi_i = 1, \; 0 < \pi_i < 1, \; i = 1 \ldots n+1.$$

The probabilities π_i define a *multinomial* distribution with

$$\theta^1 = \pi_1, \; \theta^2 = \pi_2, \ldots, \theta^n = \pi_n, \theta^{n+1} = \pi_{n+1} = 1 - \sum_{i=1}^{n} \pi_i.$$

The probability distribution function for a *single* observation is

$$p(y; \boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\delta(y-i)\theta^i\right] + \delta(y-n-1)\left(1 - \sum_{i=1}^{n} \theta^i\right)$$

where $\delta(y-i) = 1$ when $y = i$, and $= 0$ otherwise, ie., an indicator variable. Thus

$$p(y; \boldsymbol{\theta}) = \sum_{i=1}^{n+1} \delta(y-i)\theta^i$$

giving[6]

$$\ell(y; \boldsymbol{\theta}) = \sum_{i=1}^{n+1} \delta(y-i) \ln \theta^i$$

and so

$$\partial_i \ell = \frac{\delta(y-i)}{\theta^i} + (-1)\frac{\delta(y-n-1)}{\theta^{n+1}}$$

[6]Note the typographical error in Amari (1990, p24).

since

$$\theta^{n+1} = 1 - \theta^1 \ldots - \theta^n$$

Calculating $\partial_i \partial_j \ell$, gives

$$\partial_i \partial_i \ell = \frac{\delta(y-i)}{(\theta^i)^2}(-1) - \frac{\delta(y-n-1)}{(\theta^{n+1})^2}(-1)(-1)$$

and then[7]

$$\partial_i \partial_j \ell = 0 - \frac{\delta(y-n-1)}{(\theta^{n+1})^2}(-1)(-1), \quad i \neq j.$$

The metric tensor is then

$$g_{ij} = -E\partial_i \partial_j \ell = -\int p\ \partial_i \partial_j \ell\ dy = -\sum_y p\ \partial_i \partial_j \ell$$

$$= \delta_{ij}(\theta^i)^{-2} p_i + \frac{1}{(\theta^{n+1})^2} p_{n+1}$$

But $p_i = \pi_i = \theta^i$ and so

$$g_{ij} = \delta_{ij}(\pi_i)^{-1} + (\pi_{n+1})^{-1}$$

as per Amari (1990, p31). Finally

$$g_{ij} = \frac{\delta_{ij}}{\pi_i} + \frac{1}{1 - \pi_1 \ldots - \pi_n}.$$

The special case of $n = 1$ gives the Bernoulli distribution

$$\theta^1 = \pi,\ \theta^2 = 1 - \pi$$

yielding

$$g_{11} = \frac{1}{\pi} + \frac{1}{1-\pi} = i(\pi) = \frac{1}{\pi(1-\pi)}$$

and

$$g^{11} = V(\hat{\pi}) = \pi(1-\pi)$$

as expected. The multinomial distribution is an important example of a *mixture* distribution. (See Amari, 1990, p43, example 2.6.)

[7]There is a typographical error in Amari (1990, p31).

2.4.5 Example 5, Generalized Linear Model

A random sample $Y_1, \ldots, Y_n$ is taken from a distribution belonging to the special class of exponential family models whose individual contribution to the log–likelihood is given by

$$\ell = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

ie., a generalized linear model (GLM). Choose $a(\phi) = 1$ for simplicity, and so

$$\partial_1 \ell = y - b'(\theta), \;\; \partial_1 \partial_1 \ell = -b''(\theta).$$

For independent observations $y_1, \ldots, y_n$, the GLM form is close to the exponential family model defined by Equation (2.1) on page 35. Thus the log–likelihood becomes

$$\ell(\boldsymbol{y}; \boldsymbol{\theta}) = \theta^i y_i - \sum_i^n b(\theta^i) + c(\boldsymbol{y}; \phi); \;\; c(\boldsymbol{y}; \phi) = \sum_i^n c(y_i; \phi),$$

with[8]

$$y_i = y_i, \;\; \theta^i = \theta^i (= \theta^i / a(\phi)), \;\; \Psi(\boldsymbol{\theta}) = b(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_i^n b(\theta^i), \;\; \text{and} \;\; c(\boldsymbol{y}) = c(\boldsymbol{y}, \phi).$$

Now

$$\partial_i \ell = y_i - b'(\theta^i)$$

and

$$\partial_i \partial_j \ell = -b''(\theta^I) \delta_{ij}$$

where I is a nonsum index taking on the same value as i. This gives the metric tensor as

$$g_{ij}(\boldsymbol{\theta}) = -E \partial_i \partial_j \ell = b''(\theta^I) \delta_{ij}$$

so

$$g_{ij}(\boldsymbol{\theta}) = \text{diag}\left(b''(\theta^1), \ldots, b''(\theta^n)\right),$$

which reinforces the independence of the assumed sampling regime.

[8] Technically $\Psi(\boldsymbol{\theta}) = b(\boldsymbol{\theta})/a(\phi)$.

2.5 Affine Connection

If $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ are the tangent spaces corresponding to the neighbouring points $\boldsymbol{\theta}$ and $\boldsymbol{\theta}+d\boldsymbol{\theta}$, then the *affine connection* $\Gamma^k_{ji}(\boldsymbol{\theta})$ provides the means of comparing these two spaces.

A direct comparison of vector components from the two spaces is not possible since the tangent vectors from each space have differing basis vectors as shown in Figure 2.5.

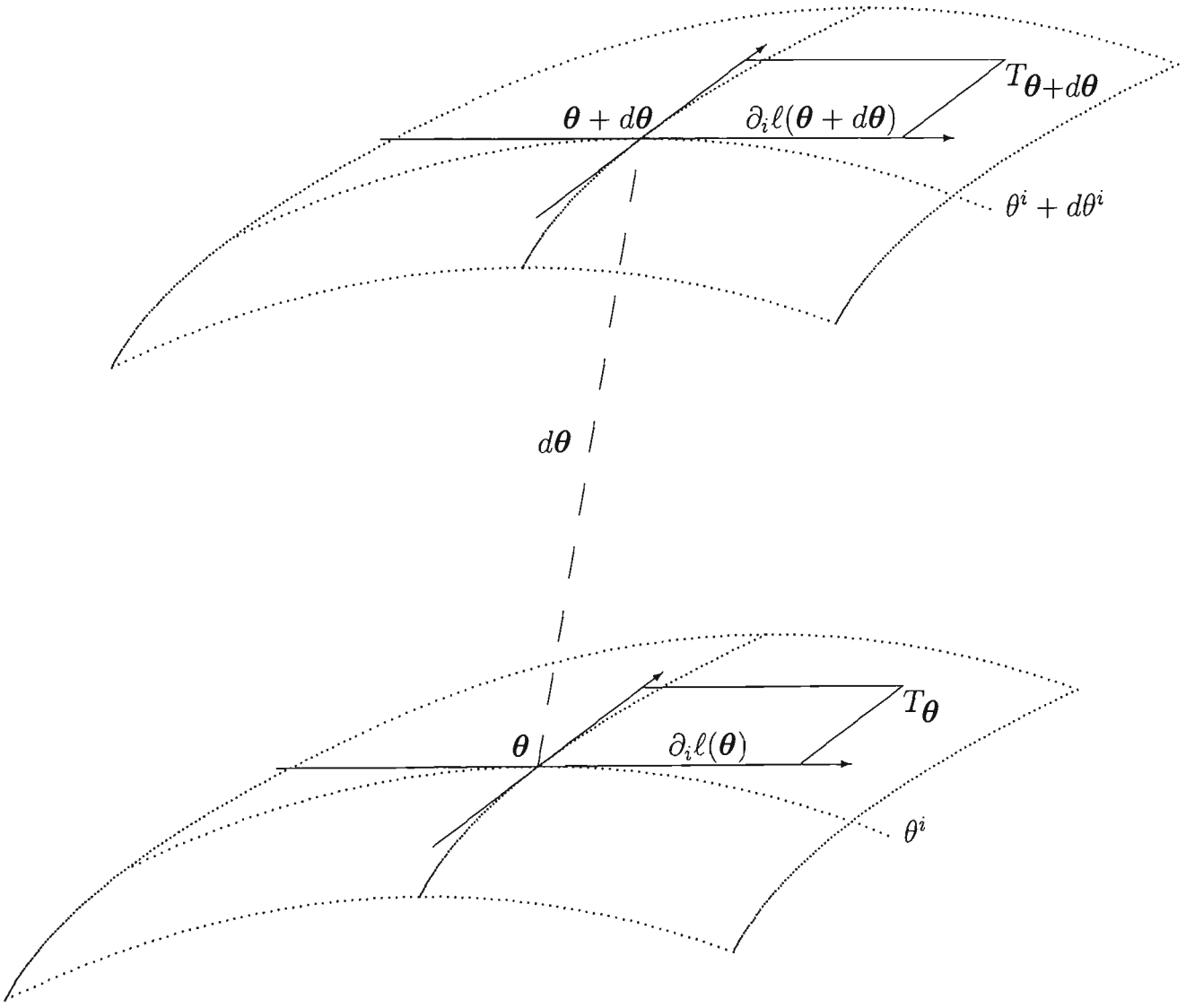

Figure 2.5: The basis vectors for neighbouring tangent spaces.

Even in Euclidean space, the basis vectors differ if the coordinate system is

curvilinear, eg., for spherical coordinates. The statistical analogue is nonlinear regression, since then the errors are Normal (implying a Euclidean space), but the coordinate system (regression parameter space) is curvilinear.

To compare vectors in $T_{\boldsymbol{\theta}}$ with those in $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ a mapping between these vector spaces is needed; such a correspondence is called *affine*, Bishop and Goldberg (1980, p220).

Choose a basis vector $\boldsymbol{e}_i(\boldsymbol{\theta}+d\boldsymbol{\theta})$ in $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$ and consider its corresponding vector in $T_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{e}_i(\boldsymbol{\theta})$ in $T_{\boldsymbol{\theta}}$, as in the Figure 2.6.

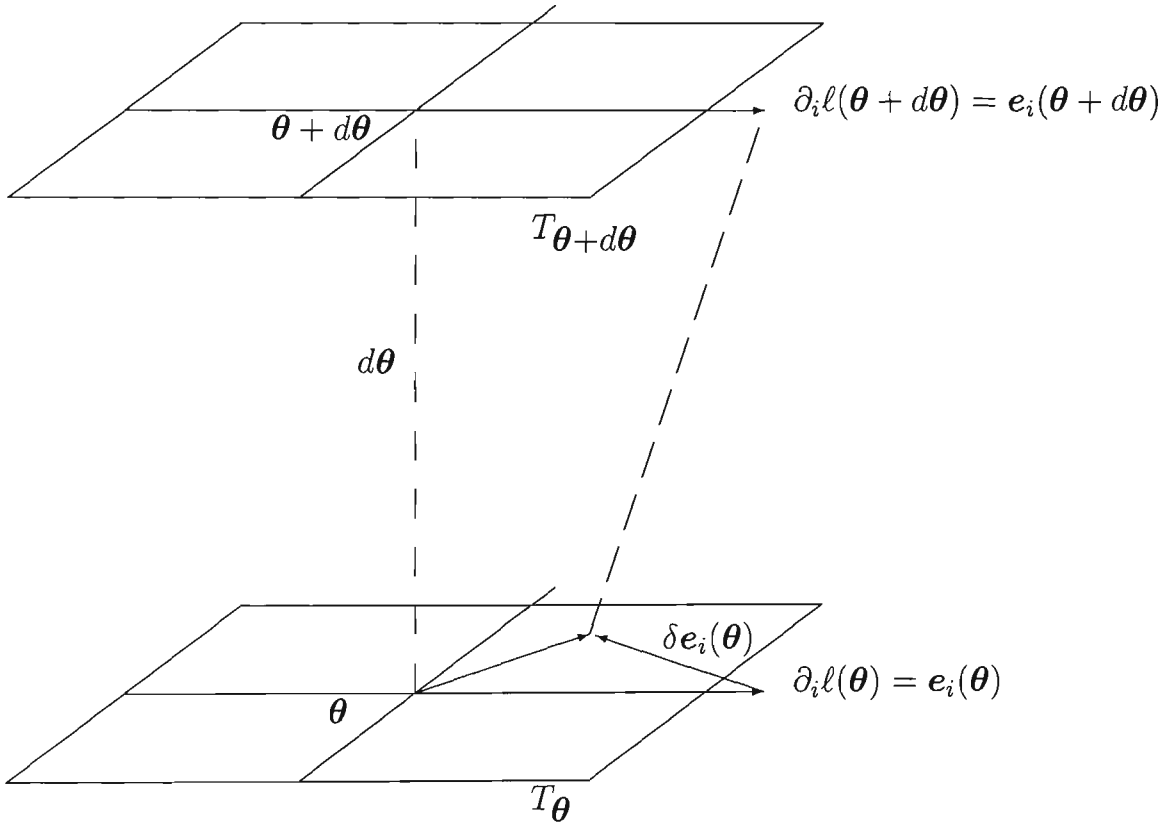

Figure 2.6: The correspondence between neighbouring basis vector spaces.

The corresponding vector in $T_{\boldsymbol{\theta}}$ is

$$\partial_i(\boldsymbol{\theta}) + \delta \boldsymbol{e}_i(\boldsymbol{\theta})$$

which means that

$$\delta \boldsymbol{e}_i(\boldsymbol{\theta}) \approx \partial_i \ell(\boldsymbol{\theta} + d\boldsymbol{\theta}) - \partial_i \ell(\boldsymbol{\theta}).$$

This difference $\delta \boldsymbol{e}_i(\boldsymbol{\theta})$ can be expressed as

$$\delta \boldsymbol{e}_i(\boldsymbol{\theta}) = d\theta^j \ \Gamma^k_{ji} \boldsymbol{e}_k(\boldsymbol{\theta}),$$

being the change in the ith basis vector while moving from $\boldsymbol{\theta}$ to $\boldsymbol{\theta} + d\boldsymbol{\theta}$. The n^3 functions Γ^k_{ji} are the coefficients of the *affine* connection, since they determine the affine correspondence between $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}+d\boldsymbol{\theta}}$.

The coefficient Γ^k_{ji} determines the influence of $\boldsymbol{e}_i$ on the change in $\boldsymbol{e}_k$ when moving a small distance in the θ^j direction. (Barndorff–Nielsen, Cox and Reid, 1986).[9]

Taking the inner product gives

$$\boldsymbol{e}_m \cdot \delta \boldsymbol{e}_i = \boldsymbol{e}_m \cdot d\theta^j \ \Gamma^k_{ji} \boldsymbol{e}_k = g_{km} \ \Gamma^k_{ji} d\theta^j = d\theta^j \ \Gamma_{jim}$$

using the identity

$$\Gamma_{jim} = \ \Gamma^k_{ji} g_{km} \ ,$$

following Kreyszig (1991, pp140–141). The forms Γ_{jim} and Γ^k_{ji} are *Christoffel symbols* of the first and second kind, having the interpretation of being inner products when the space is Euclidean.

Since a point P in the vector space is associated with

$$\ell(\boldsymbol{y}; \boldsymbol{\theta}) \stackrel{\text{def}}{=} \ln f(\boldsymbol{y}; \boldsymbol{\theta})$$

then

$$\partial_i \ell(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{\partial}{\partial \theta^i} \ell(\boldsymbol{y}; \boldsymbol{\theta})$$

[9] An alternative interpretation of the affine connection using the covariant derivative is given in Chapter 3.

will be associated with $T_{\boldsymbol{\theta}}$.

Now

$$\partial_i \ell(\boldsymbol{y}; \boldsymbol{\theta} + d\boldsymbol{\theta}) = \partial_i \ell(\boldsymbol{y}; \boldsymbol{\theta}) + \partial_j \partial_i \ell(\boldsymbol{y}; \boldsymbol{\theta}) d\theta^j$$

by Taylor's theorem, but, by virtue of the score statistic

$$E[\partial_i \ell(\boldsymbol{y}; \boldsymbol{\theta})] = 0.$$

Note the discrepancy between Amari (1982a, p361), and Amari (1990, p39), in the expansion of the derivative of the likelihood. The correct form given here is equivalent to that of Amari (1990, p39). Thus, any vector $\mathcal{L}(\boldsymbol{y})$ in $T_{\boldsymbol{\theta}}$ should satisfy

$$E[\mathcal{L}(\boldsymbol{y})] = 0$$

as well.[10] To be precise, vectors of type $\mathcal{L}(\boldsymbol{y})$ are contained in the 1–representation of the tangent space $T_{\boldsymbol{\theta}}$ denoted $T^{(1)}_{\boldsymbol{\theta}}$ defined by

$$T^{(1)}_{\boldsymbol{\theta}} = \{A(\boldsymbol{y}) | A(\boldsymbol{y}) = A^i \partial_i \ell(\boldsymbol{y}; \boldsymbol{\theta})\}.$$

Since

$$E(\partial_j \partial_i \ell) = -E(\partial_i \ell \partial_j \ell) = -g_{ji}$$

then $\partial_j \partial_i \ell$ is not contained in $T^{(1)}_{\boldsymbol{\theta}}$.

Adding $\partial_j \partial_i \ell$ to g_{ji} and to $\partial_i \ell \partial_j \ell$ yields two quantities of type $\mathcal{L} \in T^{(1)}_{\boldsymbol{\theta}}$.

These two quantities will be denoted respectively by

$$\overset{1}{\delta}_i (\boldsymbol{y}; \boldsymbol{\theta}) = \delta \boldsymbol{e}_i + g_{ji} d\theta^j = \partial_j \partial_i \ell d\theta^j + g_{ji} d\theta^j$$

and

$$\overset{2}{\delta}_i (\boldsymbol{y}; \boldsymbol{\theta}) = \delta \boldsymbol{e}_i + \partial_i \ell \partial_j \ell d\theta^j = \partial_j \partial_i \ell d\theta^j + \partial_i \ell \partial_j \ell d\theta^j.$$

[10] This notation clarifies a possible misinterpretation of equation (2.16) p361 of Amari (1982a). The use of the likelihood symbol ℓ for the expression of vectors in $T_{\boldsymbol{\theta}}$ is potentially confusing. Hence the use of $\mathcal{L}(\boldsymbol{y})$ which corresponds to the $A(\boldsymbol{x})$ of Amari (1990, pp19–20).

These two new quantities satisfy the expectation criterion

$$E(\mathcal{L}) = E(\overset{1}{\delta}_i) = E(\overset{2}{\delta}_i) = 0.$$

Any linear combination of these two new quantities would suffice. Thus the function

$$\overset{\alpha}{\delta}_i \ell = \frac{1+\alpha}{2} \overset{1}{\delta}_i + \frac{1-\alpha}{2} \overset{2}{\delta}_i$$

could be used, where α is the arbitrary constant of combination. Thus there is no unique affine connection, but a family characterised by the parameter α. Connections belonging to this family are called α–connections.

2.6 α–connections

Taking the inner product as before, but now using the α value to characterize the connection, yields

$$\boldsymbol{e}_m \cdot \overset{\alpha}{\delta}_i \ell = d\theta^j \ \overset{\alpha}{\Gamma}_{jim} = E[\partial_m \ell \ \overset{\alpha}{\delta}_i \ell]$$

Thus

$$\overset{\alpha}{\Gamma}_{jim} = E[\partial_m \ell \{ \frac{1+\alpha}{2} (\partial_j \partial_i \ell + g_{ji}) + \frac{1-\alpha}{2} (\partial_j \partial_i \ell + \partial_i \ell \partial_j \ell) \}]$$

which simplifies to

$$\overset{\alpha}{\Gamma}_{jim} = E[\partial_j \partial_i \ell \partial_m \ell] + \frac{1-\alpha}{2} E[\partial_j \ell \partial_i \ell \partial_m \ell] \tag{2.4}$$

These α–connections provide the means of comparing nearby tangent spaces that are derived from probability distributions. In short, the α–connection is the affine connection for a function that is derived from a statistical distribution.

2.7 Statistical Interpretation of α–connections

The one–parameter affine connection (α–connection) can be rewritten as

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = E[\partial_i\partial_j\ell(\boldsymbol{y};\boldsymbol{\theta})\partial_k\ell(\boldsymbol{y};\boldsymbol{\theta})] + \frac{1-\alpha}{2}E[\partial_i\ell(\boldsymbol{y};\boldsymbol{\theta})\partial_j\ell(\boldsymbol{y};\boldsymbol{\theta})\partial_k\ell(\boldsymbol{y};\boldsymbol{\theta})] \quad (2.5)$$

where α is the arbitrary constant of combination.

A statistical interpretation of these α–connections has been derived by Kass (1984), who showed that the parameterisation of one–dimensional non–linear models derived by Hougaard (1982) was related to Amari's (1982a) α–connections via

$$\delta = \frac{1-\alpha}{2}$$

where the value of δ determines the form of parameterisation in the original exponential form

$$e^{\theta' t(x) - \chi(\theta)}$$

where

$$\theta = \theta(\beta)$$

and the new parameter $\psi = g(\beta)$ is determined by solving Hougaard's equation[11], viz

$$\frac{d^2g/d\beta^2}{dg/d\beta} = \left\{\delta\frac{d^3\chi}{d\theta^3}\left(\frac{d\theta}{d\beta}\right) + \left(\frac{d^2\theta}{d\beta^2}\right)'\frac{d^2\chi}{d\beta^2}\frac{d\theta}{d\beta}\right\} \Big/ \left\{\left(\frac{d\theta}{d\beta}\right)'\frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\}. \quad (2.6)$$

Kass (1984), by the change of variable technique, was able to show the equivalence of Amari's (1982a) definition of an α–connection and Hougaard's equation[12] for one–dimensional curved exponential families.

Kass (1984) used this demonstrated equivalence between α and δ with Hougaard's (1982) results for δ to provide the statistical interpretation for Amari's (1982a) α–connections. The analysis provided below uses the equivalence shown by Kass (1984), but interprets the α–connections *directly*, via the equations due to Bartlett (1953a). This use of Bartlett's equations and the corresponding derivations are considered original.

[11] Hougaard (1982, p246, equation 2.1).

[12] Kass (1984, p92), where equations (3) (Hougaard, 1982) and (6c) (Amari, 1982a) are shown to be equivalent.

The one–dimensional procedure outlined below will be generalized to the multi–parameter case later.

2.7.1 Riemann Christoffel Curvature

The statistical interpretation of α–connections will be made in terms of those properties induced by zeroing the α–connection. The identical vanishing of an α–connection can be cast in terms of the Riemann Christoffel curvature, as described in Appendix B.4.

The Riemann Christoffel curvature tensor is basic to many of the quantities used in the differential geometric treatment of the statistical theory of curved exponential families. Key treatments are given in Amari (1990, p46), Amari (1982a, p365, 3.5), Barndorff–Nielsen, Cox and Reid (1986, p89) and Loveluck and Rund (1989, p257, 3.1 and p260, 3.16).

2.8 Equivalence of α, δ and c.

Amari (1982a) derives an α–connection as

$$\overset{\alpha}{\Gamma}_{ijk} = E\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\frac{\partial \ell}{\partial\theta_k}\right) + \frac{1-\alpha}{2}E\left(\frac{\partial \ell}{\partial\theta_i}\frac{\partial \ell}{\partial\theta_j}\frac{\partial \ell}{\partial\theta_k}\right)$$

where ℓ is the log–likelihood function.

For a *one–dimensional* model this reduces to

$$\overset{\alpha}{\Gamma}_{\theta} = E\left(\frac{\partial^2 \ell}{\partial\theta^2}\frac{\partial \ell}{\partial\theta}\right) + \frac{1-\alpha}{2}E\left(\frac{\partial \ell}{\partial\theta}\right)^3$$

Using the fact that $\alpha = 1 - 2\delta$ gives

$$\overset{\alpha}{\Gamma}_{\theta} = E\left(\frac{\partial^2 \ell}{\partial\theta^2}\frac{\partial \ell}{\partial\theta}\right) + \delta E\left(\frac{\partial \ell}{\partial\theta}\right)^3$$

which is equivalent to (6c) of Kass (1984, p88), in the one–dimensional case.

DiCiccio (1984) extends the above interpretation of properties associated with specific values of Hougaard's (1982) δ to a wider class of models in the one parameter case. The key variable used by DiCiccio (1984) was designated as c where

$$\delta = (2 - c)/3.$$

α	δ	c	Parameterisation	Transformation
-1	1	-1	Mean value	Bias reducing
-1/3	2/3	0		Skewness reducing
0	1/2	1/2		Variance stabilizing
1/3	1/3	1	$E[\partial^3\ell/\partial\psi^3] = 0$	'Normal likelihood'
1	0	2	Canonical	None – identity

Table 2.1: The interpretation of α, δ and c.

Table 2.1 summarises the special values of α, δ and c, together with their corresponding descriptions. These special values of α will be shown in Section 2.10 to be connected to special properties of the estimator, as shown under the heading 'Transformation' in Table 2.1. For example, for $\alpha = -1/3$ the special property associated with choosing a transformation that induces

$$\overset{-1/3}{\Gamma} = 0$$

will be to reduce skewness. The various properties described for each value of α are derived in the single and multi–parameter cases in Section 2.10 and Section 2.11, respectively. The choice of names (Amari, 1982a) for the connections α = -1, 0 and 1 as *mixture*, *information* and *exponential* is clear from how Table 2.1 links α, δ, c, the parameterization and the effect of the corresponding transformation.

2.9 Bartlett's Equations

Bartlett's (1953a) used the following notation for one–dimensional models[13]

$$L = \ln p(\theta; x).$$

The score statistics and other quantities were defined as

$$L_1 \equiv E\left(\frac{\partial L}{\partial \theta}\right)$$

[13]L is the log–likelihood.

$$L_2 \equiv E\left(\frac{\partial^2 L}{\partial\theta^2}\right)$$

$$_1L_2 \equiv \frac{\partial L_2}{\partial\theta}$$

$$L_1^{(2)} \equiv E\left(\frac{\partial L}{\partial\theta}\right)^2 = I$$

$$(L_1L_2) \equiv E\left(\frac{\partial L}{\partial\theta}\frac{\partial^2 L}{\partial\theta^2}\right).$$

Repeated operations on the log–likelihood yield the following relations

$$L_1 = 0 \tag{2.7}$$

$$L_2 + L_1^{(2)} = 0 \tag{2.8}$$

$$L_3 + 3(L_1L_2) + L_1^{(3)} = 0 \tag{2.9}$$

$$_1L_2 + L_1^{(3)} + 2(L_1L_2) = 0. \tag{2.10}$$

The last two Equations [(2.10) and (2.9)] can be combined to give

$$L_3 + (L_1L_2) - {}_1L_2 = 0\,. \tag{2.11}$$

Subtracting Equation (2.11) from Equation (2.10) gives

$$2_1L_2 + L_1^{(3)} + (L_1L_2) - L_3 = 0$$

which yields

$$(L_1L_2) + L_1^{(3)} = L_3 - 2_1L_2. \tag{2.12}$$

These equations can now be used to examine $\overset{\alpha}{\Gamma}_\theta$ for the key values of α, via

$$\overset{\alpha}{\Gamma}_\theta = (L_1L_2) + \frac{1-\alpha}{2}L_1^{(3)}.$$

The above results are re-expressions of results derived by Bartlett (1953a). In the next section, those properties corresponding to particular values of α will be investigated.

2.10 Interpretation of α in the one parameter case

The α–connection is now expressed in terms of a *transformed* parameter ψ, as determined indirectly by Kass (1984) in the one–parameter case[14]. In fact, the transformation used by Hougaard (1982) was from the imbedded regression coefficient β to ψ via

$$\psi = g(\beta)$$

rather than from the canonical parameter θ as used by Wedderburn, as quoted in Hougaard (1982).

The results following have been expressed in terms of this transformed parameter ψ to emphasise the interpretation associated with each α–connection via the use of equation (4) of Kass (1984), ie.,

$$\overset{\alpha}{\Gamma}_{\psi} = 0\,.$$

So, in general

$$\overset{\alpha}{\Gamma}_{\psi} = (L_1 L_2) + \frac{1-\alpha}{2} L_1^{(3)}.$$

Each case is now examined in turn, for one–parameter models.

Some of these results hold not only for curved exponential families but also for general likelihoods. The acronyms CEF (Curved Exponential Family) and GL (General Likelihood) will be appended to results which hold for each of these respective situations. These acronyms are used in Table 2.2 to summarise the results for the interpretation of each value of α.

[14]Hougaard's (1982) ψ is the γ of Kass (1984).

2.10.1 Mixture Connection

$\underline{\alpha = -1}$: $(\delta = 1)$ (GL)

Using Equation (2.12) in the form for the (-1)–connection gives

$$\overset{-1}{\Gamma}_{\psi} = (L_1 L_2) + L_1^{(3)} = L_3 - 2\,{}_1L_2$$

to give

$$\overset{-1}{\Gamma}_{\psi} = E\left(\frac{\partial^3 L}{\partial \psi^3}\right) - 2\frac{\partial}{\partial \psi}\left(E\frac{\partial^2 L}{\partial \psi^2}\right).$$

Compare this with the expansion of Bartlett (1953b), viz[15]

$$E(\widehat{\theta}) = \theta - \left\{\frac{1}{2}E\frac{\partial^3 L}{\partial \theta^3} + \frac{\partial I}{\partial \theta}\right\}/I^2 + \ldots \tag{2.13}$$

So, choosing a parameterization corresponding to $\delta = 1$ will reduce asymptotic bias, since this is equivalent to making $\overset{-1}{\Gamma}_{\psi} = 0$. Hougaard (1982, p248) shows that, for the case $\delta = 1$, the ψ parameterization can be estimated with zero bias asymptotically in a curved exponential family, viz,

$$E\widehat{\psi} = \psi + O(m^{-1})$$

where m is the number of data points[16]. In addition, this parameterization $(\delta = 1)$ minimises mean square error (MSQ) where

$$MSQ(\widehat{\psi}) = E(\widehat{\psi} - \psi)^2 = E(\widehat{\psi} - E\widehat{\psi})^2 + E(\psi - E\widehat{\psi})^2 = V(\widehat{\psi}) + bias^2(\widehat{\psi})$$

Since the transformation to ψ eliminates bias, it is almost trivial to claim that MSQ is minimised as well. This property of minimum MSQ can be demonstrated using the results of Hougaard (1982, section 3, p248).

[15] Equation (15), p310, with $I = -E\left(\partial^2 L/\partial \theta^2\right)$.

[16] The usage of O and the term *order* follow standard numerical analysis terminology, as defined in D. Kincaid and W. Cheney (1991, pp10–12). See also Amari (1990, p159) for a corresponding clarification of the term *order*.

A Taylor's series expansion

$$\widehat{\psi} = g(\widehat{\beta}) = g(\beta) + (\widehat{\beta} - \beta)g'(\beta) + (\widehat{\beta} - \beta)^2 g''(\beta)/2 + \ldots$$

forms the basis for calculating

$$E(\widehat{\psi}) = \psi + O(m^{-1})$$

to determine the order (m^{-1}) contribution to the bias. This is eliminated by choosing $\delta = 1$. Now

$$MSQ(\widehat{\psi}) = E(\widehat{\psi} - \psi)^2 = V(\widehat{\psi}) + bias^2(\widehat{\psi})$$

yielding

$$E(\widehat{\psi} - \psi)^2 = V(\widehat{\psi}) + m^{-2}[\ldots]^2$$

with the term inside $[\ldots]^2$ being zero to order (m^{-1}) if the transformation g produces the solution to Hougaard's (1982) equation with $\delta = 1$.

The assertion of asymptotic minimum MSQ and $O(m^{-1})$ unbiasedness is thus demonstrated.

2.10.2 Skewness Connection

$\underline{\alpha = -1/3}$: $(\delta = 2/3)$ (CEF)

The α–connection becomes

$$\overset{-1/3}{\Gamma}_{\psi} = (L_1 L_2) + \tfrac{2}{3} L_1^{(3)}$$

$$3\ \overset{-1/3}{\Gamma}_{\psi} = 3(L_1 L_2) + 2L_1^{(3)} = L_1^{(3)} - L_3\ .$$

Now $\kappa_3(\partial L/\partial\psi) = E\left(\partial L/\partial\psi - E(\partial L/\partial\psi)\right)^3 = E\left(\partial L/\partial\psi\right)^3 = L_1^{(3)}$, but the skewness for ψ is measured by $E(\widehat{\psi} - E\widehat{\psi})^3$ (Hougaard 1982, p248).

Amari (1990, p132 and p151) shows that this form of α–connection is directly related to the skewness of the parameter (ψ) and not the corresponding score statistic $(\partial L/\partial\psi)$.

Hougaard (1982, p248) gives a *direct* calculation in the one–parameter case showing that $\delta = 2/3$ $(\alpha = -1/3)$ produces $\kappa_3(\widehat{\psi}) = 0$ to order (m^{-1}) for a curved exponential family.

2.10.3 Information Connection

$\underline{\alpha = 0}$: $(\delta = 1/2)$ (GL)

The 0–connection is

$$2\,\overset{0}{\Gamma}_{\psi} = 2(L_1 L_2) + L_1^{(3)} = -{}_1L_2 \ .$$

So

$$\overset{0}{\Gamma}_{\psi} = 0 \Rightarrow {}_1L_2 = -\frac{\partial I}{\partial \psi} = -\frac{\partial}{\partial \psi} E(\frac{\partial L}{\partial \psi})^2 = 0 \ .$$

So the transformation g from β to ψ produces constant variance. Hougaard's (1982) derivation uses the variance of a transformed variable, viz, if $Y = g(X)$ then $Var(Y) = \sigma^2(X)\,[g'(\mu_x)]^2$. If ψ is to have constant variance after the transformation $\psi = g(\beta)$ then

$$g'(\beta) = \frac{d\psi}{d\beta} \propto \sqrt{J} \ , \quad J = i(\beta) \propto \frac{1}{\sigma^2(\beta)}$$

and the results of section 3 part 2 of Hougaard (1982) show that the key value of δ that produces the constant variance transformation is $\delta = 1/2$, ie. $\alpha = 0$.

2.10.4 'Normal' Connection

$\underline{\alpha = 1/3}$: $(\delta = 1/3)$ (GL)

The α–connection is

$$\overset{1/3}{\Gamma}_{\psi} = (L_1 L_2) + \tfrac{1}{3} L_1^{(3)} \ ,$$

to give

$$3\,\overset{1/3}{\Gamma}_{\psi} = 3(L_1 L_2) + L_1^{(3)} = -L_3 \ .$$

So $\overset{1/3}{\Gamma}_{\psi}= 0 \Rightarrow E(\partial^3 L/\partial\psi^3) = 0$, implying that the expected third derivative of the log likelihood is zero, ie., 'normal likelihood' in the terminology of Hougaard (1982, section 3, part 1, p247).

2.10.5 Exponential Connection

$\underline{\alpha = 1}$: $(\delta = 0)$ (CEF)

The 1–connection is

$$\overset{1}{\Gamma}_{\psi}= (L_1 L_2) ,$$

but since this corresponds to $\delta = 0$, the transformation is the identity, ie, the canonical parameterisation. In fact $\overset{1}{\Gamma}_{\psi}= 0$ simply means that the initial parameterisation β (from $\theta = \theta(\beta)$ in Hougaard's (1982) notation) may have the exponential family distribution in canonical form with respect to β, since $(L_1 L_2) = 0$ for a distribution of the exponential type, in terms of the canonical parameter.

If the distribution is of the exponential family type in the transformed parameter then $(L_1 L_2) = 0$ which leads to $\overset{1}{\Gamma}_{\psi}= 0$. Thus the transformed parameter is canonical.

However, if the transformation induces $\overset{1}{\Gamma}_{\psi}= 0$ this leads to $(L_1 L_2) = 0$, but it is not necessary in general for the family to be exponential if the 1–connection is zero, see Amari (1990, p152) and Appendix B.5.

2.10.6 Note

Clarification of the parameterizations in the above Section is called for. Wedderburn's formulation (Hougaard, 1982, p245) of the transformation $\psi(\theta)$

$$\psi(\theta_1) = \int_{\theta_0}^{\theta_1} \left\{ \frac{d^2}{d\theta^2} \ln \phi(\theta) \right\}^{\delta} d\theta \qquad (2.14)$$

is defined for the one–dimensional family in terms of the natural parameter θ, viz,

$$e^{\theta t(x)}/\phi(\theta).$$

A full description of Wedderburn's exponential form is given in Appendix B.6 with the cases $\delta = 0, 1/3, 1/2, 2/3$ and 1 being shown to have equivalent interpretations, albeit for the canonical (natural) parameter.

The one–dimensional submodel (curved family) is (Hougaard, 1982, p246)

$$e^{\theta(\beta)^\top t(x)}/\phi\{\theta(\beta)\}$$

The parameterization $\psi = g(\beta)$ for $\delta = 0$ now becomes the solution for g in Hougaard's (1982) equation

$$\frac{d^2g/d\beta^2}{dg/d\beta} = \left\{\delta\frac{d^3\chi}{d\theta^3}\left(\frac{d\theta}{d\beta}\right) + \left(\frac{d^2\theta}{d\beta^2}\right)'\frac{d^2\chi}{d\beta^2}\frac{d\theta}{d\beta}\right\} \Big/ \left\{\left(\frac{d\theta}{d\beta}\right)'\frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\}$$

This gives

$$\frac{d^2g/d\beta^2}{dg/d\beta} = \left\{\left(\frac{d^2\theta}{d\beta^2}\right)'\frac{d^2\chi}{d\beta^2}\frac{d\theta}{d\beta}\right\} \Big/ \left\{\left(\frac{d\theta}{d\beta}\right)'\frac{d^2\chi}{d\theta^2}\frac{d\theta}{d\beta}\right\}$$

but the canonical parameterization implies that $\psi = \theta = \beta$, so g is the identity, ie., $g = \beta$ since $\theta = \beta$. This yields

$$\frac{d^2g/d\beta^2}{dg/d\beta} = 0$$

and

$$\frac{d^2\theta}{d\beta^2} = 0$$

which trivially satisfies Hougaard's (1982, p246) equation.

2.10.7 Summary

Kass (1984) has demonstrated that the α–connections coincide only in the special case of the Normal distribution with known covariance[17]. This is confirmed by Hougaard (1982, p251) : "For the curved exponential family the four parameterizations are not identical in general, and you cannot get more than one of the properties"[18]. It follows that in the case of non–Normal errors the above special values of α will characterise the properties of the estimator. So, each property of

[17]Proposition 2, p90, Kass (1984).

[18]The case $\alpha = 1$ ($\delta = 0$) is excluded as it is the canonical parameterization (the identity!).

the estimator has to be considered separately using each key value of α, whereas for the Normal case the value of α is irrelevant. For example, if $\overset{0}{\Gamma}$ appears small, we might expect that the model will exhibit constant variance with respect to the parameter involved.

The interpretation given to the α–connections via Bartlett's equations is quite general and does not necessarily require the distribution to be of exponential type. So, interpretation of some α–connections could be made for *any* type of distribution, since $\overset{\alpha}{\Gamma}= 0$ implies special features to be associated with the choice of α, viz,

-1 (bias reduced), 0 (constant variance) and 1/3 ('normal likelihood').

Table 2.2 presents the values of α and the corresponding conditions under which the previous statistical interpretations of a zero α–connection can be made.

α	-1	-1/3	0	1/3	1
Condition	GL	CEF	GL	GL	CEF

Table 2.2: Conditions for the interpretation of α : single parameter.

Of course, the problem is to find a parameterization that will zero the α–connection. Such a parameter can be shown to exist in the one–dimensional case, (Amari, 1990, p152, Corollary).

Kass (1984) demonstrated the correspondence between parameterizations as determined from Hougaard's (1982) equation and Amari's (1982a) α–connections, whence the relation $\alpha = 1 - 2\delta$ is derived. By the use of Bartlett's (1953a) equations, the statistical interpretation of α–connections for key values of α has been demonstrated *directly*, corroborating the interpretations made by Kass (1984) in appealing to the effect of choosing the corresponding value of δ in Hougaard's (1982) equation.

Bartlett (1953b) has produced similar equations in the multi–parameter case, and similar results follow as a generalization of the one–parameter situation here described. These results are corroborated by Amari (1990, pp150–152), using an

entirely different approach.

2.11 Interpretation of α in the multi-parameter case

When more than one parameter is involved, the general form of α–connection is required, viz

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = E[\partial_i\partial_j\ell(\boldsymbol{y};\boldsymbol{\theta})\partial_k\ell(\boldsymbol{y};\boldsymbol{\theta})] + \frac{1-\alpha}{2}E[\partial_i\ell(\boldsymbol{y};\boldsymbol{\theta})\partial_j\ell(\boldsymbol{y};\boldsymbol{\theta})\partial_k\ell(\boldsymbol{y};\boldsymbol{\theta})] \,.$$

The same key values of α are associated with special properties of the estimators as outlined in Amari (1990, pp150–152). The interpretation in general depends on advanced analysis of estimator behaviour, which will be treated in later chapters. Some of the properties described using the one–dimensional approach via Bartlett's (1953a) equations carry over and so will be expounded. The generalization of these equations is given in Bartlett (1953b, pp306–307); again only the first three derivatives are necessary to produce the multi–dimensional analogue of the equations for one parameter models.

A modification of notation from the one–parameter case is necessary to enable similar manipulation of identities as used in the interpretation of α–connections for one–parameter models. So, the equations of Bartlett (1953b, pp306–307) are represented in the following *extended* notation

$$L_{\mathsf{a}} = E(\frac{\partial L}{\partial\theta_{\mathsf{a}}})$$

$$(L_{\mathsf{a}}L_{\mathsf{b}}) = E(\frac{\partial L}{\partial\theta_{\mathsf{a}}}\frac{\partial L}{\partial\theta_{\mathsf{b}}})$$

$$(L_{\mathsf{ab}}) = E(\frac{\partial^2 L}{\partial\theta_{\mathsf{a}}\partial\theta_{\mathsf{b}}})$$

$$(L_{\mathsf{a}}L_{\mathsf{b}}L_{\mathsf{c}}) = E(\frac{\partial L}{\partial\theta_{\mathsf{a}}}\frac{\partial L}{\partial\theta_{\mathsf{b}}}\frac{\partial L}{\partial\theta_{\mathsf{c}}})$$

$$(L_{\mathsf{abc}}) = E(\frac{\partial^3 L}{\partial\theta_{\mathsf{a}}\partial\theta_{\mathsf{b}}\partial\theta_{\mathsf{c}}})$$

$$(L_{\mathsf{a}}L_{\mathsf{bc}}) = E(\frac{\partial L}{\partial\theta_{\mathsf{a}}}\frac{\partial^2 L}{\partial\theta_{\mathsf{b}}\partial\theta_{\mathsf{c}}})$$

$${}_{\mathsf{a}}L_{\mathsf{bc}} = \frac{\partial}{\partial\theta_{\mathsf{a}}}[E(\frac{\partial^2 L}{\partial\theta_{\mathsf{b}}\partial\theta_{\mathsf{c}}})]$$

$$\vdots$$

Note that the *type* used for the subscripts is designed to allow different parameters to be addressed. In later developments, the subscripts ijk will be used to denote natural parameters in an exponential family model, eg., a GLM, while subscripts abc will denote regression coefficients contained within those natural parameters via the fitted values. The notation used here, viz, abc, is meant to show that either set of parameters can be intended, or even a set of transformed parameters, as in the one dimensional case.

Bartlett's (1953b) equations are (in full form)

$$E(\frac{\partial L}{\partial\theta_i}) = 0$$

$$E(\frac{\partial L}{\partial\theta_i}\frac{\partial L}{\partial\theta_j}) = -E(\frac{\partial^2 L}{\partial\theta_i\partial\theta_j})$$

$$E(\frac{\partial L}{\partial\theta_i}\frac{\partial L}{\partial\theta_j}\frac{\partial L}{\partial\theta_k}) = -E(\frac{\partial^3 L}{\partial\theta_i\partial\theta_j\partial\theta_k}) - E(\frac{\partial L}{\partial\theta_i}\frac{\partial^2 L}{\partial\theta_j\partial\theta_k}) - E(\frac{\partial L}{\partial\theta_j}\frac{\partial^2 L}{\partial\theta_i\partial\theta_k}) - E(\frac{\partial L}{\partial\theta_k}\frac{\partial^2 L}{\partial\theta_i\partial\theta_j})$$

where $\theta_i, \theta_j, \theta_k$ are *arbitrary* parameters.

An additional equation is generated by differentiating the second order equation. The resulting four equations can be combined to obtain another relation

used to simplify expressions involving key quantities such as the skewness tensor $(L_{\mathsf{a}}L_{\mathsf{b}}L_{\mathsf{c}})$. Using the extended notation defined earlier, these five equations (Bartlett, 1953b, pp306–307) now become

$$L_{\mathsf{a}} = 0$$

$$(L_{\mathsf{a}}L_{\mathsf{b}}) + (L_{\mathsf{ab}}) = 0$$

$$(L_{\mathsf{a}}L_{\mathsf{b}}L_{\mathsf{c}}) + (L_{\mathsf{abc}}) + (L_{\mathsf{a}}L_{\mathsf{bc}}) + (L_{\mathsf{b}}L_{\mathsf{ac}}) + (L_{\mathsf{c}}L_{\mathsf{ab}}) = 0$$

$${}_{\mathsf{a}}L_{\mathsf{bc}} + (L_{\mathsf{a}}L_{\mathsf{b}}L_{\mathsf{c}}) + (L_{\mathsf{ab}}L_{\mathsf{c}}) + (L_{\mathsf{ac}}L_{\mathsf{b}}) = 0$$

$$(L_{\mathsf{a}}L_{\mathsf{b}}L_{\mathsf{c}}) = 2(L_{\mathsf{ab}}L_{\mathsf{c}}) - {}_{\mathsf{a}}L_{\mathsf{bc}} - {}_{\mathsf{b}}L_{\mathsf{ac}} - {}_{\mathsf{c}}L_{\mathsf{ab}} \, .$$

Various adaptations are possible using different combinations of the indices abc.

The α–connection now becomes

$$\overset{\alpha}{\Gamma}_{\mathsf{abc}} = (L_{\mathsf{ab}}L_{\mathsf{c}}) + \frac{1-\alpha}{2}(L_{\mathsf{a}}L_{\mathsf{b}}L_{\mathsf{c}}) \, .$$

The key values of α have the same interpretation as in the single parameter case (Amari, 1990, pp150–152). The full derivation of some of these results depends on later developments, but all values of α are reported for completeness. The interpretations of the cases $\alpha = 0, 1/3$ and 1 can be derived directly using Bartlett's (1953b) equations.

Similarly to the one parameter case, the solution of the equation

$$\overset{\alpha}{\Gamma}_{\mathsf{abc}} = 0$$

will be in terms of transformed parameters in a curved exponential family. Thus the subscripts abc will refer to transformed parameters.

Again, the term CEF stands for 'Curved Exponential Family', and GL means 'General Likelihood'. These terms are also used in Table 2.3.

2.11.1 Mixture Connection

$\underline{\alpha = -1}$: CEF

The (-1)–connection is

$$\overset{-1}{\Gamma}_{\mathsf{abc}} \overset{\text{def}}{=} \overset{m}{\Gamma}_{\mathsf{abc}} = (L_{\mathsf{ab}}L_{\mathsf{c}}) + (L_{\mathsf{a}}L_{\mathsf{b}}L_{\mathsf{c}}) = (L_{\mathsf{abc}}) - {}_{\mathsf{a}}L_{\mathsf{bc}} - {}_{\mathsf{b}}L_{\mathsf{ac}} \ .$$

Together with other variations, this equation is a generalization of the connection from the one–parameter case examined earlier. The corresponding results for bias are given in Amari (1990, p131 and p150), while the minimum mean square error result is given in Amari (1990, p133 and p150), with both results being written in terms of the mixture connection $\left(\overset{m}{\Gamma}\right)$.

2.11.2 Skewness Connection

$\underline{\alpha = -1/3}$: CEF

The skewness connection is

$$\overset{-1/3}{\Gamma}_{\mathsf{abc}} = (L_{\mathsf{ab}}L_{\mathsf{c}}) + \frac{2}{3}(L_{\mathsf{a}}L_{\mathsf{b}}L_{\mathsf{c}}) \ .$$

The term $-3\overset{-1/3}{\Gamma}_{\mathsf{abc}}$ is shown to be precisely the third cumulant K_{abc} (Amari, 1990, p132), and so the parameterization that zeros the $(-1/3)$–connection produces zero asymptotic skewness.

2.11.3 Information Connection

$\underline{\alpha = 0}$: *GL*

The 0–connection is

$$\overset{0}{\Gamma}_{\mathsf{abc}} = (L_{\mathsf{ab}} L_{\mathsf{c}}) + \frac{1}{2}(L_{\mathsf{a}} L_{\mathsf{b}} L_{\mathsf{c}}) \, .$$

Thus

$$2 \ \overset{0}{\Gamma}_{\mathsf{abc}} = 2(L_{\mathsf{ab}} L_{\mathsf{c}}) + (L_{\mathsf{a}} L_{\mathsf{b}} L_{\mathsf{c}}) = \ {}_{\mathsf{c}}L_{\mathsf{ab}} - \ {}_{\mathsf{a}}L_{\mathsf{cb}} - \ {}_{\mathsf{b}}L_{\mathsf{ca}} \, .$$

Now if

$$2 \ \overset{0}{\Gamma}_{\mathsf{abc}} = 0$$

this implies that

$$\frac{\partial}{\partial \theta_{\mathsf{c}}}(E \frac{\partial^2 L}{\partial \theta_{\mathsf{a}} \partial \theta_{\mathsf{b}}}) = \frac{\partial}{\partial \theta_{\mathsf{a}}}(E \frac{\partial^2 L}{\partial \theta_{\mathsf{c}} \partial \theta_{\mathsf{b}}}) + \frac{\partial}{\partial \theta_{\mathsf{b}}}(E \frac{\partial^2 L}{\partial \theta_{\mathsf{c}} \partial \theta_{\mathsf{a}}})$$

Using the other combinations of a, b and c yields similar equations, viz

$$a = b + c$$

$$c = a + b$$

$$b = a + c$$

where

$$a = \ {}_{\mathsf{c}}L_{\mathsf{ab}}, \ b = \ {}_{\mathsf{a}}L_{\mathsf{cb}}, \ c = \ {}_{\mathsf{b}}L_{\mathsf{ca}}.$$

The only solution is

$$a = b = c = 0$$

which implies that the matrix

$$E(\frac{\partial^2 L}{\partial \theta_{\mathsf{a}} \partial \theta_{\mathsf{b}}})$$

is constant, for arbitrary a and b. So a transformation which zeros the 0–connection produces constant (co)–variance with respect to the new parameterization. Note

that this result does not require that the likelihood be derived from a curved exponential family necessarily, nor that it be exponential in the transformed parameter.

2.11.4 'Normal' Connection

$\underline{\alpha = 1/3}$ *GL*

The normal connection is

$$\overset{1/3}{\Gamma}_{abc} = (L_{ab}L_c) + \frac{1}{3}(L_aL_bL_c) .$$

Cycling the indices yields

$$\overset{1/3}{\Gamma}_{bac} = (L_{ba}L_c) + \frac{1}{3}(L_bL_aL_c)$$

and

$$\overset{1/3}{\Gamma}_{cab} = (L_{ca}L_b) + \frac{1}{3}(L_cL_aL_b) .$$

Summing these three ($= \Sigma$), implies that if the (1/3)-connections are zeroed, then

$$\Sigma = 0 = (L_{ab}L_c) + (L_{ba}L_c) + (L_{ca}L_b) + (L_aL_bL_c) = -(L_{abc}) .$$

Invoking the third equation of Bartlett (1953b), viz

$$-(L_aL_bL_c) = (L_{abc}) + (L_aL_{bc}) + (L_bL_{ac}) + (L_cL_{ab}) ,$$

means that if $\overset{1/3}{\Gamma}_{abc} = 0$ then $(L_{abc}) = 0$, to give

$$E(\frac{\partial^3 L}{\partial\theta_a\partial\theta_b\partial\theta_c}) = 0 .$$

Thus, the matrix of expected third derivatives of the log–likelihood is zero, producing 'normal' likelihood in the terminology of Hougaard (1982).

This result holds for *general* likelihoods, not just for the transformed parameters of curved exponential families.

2.11.5 Exponential Connection

$\underline{\alpha = 1}$ *CEF*

If the exponential family distribution is canonical in the transformed parameter, then

$$(L_{\mathsf{ab}}L_{\mathsf{c}}) = 0 \rightarrow \overset{1}{\Gamma}_{\mathsf{abc}} = 0$$

If the transformation induces $\overset{1}{\Gamma}_{\mathsf{abc}} = 0$ then $(L_{\mathsf{ab}}L_{\mathsf{c}}) = 0$, but it is not necessarily true that the family be (curved) exponential, as shown in Appendix B.5 and Amari (1990, p152). The same comments applied to the one–parameter case.

Comments

It should be noted that in line with the one–parameter case, the cases $\alpha = 0$ and $\alpha = 1/3$ produce results that hold for general likelihoods.

Table 2.3 presents the values of α and the corresponding conditions under which the previous statistical interpretations of a zero α–connection can be made.

α	-1	-1/3	0	1/3	1
Condition	CEF	CEF	GL	GL	CEF

Table 2.3: Conditions for the interpretation of α : multi–parameter.

Of course, the problem is to find a parameterization that will zero the α–connection. In general, such parameters need not exist. However, *local* parameterizations can be defined that satisfy the conditions required in a small neighbourhood of specific values. Thus, any of the given conditions can be fulfilled locally by a particular point; see Amari (1990, pp150–152).

For example, using Normal errors and a nonlinear response (nonlinear regression), a Taylor's expansion can produce an approximating model that is linear in the parameters with

$$\overset{1}{\Gamma}_{abc} = 0 \, .$$

So, in a small neighbourhood of the expansion the parameters are locally canonical (linear). This linearity forms the basis of iterative solutions to the fitting of nonlinear regression models to data.

2.12 Dual Space

For the general exponential family defined by

$$\ell = \theta^i y_i - \psi(\boldsymbol{\theta}) + c(\boldsymbol{y})$$

the parameter space defined by $\boldsymbol{\theta}$ is called the *natural* parameter space.

The *dual* space contains the space of expectations; see Amari (1982a, p366). This is the space used by Bates and Watts (1980) in examining the differential geometry of the nonlinear regression model. The expectation surface is characterised by

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$$

with

$$\xi_i = E_\theta(y_i) = \partial_i \psi(\boldsymbol{\theta}).$$

There is a 1–to–1 correspondence between $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. The space of expectations is exactly the space of fitted values, and is of importance since it is the coordinate system in which the Cramer–Rao bound is attained. It has the further property of causing the mixture (-1)–connection to vanish in the same way that the exponential (1)–connection vanishes for the natural coordinate system.

2.13 Generalized Linear Models

The essential consideration in examining generalized linear models (GLMs) is the extension from Normality to statistical distributions of other types. Much effort and ingenuity has been expended on Normal linear theory, and the success of GLMs has been partly due to many results of Normal linear theory and applications being subsumed in the theory of generalized linear models. Two such examples are the

implementation of the GLIM fitting algorithm, and the analysis of deviance of sequentially fitted nested models of increasing complexity.

The differential geometric approach offers a vehicle for expanding the results of Normal theory into the non–normal structures of GLMs. Consequently, frequent reference to the Normal case will be made as a special example of a GLM.

For a GLM, the contribution to the log–likelihood ℓ for a single observation is, using the notation of McCullagh and Nelder (1989),

$$\ln f(y;\theta) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi).$$

A general exponential family is represented by

$$\ell = c(\boldsymbol{y}) + \theta^i y_i - \psi(\boldsymbol{\theta})$$

yielding

$$\partial_k \ell = y_k - \partial_k \psi$$

and

$$\partial_i \partial_j \ell = -\partial_i \partial_j \psi = -g_{ij}$$

So, taking expectations gives

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = \frac{1-\alpha}{2} E(\partial_i \ell \partial_j \ell \partial_k \ell)$$

The skewness tensor T_{ijk} can be written as

$$T_{ijk} \overset{\text{def}}{=} E(\partial_i \ell \partial_j \ell \partial_k \ell) = -E(\partial_i \partial_j \partial_k \ell). \tag{2.15}$$

from Bartlett (1953b, p306), and Amari (1982a, p365).

For a GLM with unit scale parameter, $(a(\phi) = 1)$,

$$g_{ij} = b''(\theta^I)\delta_{ij} \tag{2.16}$$

giving[19]

[19] δ_{ij} is the Kronecker delta, viz

$$\begin{aligned} \delta_{ij} &= 1, i = j \\ &= 0, i \neq j\ . \end{aligned}$$

$$T_{ijk} = b'''(\theta^K)E_{ijk} \tag{2.17}$$

where E_{ijk} is defined by

$$E_{ijk} = 1, \text{ if } i = j = k,$$

$$= 0 \text{ else}$$

giving

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = \frac{1-\alpha}{2}T_{ijk}, \tag{2.18}$$

where I and K are nonsum indices, as described in Section 1.8.2. However, Equation (2.18) is for the *canonical* parameter $\boldsymbol{\theta}$. The only GLMs for which these canonical parameters are of specific interest are simple analysis of variance models, ie., models using categorical predictors only[20]. Usually a transformation to the parameters of interest ($\boldsymbol{\beta}$) will be required, assuming $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\beta})$, via $\boldsymbol{\eta} = \boldsymbol{X\beta}$ and the link function g defined by $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. There are various equivalent forms for the linear predictor $\boldsymbol{\eta}$, viz,

$$\boldsymbol{\eta} = \boldsymbol{X\beta}$$

and

$$\eta_i = g(\mu_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta} = \sum_j X_{ij}\beta_j = X_{ij}\beta^j.$$

Any of these forms may be used to describe the linear predictor.

Note

The definition of E_{ijk} corresponds to the generalized Kronecker delta $\delta_{i,j,k}$ of Amari (1990, p43). See also McCullagh and Nelder (1983, p237, Appendix D, A.21a) for an equivalent definition of the skewness, allowing for differences due to the scale parameter.

[20]The choice of link function is irrelevant as the fitted values are effectively the parameters.

2.13.1 One–dimensional GLMS

The application of this differential geometric technique to the generalized linear models of Nelder and Wedderburn (1972), in the one dimensional case, is now investigated.

In the trivial case of a null model, $\theta_i = \theta \; \forall \, i$ so

$$\overset{\alpha}{\Gamma}_{\theta} = \frac{1-\alpha}{2} E\left(\frac{\partial \ell}{\partial \theta}\right)^3 = \frac{1-\alpha}{2} E\left(-\frac{\partial^3 \ell}{\partial \theta^3}\right) = \frac{1-\alpha}{2}\kappa_3 = \frac{1-\alpha}{2}\frac{b'''(\theta)}{a(\phi)}$$

using the results of the previous section, or Bartlett (1953a, p13). For example, given Normal errors

$$\overset{\alpha}{\Gamma}_{\theta} = 0 \;\; \forall \, \theta,$$

since

$$\theta = \mu, \;\; b(\theta) = \theta^2/2 \;\Rightarrow b'''(\theta) = 0.$$

This give zero skewness, as is expected for a symmetric distribution.

The one–dimensional GLM of interest is of course one for which the single regression parameter (β) produces a 'line of means' $\mu_i = b'(\theta_i)$ via the link function

$$g(\mu_i) = \boldsymbol{X_i}^{\top} \beta = \eta_i.$$

Such a model is a special case of a *curved exponential family*, in the terminology of Efron (1975).

2.14 Regression coefficients in GLMs

For a curved exponential family imbedded in a multi dimensional parameter space, the regression coefficients of interest ($\boldsymbol{\beta}$) are imbedded in the space of natural parameters ($\boldsymbol{\theta}$). For a GLM, the relation describing this imbedding is

$$\theta^i = \mathsf{f}(X_{ij}\beta^j).$$

A full notational description of this relationship and the role played by the linear predictor is given in Appendix B.7. Amari (1982a, Theorem 3, 4.6, p370) gives

the mechanism for writing $\overset{\alpha}{\Gamma}_{\boldsymbol{\beta}}$ in terms of $\overset{\alpha}{\Gamma}_{\boldsymbol{\vartheta}}$, $\boldsymbol{\vartheta}$ being the natural parameters (coordinates).[21] This result will be referred to as the 'imbedding theorem', as it will be cited frequently.[22]

2.14.1 Imbedding

A set of m regression coefficients $(\boldsymbol{u})$ is contained within the set of n natural parameters $(\boldsymbol{\vartheta})$ where $m < n$. So $\vartheta^i = \vartheta^i(\boldsymbol{u})$ [$\theta^i = \theta^i(\boldsymbol{\beta})$ for a GLM]. Now

$$f(\boldsymbol{y};\boldsymbol{u}) = f(\boldsymbol{y};\boldsymbol{\vartheta}(\boldsymbol{u}))$$

and

$$\partial_a \overset{\text{def}}{=} \frac{\partial}{\partial u^a}, \quad a = 1, 2, \ldots, m$$

giving

$$\partial_a \ell(\boldsymbol{y};\boldsymbol{u}) = \mathrm{B}^i_a(\boldsymbol{u}) \partial_i \ell(\boldsymbol{y};\boldsymbol{\vartheta}(\boldsymbol{u}))$$

where

$$\mathrm{B}^i_a(\boldsymbol{u}) = \frac{\partial \vartheta^i}{\partial u^a} .$$

Subscripts abc will be associated with $\boldsymbol{u}$, ijk with $\boldsymbol{\vartheta}$. Corresponding GLM relations can be written for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, using $\boldsymbol{u} = \boldsymbol{\beta}$ and $\boldsymbol{\theta} = a(\phi)\boldsymbol{\vartheta}$.

2.14.2 Imbedding Theorem

The expression of the α–connection $(\overset{\alpha}{\Gamma}_{abc})$ for the regression coefficients in terms of the α–connection $(\overset{\alpha}{\Gamma}_{ijk})$ of the natural parameters (or coordinates), is given by

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = (\partial_a \mathrm{B}^i_b)\mathrm{B}^j_c g_{ij}(\boldsymbol{\vartheta}(\boldsymbol{u})) + \mathrm{B}^i_a \mathrm{B}^j_b \mathrm{B}^k_c \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\vartheta}(\boldsymbol{u})). \tag{2.19}$$

A derivation of this relation is given in Appendix B.8. For a GLM, this relation becomes

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \left(\partial_a B^i_b\right) B^j_c g_{ij}(\boldsymbol{\theta}(\boldsymbol{\beta})) + B^i_a B^j_b B^k_c \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}(\boldsymbol{\beta})) \tag{2.20}$$

[21] The variable ϑ will be used for the natural parameter in Amari's (1982a) formulation when ambiguity with the canonical GLM parameter θ arises.

[22] Kass (1984, p89) uses the term 'inheritance' relation.

as $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ only differ by the scale parameter $a(\phi)$, since $\boldsymbol{\vartheta} = \boldsymbol{\theta}/a(\phi)$. The terms B_b^i in Equation (2.20) are defined for $\boldsymbol{\theta}$ so that

$$B_b^i = \frac{\partial \theta^i}{\partial \beta^b},$$

whereas the terms B_b^i in Equation (2.19) are defined for $\boldsymbol{\vartheta}$ so that

$$\mathrm{B}_b^i = \frac{\partial \vartheta^i}{\partial \beta^b}$$

as given in Section 2.14.1.

To reconcile Equation (2.20) and Equation (2.19), the expression for an affine connection in one coordinate system in terms of the affine connection in another coordinate system is invoked (Amari, 1982a, 2.28, p364).[23] The transformation equation for the α–connection (Amari, 1982a, 2.28, p364) is rearranged to give Equation (2.21). The coordinate transformation defined by Amari was

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta})$$

with the prime indices being associated with $\boldsymbol{\eta}$. The transformation equation (2.28) of Amari (1982a) was

$$\Gamma_{i'j'k'} = \left(\partial_{i'} B_{j'}^m\right) B_{k'}^l g_{lm} + B_{i'}^l B_{j'}^m B_{k'}^n \Gamma_{lmn} \tag{2.21}$$

where

$$B_{i'}^l = \frac{\partial \theta^l}{\partial \eta^{i'}}.$$

The coordinate transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\vartheta}$ implies corresponding indices lmn to $i'j'k'$, in the notation of Amari (1982a, p364). The use of the prime indicates that the transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\vartheta}$ is 1:1. The 1:1 transformation involved in $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(\boldsymbol{\theta})$ is different to the imbedding relation $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(\boldsymbol{\beta})$, and so Equation (2.21) is different to Equation (2.19), even though their forms are similar. For a GLM

$$\boldsymbol{\theta} = a(\phi)\boldsymbol{\vartheta}$$

[23] The trailing term in equation (2.28) of Amari (1982a, p364) *pre-empts* this study.

giving

$$B^l_{i'} = \frac{\partial \theta^l}{\partial \vartheta^{i'}} = a(\phi) \ .$$

This implies that

$$\partial_{i'} B^m_{j'} = 0$$

giving the α–connection in terms of ϑ as

$$\Gamma_{i'j'k'}(\boldsymbol{\vartheta}) = B^l_{i'} B^m_{j'} B^n_{k'} \Gamma_{lmn} = a^3(\phi) \Gamma_{lmn}(\boldsymbol{\theta}).$$

Metric

The metric g_{ab} for the regression coefficients ($\boldsymbol{u} \equiv \boldsymbol{\beta}$) in terms of the metric g_{ij} for the natural parameters $\boldsymbol{\vartheta}$ becomes

$$g_{ab}(\boldsymbol{u}) = \mathrm{B}^i_a \mathrm{B}^j_b g_{ij}(\boldsymbol{\vartheta}(\boldsymbol{u})). \tag{2.22}$$

The derivation follows from the definition of the metric, namely

$$g_{ab}(\boldsymbol{\beta}) = E\left[\partial_a \ell \partial_b \ell\right] = E\left[\mathrm{B}^i_a \partial_i \ell \mathrm{B}^j_b \partial_j \ell\right] = \mathrm{B}^i_a \mathrm{B}^j_b E\left[\partial_i \ell \partial_j \ell\right] = \mathrm{B}^i_a \mathrm{B}^j_b g_{ij}(\boldsymbol{\vartheta}).$$

This result will be used later in conjunction with the 'imbedding' theorem. The metric for $\boldsymbol{\vartheta}$ expressed as a function of the metric for $\boldsymbol{\theta}$ is

$$g_{i'j'}(\boldsymbol{\vartheta}) = B^k_{i'} B^l_{j'} g_{kl}(\boldsymbol{\theta})$$

where the prime again denotes the $\boldsymbol{\vartheta}$ coordinate system. Since

$$B^k_{i'} = \frac{\partial \theta^k}{\partial \vartheta^{i'}} = a(\phi)$$

this gives

$$g_{i'j'}(\boldsymbol{\vartheta}) = a^2(\phi) g_{kl}(\boldsymbol{\theta}).$$

Scale Parameter

Gathering results for the α–connection and the metric gives

$$\Gamma_{ijk}(\boldsymbol{\vartheta}) = a^3(\phi) \Gamma_{ijk}(\boldsymbol{\theta})$$

and

$$g_{ij}(\boldsymbol{\vartheta}) = a^2(\phi) g_{ij}(\boldsymbol{\theta}). \tag{2.23}$$

To convert Equation (2.19) to Equation (2.20), it is necessary to convert from $\boldsymbol{\vartheta}$ to $\boldsymbol{\theta}$. In Equation (2.19), B_b^i is

$$\mathrm{B}_b^i = \frac{\partial \vartheta^i}{\partial \beta^b} = \frac{\partial \vartheta^i}{\partial \theta^I} \frac{\partial \theta^I}{\partial \beta^b} = \frac{1}{a(\phi)} B_b^i$$

where the index I is nonsum taking the same value as the index i. Continued use of this chain rule leads to cancellation of all generated terms in $a(\phi)$, leading to Equation (2.20).

The interpretation of this phenomenon is that the scale parameter affects the α–connection for the natural parameters (Γ_{ijk}), but not the α–connection for the regression coefficients (Γ_{abc}). This is due to Γ_{ijk} being a function of derivatives of the likelihood with respect to the natural parameters, while Γ_{abc} is related to derivatives of the likelihood with respect to the regression coefficients, and so Γ_{abc} is independent of changes of scale in the natural parameters.

2.14.3 Normal Distribution

For the Normal distribution,

$$\theta^i = \mu^i \ , \ b(\theta^i) = (\theta^i)^2/2 \ , \ \rightarrow b''(\theta^i) = 1 \ , \ \rightarrow b'''(\theta^i) = 0 \ .$$

This gives the α–connection for the natural parameters as

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) \ = \ 0,$$

since

$$\overset{\alpha}{\Gamma}(\boldsymbol{\theta}) \ \propto \ b'''(\boldsymbol{\theta})$$

as given in Equation (2.17). Reworking Equation (2.23), and substituting $a(\phi) = \sigma^2$ gives

$$g_{ij}(\boldsymbol{\theta}) = \frac{1}{a^2(\phi)} g_{ij}(\boldsymbol{\vartheta}) = \frac{1}{\sigma^4} g_{ij}(\boldsymbol{\vartheta}) = \frac{1}{\sigma^4}(\sigma^2 \delta_{ij}) = \frac{\delta_{ij}}{\sigma^2}.$$

So the α–connection for the regression coefficients becomes

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \left(\partial_a B_b^i\right) B_c^j g_{ij} = \left(\partial_a B_b^i\right) B_c^j \delta_{ij}/\sigma^2,$$

as in Section 3.5.1.

2.14.4 Normal Linear Models

For these models, the link function is the identity and the errors are Normal giving

$$\theta^i = \mu^i = X_{ij}\beta^j .$$

In this case the distribution is symmetric, so the skewness tensor T_{ijk} vanishes[24]. This makes $\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = 0$, as shown in Section 2.14.3. Since a linear model implies a canonical link for the Normal distribution,

$$B_b^i = \frac{\partial \theta^i}{\partial \beta^b} = \frac{\partial \mu^i}{\partial \beta^b} = X_{ij}$$

to give

$$\partial_a B_b^i = 0$$

which gives

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = 0.$$

So all the α–connections for $\boldsymbol{\beta}$ vanish, showing that all the conditions associated with key values of α such as unbiasedness, zero skewness, constant variance and 'normal' likelihood hold for the Normal linear model without the need for transformation, as stated in Hougaard (1982, p249).

2.14.5 Nonlinear Regression

For the nonlinear regression problem, the mean is a nonlinear function of the predictors, giving

$$\mu^i = \theta^i = \mathsf{f}(\boldsymbol{X}_i; \boldsymbol{\beta})$$

[24] $\partial_i \partial_j \ell = -\delta_{ij} \Rightarrow \partial_i \partial_j \partial_k \ell = 0.$

and

$$B_b^i = \frac{\partial \theta^i}{\partial \beta^b} = \frac{\partial \mathsf{f}(\boldsymbol{X}_i;\boldsymbol{\beta})}{\partial \beta^b} \stackrel{\text{def}}{=} \mathsf{f}'_{ib}.$$

So

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \left(\partial_a B_b^i\right) B_c^j g_{ij} = (\partial_a \mathsf{f}'_{ib})\mathsf{f}'_{jc}\delta_{ij}/\sigma^2$$

even though $\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = 0$, due to the distribution being Normal. As in the Normal linear case, all the α–connections are identical, since $\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta})$ is not related to α. This of course means that any transformation which zeros one of the α–connections zeros them for all α. So all the properties corresponding to special values of α can all be satisfied simultaneously, as shown by Hougaard (1982, p246), for one–parameter models. As outlined by Kass (1984), these properties include unbiasedness, stability of variance, lack of skewness and normality of likelihood (zero expected third derivative of the log–likelihood).

This simultaneity is a special property of the Normal distribution, which does not necessarily hold for general non–normal errors, since $\overset{\alpha}{\Gamma}_{ijk} \neq 0$ in general. For corroboration see Amari (1990, p156), and Kass (1984, p90, Proposition 2).

Note

The class of models defined by Wei (1994) are related asymptotically to nonlinear regression models. This is due to the constraint imposed by the regularity condition [Wei, 1994, p328, (a)], viz,

$$\|L^{(3)}_{abc}(\theta)\| < M(\theta)/n$$

with

$$EM(\theta) < K$$

which imply that the skewness tensor vanishes asymptotically.[25] As demonstrated by Kass (1984), the only family of distributions for which this occurs are Normal with known (co)–variance. Hence it would appear that 'close to Normal'

[25] In Amari's notation $T_{ijk} \to 0$.

families only are being addressed by this class of model. Since asymptotic Normality is implied, the skewness tensor T_{ijk} vanishes asymptotically and a 'common' affine connection is implied, i.e., $\overset{\alpha}{\Gamma}_{abc}$ is independent of α. So all the properties of nonlinear regression models will be inherited asymptotically, see Kass (1984) and Amari(1990). Hence the question of which connection to use is irrelevant (asymptotically).

2.14.6 Generalized Linear Models

A feature of Generalized Linear Models is the *factorisation* that occurs in B, viz,

$$B_b^i = \frac{\partial \theta^i}{\partial \beta^b} = \frac{\partial \mathsf{f}(\boldsymbol{X}_i^\top \boldsymbol{\beta})}{\partial \beta^b} = \frac{\partial \mathsf{f}(X_{ij}\beta^j)}{\partial \beta^b} = \mathsf{f}'_{ib}$$

and so

$$B_b^i = \frac{\partial \mathsf{f}_i}{\partial \eta_I}\frac{\partial \eta_I}{\partial \beta^b} = \frac{\partial \mathsf{f}_i}{\partial \eta_I} X_{Ib}.$$

The factorisation can be expressed as

$$\mathsf{f}'_{ib} = \frac{\partial \mathsf{f}_i}{\partial \eta_I} X_{Ib}.$$

and is related to simplifications used in the fitting procedure of the GLIM algorithm. To elaborate, fitting general models of the form

$$E(\boldsymbol{Y}) = \boldsymbol{f}(\boldsymbol{\theta})$$

to non–Normal data, as reported in Seber and Wild (1989, p34), can be effected via the IRLS (Iteratively Reweighted Least Squares) algorithm due to Green (1984). At the core of this method is the factorisation

$$\frac{\partial \ell}{\partial \theta^i} = \frac{\partial \ell}{\partial f_j}\frac{\partial f_j}{\partial \theta^i}\,.$$

For the GLIM algorithm, the analogous development is to factorise the derivative of the log–likelihood as

$$\frac{\partial \ell}{\partial \beta^i} = \frac{\partial \ell}{\partial \eta^j}\frac{\partial \eta^j}{\partial \beta^i} = \frac{\partial \ell}{\partial \eta^j} X_{ij}$$

using the notation of McCullagh and Nelder (1989, p41). Further simplification of $\partial\ell/\partial\eta$ leads to the form of estimating equation peculiar to the GLIM algorithm. In particular, for canonical links ($\theta = \eta$)

$$\frac{\partial \ell}{\partial \eta^j} = \frac{\partial \ell}{\partial \theta^j}\frac{d\theta^J}{d\mu^J}\frac{d\mu^J}{d\eta^J} = \frac{\partial \ell}{\partial \theta^j}$$

leading to a simplification in the Hessian matrix (matrix of expected second derivatives of the log–likelihood with respect to the regression coefficients).

The α–connection for the regression coefficients $\boldsymbol{\beta}$ is

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \left(\partial_a B_b^i\right) B_c^j g_{ij} + B_a^i B_b^j B_c^k \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}).$$

The application of the above factorisation to this general form of the α–connection for a GLM is further developed in Section 2.15.7.

The partitioning of model/link effects and distribution effects for a GLM adds a layer of complexity on top of that experienced for Normal errors. The error distribution affects Γ_{abc} via Γ_{ijk} and the metric g_{ij}. The model/link function affects the terms B_b^i only. Thus, both terms in the α–connection are affected by the error distribution and the form of model as determined by the link function. Since the first term can be zeroed by choosing a canonical link, it is termed 'model' dependent, whereas the second term is deemed to be 'error' dependent since it can be zeroed by choosing Normal errors. So the situation for GLMs is an extension of Normal error models where intrinsic and parameter–effect curvatures are model dependent, but the distributional contribution Γ_{ijk} is zero.

2.14.7 Canonical Links

For a GLM, the link being canonical implies that

$$\theta^i = \eta^i = X_{ij}\beta^j$$

so

$$B_b^i = \frac{\partial \theta^i}{\partial u^b} = \frac{\partial \eta^i}{\partial \beta^b} = \frac{\partial X_{ij}\beta^j}{\partial \beta^b} = X_{ib}.$$

This implies that

$$\partial_a B_b^i = 0,$$

giving

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = B_a^i B_b^j B_c^k \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = X_{ia} X_{jb} X_{kc} \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta})$$

as the expression for the α–connection for the regression coefficients.

Notes

- As pointed out by Kass (1984), the 'inheritance' relation, Equation (2.19), is true not only for the exponential family of distributions, but also for general likelihood functions.
- The first component in Equation (2.20) is model dependent, since it can be zeroed by choosing a canonical link.
- The second component in Equation (2.20) is distribution dependent, since it can be zeroed by choosing Normal errors.
- A related but not identical relation to Equation (2.19) has been derived independently by Kass (1984) for the reparameterization of one–dimensional models. In this case the transformation is one–to–one, but as can be seen in Appendix B.9, a similar relationship to Amari's (1982a) imbedding theorem is obtained.

2.14.8 Summary

The general form of the α–connection for the regression coefficients is given by

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \left(\partial_a B_b^i\right) B_c^j g_{ij} + B_a^i B_b^j B_c^k \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}).$$

The form of this α–connection Γ_{abc} is given below for several special cases.

Normal errors (nonlinear regression)

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \left(\partial_a B_b^i\right) B_c^j g_{ij} = \left(\partial_a B_b^i\right) B_c^j \delta_{ij}/\sigma^2 + 0$$

Normal errors (linear model)

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = 0$$

GLM

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \left(\partial_a B_b^i\right) B_c^j b''(\theta^I)\delta_{ij}/a(\phi) + \frac{1-\alpha}{2} B_a^i B_b^j B_c^k b'''(\theta^K) E_{ijk}/a(\phi).$$

GLM (canonical link)

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = 0 + B_a^i B_b^j B_c^k \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = \frac{1-\alpha}{2} B_a^i B_b^j B_c^k b'''(\theta^K) E_{ijk}/a(\phi).$$

So, both terms in the expression for Γ_{abc} contain expressions such as B_b^i which are affected by the model parameterization as determined by the link function. The first term can be called 'model' dependent and the second can be termed 'error' dependent, due to the conditions that cause them to be zero.

2.15 Exponential Connection and GLMs

2.15.1 Theorem

The exponential connection in terms of the regression coefficients is zero if and only if the link function is canonical for a generalized linear model.

2.15.2 Preliminaries

The proof is almost trivial in the sense that a canonical link for a GLM implies exponentiality with respect to the regression coefficients[26]. Thus the exponential connection associated with the regression coefficients must be zero, since the distributional form is then of the exponential type with respect to the regression coefficients. Following the notation of Amari (1982a) for curved exponential families defined by the distribution function

$$f(\boldsymbol{y}; \boldsymbol{\theta}(\boldsymbol{u}))$$

[26] A canonical link also implies sufficiency.

with[27]

$$\theta^i = \mathsf{f}(X_{ij}\beta^j)$$

for a GLM, it follows that $\boldsymbol{u} = \boldsymbol{\beta}$.

Subscripts abc will be used for the regression coefficients $\boldsymbol{u}$,

and subscripts ijk will be used for the natural parameters $\boldsymbol{\theta}$.

Connections can be defined in terms of the *regression* coefficients, viz

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = E(\partial_a\partial_b\ell\partial_c\ell) + \delta E(\partial_a\ell\partial_b\ell\partial_c\ell)$$

$$\overset{e}{\Gamma}_{abc}(\boldsymbol{u}) = E(\partial_a\partial_b\ell\partial_c\ell)$$

$$\overset{m}{\Gamma}_{abc}(\boldsymbol{u}) = \overset{e}{\Gamma}_{abc}(\boldsymbol{u}) + E(\partial_a\ell\partial_b\ell\partial_c\ell)$$

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = \delta\ \overset{m}{\Gamma}_{abc}(\boldsymbol{u}) + (1-\delta)\ \overset{e}{\Gamma}_{abc}(\boldsymbol{u}).$$

2.15.3 Proposition

The precise mathematical statement of the theorem can now be given as

1. For a GLM with a canonical link, $\overset{e}{\Gamma}_{abc}(\boldsymbol{u}) = 0$.
2. If $\overset{e}{\Gamma}_{abc}(\boldsymbol{u}) = 0$ for a GLM, then the link function is canonical.

2.15.4 Proof

1. In terms of the natural parameters $\boldsymbol{\theta}$,

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = E(\partial_i\partial_j\ell\partial_k\ell) + \delta E(\partial_i\ell\partial_j\ell\partial_k\ell)$$

with

$$\overset{e}{\Gamma}_{ijk}(\boldsymbol{\theta}) = \overset{1}{\Gamma}_{ijk}(\boldsymbol{\theta}) = E(\partial_i\partial_j\ell\partial_k\ell).$$

For the exponential family

$$\ell = c(\boldsymbol{y}) + \theta^i y_i - \psi(\boldsymbol{\theta})$$

[27]Using the notation Appendix B.7.

$$\overset{e}{\Gamma}_{ijk}(\boldsymbol{\theta}) = E(\partial_i\partial_j\ell\partial_k\ell) = E(-\partial_i\partial_j\psi\partial_k\ell) = E(-g_{ij}\partial_k\ell) = 0$$

by virtue of the score statistic. From the imbedding theorem (Amari 1982a)

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = (\partial_a B^i_b)B^j_c g_{ij} + B^i_a B^j_b B^k_c \ \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}(\boldsymbol{u}))$$

giving for $\alpha = 1$ (the exponential connection)[28]

$$\overset{e}{\Gamma}_{abc}(\boldsymbol{u}) = (\partial_a B^i_b)B^j_c g_{ij} + B^i_a B^j_b B^k_c \ \overset{e}{\Gamma}_{ijk}(\boldsymbol{\theta}(\boldsymbol{u})).$$

Now

$$\overset{e}{\Gamma}_{ijk}(\boldsymbol{\theta}) = 0$$

for exponential family models, giving

$$\overset{e}{\Gamma}_{abc}(\boldsymbol{u}) = (\partial_a B^i_b)B^j_c g_{ij}.$$

Since the link is canonical,

$$B^i_b = \frac{\partial\theta^i}{\partial u^b} = \frac{\partial\eta^i}{\partial\beta^b} = X_{ib}$$

giving

$$\partial_a B^i_b = 0.$$

Thus

$$\overset{e}{\Gamma}_{abc}(\boldsymbol{u}) = 0$$

as expected.

2. If the exponential connection with respect to $\boldsymbol{\beta}$ vanishes, then

$$\overset{e}{\Gamma}_{abc}(\boldsymbol{u}) = 0$$

implying that

$$(\partial_a B^i_b)B^j_c g_{ij} + B^i_a B^j_b B^k_c \ \overset{e}{\Gamma}_{ijk}(\boldsymbol{\theta}) = 0.$$

However

$$\overset{e}{\Gamma}_{ijk}(\boldsymbol{\theta}) = 0$$

[28]It is assumed that the scale parameter $a(\phi) = 1$.

since the family is exponential with respect to $\boldsymbol{\theta}$,

$$(\partial_a B_b^i) B_c^j g_{ij} = 0.$$

Excluding trivial models, this implies that

$$\partial_a B_b^i = 0$$

ie., B_b^i is constant. Now[29]

$$B_b^i = \frac{\partial \mathsf{f}\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}\right)}{\partial \beta^b} = \frac{\partial \mathsf{f}_i}{\partial \eta_I} \frac{\partial \eta_I}{\partial \beta^b},$$

so

$$\frac{\partial \mathsf{f}_i}{\partial \eta_I} X_{Ib}$$

is constant[30].

This implies $\mathsf{f} \propto \eta$ which defines f as a canonical link, as $\theta = \mathsf{f} \rightsquigarrow \theta \propto \eta$.

2.15.5 Interpretation

For exponential families of distributions

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = \frac{1-\alpha}{2} T_{ijk}$$

and

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = (\partial_a B_b^i) B_c^j g_{ij} + B_a^i B_b^j B_c^k \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}).$$

The results of the previous Section will now be used to investigate these α–connections for GLMs.

[29] Using the notation of Bishop and Goldberg (1980), a repeated *upper case* superscript and subscript will be taken as a nonsum index. Such an upper case index will take on the same value as its lower case counterpart.

[30] $\mathsf{f}_i \overset{\text{def}}{=} \mathsf{f}(\boldsymbol{X}_i^\top \boldsymbol{\beta})$.

2.15.6 Canonical Link

If the link is canonical, then

$$\partial_a B_b^i = 0$$

giving

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = B_a^i B_b^j B_c^k \delta T_{ijk}(\boldsymbol{\theta}).$$

Now in general

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = \delta \ \overset{m}{\Gamma}_{abc}(\boldsymbol{u}) + (1-\delta) \ \overset{e}{\Gamma}_{abc}(\boldsymbol{u})$$

i.e.

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = \delta[\ \overset{m}{\Gamma}_{abc}(\boldsymbol{u}) - \ \overset{e}{\Gamma}_{abc}(\boldsymbol{u})] + \ \overset{e}{\Gamma}_{abc}(\boldsymbol{u}) = \delta T_{abc}$$

where

$$T_{abc} = E(\partial_a \ell \partial_b \ell \partial_c \ell) = B_a^i B_b^j B_c^k E(\partial_i \ell \partial_j \ell \partial_k \ell) = B_a^i B_b^j B_c^k T_{ijk}.$$

So

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = B_a^i B_b^j B_c^k \delta T_{ijk}(\boldsymbol{\theta}) = \delta T_{abc}$$

and, since the link is canonical,

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \frac{1-\alpha}{2} X_{ia} X_{jb} X_{kc} T_{ijk}(\boldsymbol{\theta}). \tag{2.24}$$

In agreement with Amari (1982b, p4, 2.10), the following definition is given for the skewness of the score function

$$\overset{C}{T}_{abc} = X_{ia} X_{jb} X_{kc} T_{ijk} \ . \tag{2.25}$$

2.15.7 Non–Canonical Link

When the link is non–canonical,

$$\overset{\alpha}{\Gamma}_{abc} = \delta T_{abc} + \ \overset{e}{\Gamma}_{abc}$$

$$\overset{\alpha}{\Gamma}_{abc} = \delta B_a^i B_b^j B_c^k T_{ijk} + (\partial_a B_b^j) B_c^k g_{jk} + B_a^i B_b^j B_c^k \ \overset{e}{\Gamma}_{ijk}.$$

Now

$$B_a^i = \frac{\partial \theta^i}{\partial u^a} = \frac{\partial \mathsf{f}\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}\right)}{\partial \beta^a} = \frac{\partial \mathsf{f}_i}{\partial \eta_I} X_{Ia}$$

giving

$$\overset{\alpha}{\Gamma}_{abc} = \delta \overset{\varphi}{T}_{abc} + \partial_a \left(\frac{\partial \mathsf{f}_j}{\partial \eta_J} X_{Jb} \right) \frac{\partial \mathsf{f}_k}{\partial \eta_K} X_{Kc} g_{jk}$$

since $\overset{e}{\Gamma}_{ijk} = 0$. The skewness $\overset{\varphi}{T}_{abc}$ is defined as

$$\overset{\varphi}{T}_{abc} = B_a^i B_b^j B_c^k T_{ijk} \left(= X_{Ia} X_{Jb} X_{Kc} \frac{\partial \mathsf{f}_i}{\partial \eta_I} \frac{\partial \mathsf{f}_j}{\partial \eta_J} \frac{\partial \mathsf{f}_k}{\partial \eta_K} T_{ijk} = \frac{\partial \mathsf{f}_i}{\partial \eta_I} \frac{\partial \mathsf{f}_j}{\partial \eta_J} \frac{\partial \mathsf{f}_k}{\partial \eta_K} \overset{C}{T}_{abc} \right) \quad (2.26)$$

by extending the previous notation for $\overset{C}{T}_{abc}$. The α–connection is now

$$\overset{\alpha}{\Gamma}_{abc} = \delta \frac{\partial \mathsf{f}_i}{\partial \eta_I} X_{Ia} \frac{\partial \mathsf{f}_j}{\partial \eta_J} X_{Jb} \frac{\partial \mathsf{f}_k}{\partial \eta_K} X_{Kc} T_{ijk} + \frac{\partial}{\partial \beta^a} \left(\frac{\partial \mathsf{f}_j}{\partial \eta_J} \right) X_{Jb} \frac{\partial \mathsf{f}_k}{\partial \eta_K} X_{Kc} g_{jk},$$

which gives

$$\overset{\alpha}{\Gamma}_{abc} = \tfrac{1-\alpha}{2} X_{Ia} X_{Jb} X_{Kc} T_{ijk} \frac{\partial \mathsf{f}_i}{\partial \eta_I} \frac{\partial \mathsf{f}_j}{\partial \eta_J} \frac{\partial \mathsf{f}_k}{\partial \eta_K} + \frac{\partial}{\partial \eta_I} \left(\frac{\partial \mathsf{f}_j}{\partial \eta_J} \right) X_{Ia} X_{Jb} X_{Kc} \frac{\partial \mathsf{f}_k}{\partial \eta_K} g_{jk},$$

and, finally

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \tfrac{1-\alpha}{2} X_{Ia} X_{Jb} X_{Kc} T_{ijk}(\boldsymbol{\theta}) \left(\frac{\partial \mathsf{f}_i}{\partial \eta_I} \frac{\partial \mathsf{f}_j}{\partial \eta_J} \frac{\partial \mathsf{f}_k}{\partial \eta_K} \right) + X_{Ia} X_{Jb} X_{Kc} g_{jk} \left(\frac{\partial^2 \mathsf{f}_j}{\partial \eta_I \partial \eta_J} \frac{\partial \mathsf{f}_k}{\partial \eta_K} \right) . \quad (2.27)$$

2.15.8 Discussion

The two Sections 2.15.6 and 2.15.7 can now be compared. The following points can be made

- Equation (2.27) reduces to Equation (2.24) when the link is canonical.
- The second term in Equation (2.27) is independent of α, and is just the exponential connection $\overset{e}{\Gamma}_{abc}(\boldsymbol{\beta})$.
- Because the second term in Equation (2.27) can be zeroed by choosing a canonical link, this term is *model* dependent.
- The first term in Equation (2.27) is *error* dependent, since it can be zeroed by choosing Normal errors. This term is proportional to the skewness tensor $\overset{\varphi}{T}_{abc}$, and will be exactly equal to the skewness term in Equation (2.24), when the link is canonical.

The results obtained in Equation (2.24) and Equation (2.27) can be compared with those of Pregibon (1980) and Efron (1975).

1. The α–connections derived in Equation (2.24) and Equation (2.27) form the basis of a test of link adequacy, by comparing various link functions with the canonical link, which some statistical packages use as the default. Pregibon (1980) has proposed a goodness–of–link test that subsumes the 'correct' link function into a family of types and compares alternatives via the deviance. The proposed test would have a different focus, since it involve comparisons of link functions with the default (canonical). The idea for the type of test is suggested by the form of the α–connections for canonical and non–canonical link as given in Equation (2.24) and Equation (2.27) respectively. As shown in Section 2.15.6 and Section 2.15.7, these α–connections are related to the skewness of the score function. In particular, the skewness of the score with respect to the regression coefficients can be written in terms of the skewness with respect to the natural parameters. It is this relationship that forms the basis of the test for a canonical link function. The test proceeds by fitting the canonical link to subsets of the data to determine if the relation between T_{abc} and T_{ijk} is linear, as suggested by Equation (2.25). Departure from this suggested form, would be taken as evidence of the link function being non–canonical. This test is described via examples in Section 2.15.9.

2. Efron(1975) has defined a measure of statistical curvature that vanishes for exponential families. The measure of statistical curvature γ_θ is defined by

$$\gamma_\theta^2 = \frac{\nu_{02}}{i_\theta^2} - \frac{\nu_{11}^2}{i_\theta^3}$$

where

$$i_\theta = -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = E\left(\frac{\partial \ell}{\partial \theta}\right)^2,$$

$$\nu_{02} = E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right)^2 - i_\theta^2$$

and

$$\nu_{11} = E\left(\frac{\partial \ell}{\partial \theta}\frac{\partial^2 \ell}{\partial \theta^2}\right).$$

The single parameter of interest is θ. The estimate of statistical curvature γ_θ is designed to be a measure of the nearness of the model to an exponential family type. Such exponential family models exhibit desirable statistical properties such as allowing the application of linear methods, encompassing locally most powerful tests and efficient estimation (Seber and Wild, 1989, p160). So, models with low statistical curvature could be expected to behave in a similar manner to such exponential family models and to inherit their corresponding good statistical properties. For an exponential family, the log–likelihood is defined by

$$\ell = y\theta - \psi(\theta) = c(y)$$

to give

$$\partial \ell/\partial \theta = \dot{\ell} = y - \psi'(\theta)\ ,\ \partial^2 \ell/\partial \theta^2 = \ddot{\ell} = -\psi''(\theta).$$

The square of the statistical curvature γ_θ becomes

$$\gamma_\theta^2 = \frac{\nu_{02}}{i_\theta^2} - \frac{\nu_{11}}{i_\theta^3}$$

where $i_\theta = \psi''(\theta)$. Since

$$\nu_{02} = E\ddot{\ell}^2 - i_\theta^2 = 0$$

and

$$\nu_{11} = E\dot{\ell}\ddot{\ell} = -\psi'' E\dot{\ell} = 0$$

by the score statistic, then $\gamma_\theta = 0$ for an exponential family. This measure of statistical curvature is defined for one parameter models subject to regularity conditions similar to those described in Section 2.1.2, (Efron, 1975, p1191 and p1196). One parameter models which are curved subsets of higher dimensional exponential families are called 'curved exponential families', as defined in Section 1.7. For such models, Efron (1975) showed that the bias and asymptotic variance of the MLE contain terms involving γ_θ and 'naming

curvature', so called as it is dependent on the form of model parameterization. An interpretation of γ_θ is that the loss of information due to curvature reduces the sample size from n to $n-\gamma_\theta^2$, since the MLE extracts all but $i_\theta\gamma_\theta^2$ of the information in the sample. Extension of Efron's curvature to the multi–parameter case was considered by Reeds (1975), in the discussion of Efron's paper. The extension was to multi–parameter curved exponential families. Reeds (1975) noted that, in the single parameter case, a transformation can always be found that will eliminate naming curvature (parameter–effects curvature), but that this is not necessarily so in the multi–parameter case. This transformation feature is reported also by Amari (1990), Kass (1984), and Hougaard (1982), for curved exponential families. The measures of statistical curvature and 'naming curvature' of Efron can be shown to be related to those of Bates and Watts (1980), for the case of Normal errors. The statistical curvature of Efron then becomes the intrinsic curvature of Bates and Watts (1980), as described in Seber and Wild (1989, p160). The corresponding 'naming curvature' becomes the parameter–effects curvature of Bates and Watts(1980), as described in Bates and Watts (1981, p1166) and Seber and Wild (1989, p164).

The acceleration and velocity terms given by Efron (1975, p1196, equation 4.2) can be used from the general case to verify these assertions for Normal Errors. Using the notation of Bates and Watts (1980), the intrinsic curvature is

$$\gamma^N = \frac{|\ddot{\eta}^N|}{|\dot{\eta}|^2}$$

while the parameter–effects curvature is

$$\gamma^T = \frac{|\ddot{\eta}^T|}{|\dot{\eta}|^2} .$$

Using the orthogonal decomposition of the acceleration in (4.2) of Efron (1975, p1196) gives

$$\gamma^T = |\frac{\nu_{11}}{i^{1/2}}|/i = \frac{\nu_{11}}{i^{3/2}}$$

and

$$\gamma^N = \frac{|i\gamma_\theta|}{i} = \gamma_\theta$$

as required. A unit scale has been used throughout, since from (4.1) of Efron (1975, p1196), $\Sigma = I$.

Although Efron's measure of statistical curvature is defined for one–dimensional models, several observations can be made that extend to cases of general interest and application.

- For a GLM that consists of a constant predictor only (the null model), $\theta_i = \theta \; \forall \; i$, the likelihood is

$$\ln f(y_i; \theta) = y_i\theta - b(\theta) + c(y_i)$$

giving $\nu_{11} = \nu_{02} = 0$ to yield

$$M_\theta = \begin{pmatrix} \nu_{20} & \nu_{11} \\ \nu_{11} & \nu_{02} \end{pmatrix} = \begin{pmatrix} \mathrm{x} & 0 \\ 0 & 0 \end{pmatrix}$$

causing $\gamma_\theta = 0$ as expected, since θ is then the natural (canonical) parameter. The form of link function is irrelevant.

- Consider the Poisson regression problem of Efron (1975, p1193). For the stated problem

$$\eta_\theta = \ln(a + \theta b_i)$$

where η is the natural parameter in Efron's notation. If the *canonical* link (log) is chosen, then

$$\eta_\theta = a + \beta b_i$$

and M_θ and γ_θ have the same values as the null model, as expected.

- The one–parameter family described by Efron (1975, p1194) as

$$\eta_\theta = a + b\tau(\theta)$$

is a superfamily of one–dimensional GLMs, since for a GLM the natural parameter η is in linear form (ie, 'canonical', in GLM terminology) as a function of θ.[31] Hence the statistical curvature γ_θ is zero. This highlights the fact that Efron's (1975) statistical curvature is the exponential curvature, which explains the alternative connotation of 'Efron' for the exponential connection, as defined by Dawid (1975). Note that η is the natural parameter in Efron's (1975) notation.

- In the one dimensional case, the exponential connection is

$$\overset{e}{\Gamma}_\theta = E(\frac{\partial \ell}{\partial \theta})(\frac{\partial^2 \ell}{\partial \theta^2}) = \nu_{11}(\theta)$$

where ν_{11} is as given by (3.21) of Efron (1975, p1195). Since $\overset{e}{\Gamma}_\theta$ is the exponential connection, then $\nu_{11}(\theta) = 0$ for exponential family models, again reinforcing the name 'Efron' for the exponential connection given by Dawid (1975).

2.15.9 Link Adequacy

This section describes a test for judging the adequacy of a canonical link in the fitting of a GLM to data. From Equation (2.25), the form of the skewness of the score function for a GLM with canonical link is

$$\overset{C}{T}_{abc} = X_{ia}X_{jb}X_{kc}T_{ijk} \tag{2.28}$$

whereas for a non–canonical link the skewness relation becomes [from Equation (2.26)]

$$\overset{\mathscr{C}}{T}_{abc} = \frac{\partial \mathsf{f}_i}{\partial \eta_I}\frac{\partial \mathsf{f}_j}{\partial \eta_J}\frac{\partial \mathsf{f}_k}{\partial \eta_K}T_{ijk} = \frac{\partial \mathsf{f}_i}{\partial \eta_I}\frac{\partial \mathsf{f}_j}{\partial \eta_J}\frac{\partial \mathsf{f}_k}{\partial \eta_K}\overset{C}{T}_{abc} \ . \tag{2.29}$$

So, the rationale of the test is to calculate the LHS in Equation (2.28) from the data and to compare this with the RHS in Equation (2.28) as calculated from the data, using the form of T_{ijk} suggested by a GLM with canonical link. Departures from a linear relation between T_{abc} and T_{ijk} will be taken as evidence of the link being non–canonical, since such a departure is suggested by Equation (2.26). Some recasting

[31] For a GLM, $\eta_\theta = \theta = \tau(\theta)$ and $b = \frac{1}{a(\phi)}$.

of Equation (2.28) is needed to allow the calculations to be performed in terms of the regression coefficients and not the score function based on those coefficients. To simplify this recasting, the calculations will be demonstrated on a GLM with Poisson errors, since this allows some simplification over other error types. As the test is based on the assumption of the link being canonical, Equation (2.28) will be used. A one dimensional derivation will be presented to highlight the approach with a minimum of complexity. The first step is to convert from skewness of the score function to skewness of the regression coefficients, since

$$T_{abc} = E\partial_a \ell \partial_b \ell \partial_c \ell.$$

Using a result due to Bartlett (1953a, p315, (27)),

$$\frac{\partial L}{\partial \widehat{\beta}} = \frac{\partial L}{\partial \beta} + \frac{\partial^2 L}{\partial \beta^2}\left(\widehat{\beta} - \beta\right) + \ldots \tag{2.30}$$

where $L(\equiv \ell)$ is the log–likelihood in Bartlett's(1953b) notation, and β and θ follow from the formulation of a GLM due to McCullagh and Nelder (1989). This gives

$$0 = \frac{\partial L}{\partial \beta} + \frac{\partial^2 L}{\partial \beta^2}\left(\widehat{\beta} - \beta\right) + \ldots$$

which becomes

$$\frac{\partial L}{\partial \beta} = -\frac{\partial^2 L}{\partial \beta^2}\left(\widehat{\beta} - \beta\right) + \ldots$$

The second derivative can be expressed as

$$\frac{\partial^2 L}{\partial \beta^2} = \frac{\partial}{\partial \beta}\left(\frac{\partial L}{\partial \beta}\right) = \frac{\partial}{\partial \beta}\left(\frac{\partial L}{\partial \theta}\frac{\partial \theta}{\partial \beta}\right) = \frac{\partial}{\partial \beta}\left(\frac{\partial L}{\partial \theta}\right)\frac{\partial \theta}{\partial \beta} + \frac{\partial L}{\partial \theta}\frac{\partial^2 \theta}{\partial \beta^2}$$

$$= \frac{\partial}{\partial \theta}\left(\frac{\partial L}{\partial \theta}\right)\frac{\partial \theta}{\partial \beta}\frac{\partial \theta}{\partial \beta} + \frac{\partial L}{\partial \theta}\frac{\partial^2 \theta}{\partial \beta^2} = \frac{\partial^2 L}{\partial \theta^2}\left(\frac{\partial \theta}{\partial \beta}\right)^2 + \frac{\partial L}{\partial \theta}\frac{\partial^2 \theta}{\partial \beta^2}.$$

If the link is canonical in the GLM, then

$$L = \frac{y\theta - b(\theta)}{a(\phi)} + \ldots$$

giving

$$\frac{\partial L}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$$

and

$$\frac{\partial^2 L}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)},$$

with $\theta = \eta = X\beta$ leading to

$$\frac{\partial \theta}{\partial \beta} = X$$

and thus

$$\frac{\partial^2 \theta}{\partial \beta^2} = 0.$$

Hence

$$\frac{\partial^2 L}{\partial \beta^2} = \frac{\partial^2 L}{\partial \theta^2} X^2 = -\frac{b''(\theta)}{a(\phi)} X^2$$

giving

$$\frac{\partial L}{\partial \beta} = -\frac{\partial^2 L}{\partial \beta^2}\left(\widehat{\beta} - \beta\right) = \frac{b''(\theta)}{a(\phi)} X^2 \left(\widehat{\beta} - \beta\right).$$

So, now

$$E\left(\frac{\partial L}{\partial \beta}\right)^3 = E\left(\frac{b''(\theta)}{a(\phi)}\right)^3 X^6 \left(\widehat{\beta} - \beta\right)^3$$

giving the skewness as

$$T_{abc} = E\left(\frac{\partial L}{\partial \beta}\right)^3 = E\left(\widehat{\beta} - \beta\right)^3 \left(\frac{b''(\theta)}{a(\phi)}\right)^3 X^6 = X_{ia} X_{jb} X_{kc} T_{ijk} = X^3 b'''(\theta)/a(\phi)^3$$

using Equation (2.17) in Section 2.13. For Poisson errors, $a(\phi) = 1$ and $b(\theta) = e^\theta$ yielding

$$b'(\theta) = b''(\theta) = b'''(\theta) = b(\theta) = \lambda$$

where λ represents the mean value (fitted value). If this mean is denoted by f, then

$$E\left(\widehat{\beta} - \beta\right)^3 = X^{-3} f^{-2}. \tag{2.31}$$

So, a plot of the (raw) skewness against $X^{-3}f^{-2}$ should be linear if the link is canonical.

Rationale of the Test

If the link function is canonical, the resulting graph of estimated skewness against expected value (of the skewness) should be linear, by virtue of Equation (2.28).

If the link function is non–canonical, then the resulting graph will not necessarily be linear, as predicted by Equation (2.29). The following set of Examples have been chosen to demonstrate these two cases (canonical and non–canonical link) empirically. In practice, for a given set of data, various link functions would be tried, and the corresponding graphs used to decide on the appropriate link function. Ultimately this test judges the link function as being canonical by the linearity of the plot of observed versus expected skewness. As the user will not know the true link function in reality, another variant of the simulation could fit the canonical link function to data generated from a non–canonical link. Likewise, a non–canonical link could be fitted to data generated from a canonical link. Both of these variants on the simulation would check the ability of the test to determine departures from canonicality.

Test Examples

Several simulations were run to demonstrate the workings of the approach. The examples are adapted from the problem given in Dobson (1993, p42). The same experimental design has been used in all the Examples. Simulated data (Y) from the Poisson distribution were generated twice at each of 5 levels of the predictor variable ($X = 1, 2, 3, 4, 5$), giving 10 observations in all, with different link functions expressing the relation between the expected value of the distribution and the predictor, eg, for the identity link, $E(Y) = \beta X$. The link functions used in each of the examples are given in Table 2.4.

Example	Link
1	Reciprocal
2	Identity
3	Logarithm
4	Square Root

Table 2.4: Link functions used in the Examples.

A sub–sampling scheme was used whereby 3 data points corresponding to consecutive predictor values were used and the known model fitted to the data. This was repeated for all possible combinations within each cell of 3 consecutive X values, generating 8 sets of 3 data points covering $X = (1, 2, 3)$, $(2, 3, 4)$ and $(3, 4, 5)$. The 8 regression coefficients so obtained were used to estimate the skewness of the regression coefficient $\widehat{\beta}$ in 2 lots each based on 4 regression estimates. These two skewness estimates were taken as representing the skewness of the regression coefficient in the model centred at the median of the spread of predictor values. Thus, skewness estimates were obtained for $X = 2, 3, 4$. The skewness estimates found were plotted against $X^{-3}f^{-2}$ and the relation judged for linearity. Departure from linearity in the plot of observed skewness against expected skewness should indicate that the link function differs from the canonical.

Example 1

The data given in Table 2.5 were generated from a GLM with reciprocal link and Poisson errors. The expected value was given by

$$E(Y) = 1/\beta X, \quad X = 1, 2, \ldots 5$$

with $\beta = 0.1$. Using the procedure described in 'Test Examples', the skewness $[K_3(s)]$ of the regression coefficients centred on $X = 2, 3, 4$ was estimated using the statistical package SPSS (Norusis, 1993). These estimates of the skewness are given in Table 2.6, together with their standard errors (SE).

Since this link function was chosen to be unlike the log link (canonical link for Poisson errors), it is expected that the plot of observed skewness versus expected skewness should be unlike a linear relation. This is verified from the plot given in Figure 2.7, where the pattern of response is clearly nonlinear. The label 'skewness' refers to observed skewness(standardized in SPSS), while the label 'expected' corresponds to $X^{-3}f^{-2}$. The difference in scale between 'skewness' and 'expected' is due to the standardization of observed skewness by SPSS.

Y	12,17	8,6	4,3	4,1	1,3
X	1	2	3	4	5

Table 2.5: Example 1 : Poisson data with reciprocal link

	X					
	2		3		4	
$K_3(s)$	0.17	0.13	0.22	0.25	0.91	1.01
SE	0.08	0.07	0.08	0.09	0.10	0.11

Table 2.6: Example 1 : Skewness and the standard error of the coefficients

This Example has demonstrated empirically the behaviour predicted by Equation (2.26), ie., the form of skewness for non–canonical link.

Example 2

The data given in Table 2.7 were generated from a GLM with identity link and Poisson errors. The expected value was given by

$$E(Y) = \beta X, \quad X = 1, 2, \ldots 5$$

with $\beta = 5$. Using the procedure described in 'Test Examples', the skewness $[K_3(s)]$ of the regression coefficients centred on $X = 2, 3, 4$ was estimated using the statistical package SPSS (Norusis, 1993). These estimates of the skewness are given in Table 2.8, together with their standard errors (SE).

For this link function, no obvious skewness is present and so a plot is not given. It is clear that there is no obvious relation between the skewness and the expected value as calculated under the assumption of a canonical link. This link function is obviously not similar to the log link, which is the canonical link for the Poisson distribution.

Y	5,6	12,11	15,11	18,27	28,24
X	1	2	3	4	5

Table 2.7: Example 2 : Poisson data with identity link

	X					
	2		3		4	
$K_3(s)$	-0.001	0.001	0.000	0.000	0.000	0.000
SE	0.40	0.40	0.63	0.63	0.47	0.47

Table 2.8: Example 2 : Skewness and the standard error of the coefficients

Again, this Example shows that the pattern of skewness for non–canonical link as described by Equation (2.26) will be different to that for canonical link as described by Equation (2.25).

Example 3

The data given in Table 2.9 were generated from a GLM with log link and Poisson errors. The expected value was given by

$$E(Y) = e^{\beta X}, \quad X = 1, 2, \ldots 5$$

with $\beta = 1$. Using the procedure described in 'Test Examples', the skewness $[K_3(s)]$ of the regression coefficients centred on $X = 2, 3, 4$ was estimated using the statistical package SPSS (Norusis, 1993). These estimates of the skewness are given in Table 2.10, together with their standard errors (SE). This link (log) is the canonical link function for Poisson errors, and so a linear relation between the observed skewness and expected skewness should be evident. The plot (Figure 2.8) does not show the departure from a linear relation that was shown by the non–canonical links (reciprocal and identity). The label 'skewness' denotes observed skewness(standardized in SPSS), while the label 'expected' denotes $X^{-3}f^{-2}$. Several other plots were produced, some using the unstandardized skewness, but all gave similar results.

Y	4,5	5,11	19,14	60,52	139,160
X	1	2	3	4	5

Table 2.9: Example 3 : Poisson data with log link

	X					
	2		3		4	
$K_3(s)$	-0.36	-0.36	-0.11	-0.09	-0.06	-0.07
SE	0.09	0.09	0.02	0.01	0.01	0.02

Table 2.10: Example 3 : Skewness and the standard error of the coefficients

The pattern of skewness for this Example is in line with the predictions of Equation (2.25) for a canonical link.

Example 4

The data given in Table 2.11 were generated from a GLM with square root link and Poisson errors. The expected value was given by

$$E(Y) = (\beta X)^2, \quad X = 1, 2, \ldots 5$$

with $\beta = 3$. Using the procedure described in 'Test Examples', the skewness $[K_3(s)]$ of the regression coefficients centred on $X = 2, 3, 4$ was estimated using the statistical package SPSS (Norusis, 1993). These estimates of the skewness are given in Table 2.12, together with their standard errors (SE).

The plot (Figure 2.9) is qualitatively similar to the log link (Example 3), possibly reflecting similarities between the log and square root functions over some range of the X values. The label 'skewness' refers to observed skewness(standardized in SPSS), while the label 'expected' gives $X^{-3}f^{-2}$.

Y	5,9	44,34	89,79	132,141	198,220
X	1	2	3	4	5

Table 2.11: Example 4 : Poisson data with square root link

	X					
	2		3		4	
$K_3(s)$	-0.08	-0.08	0.00	-0.03	-0.02	-0.02
SE	0.10	0.10	0.04	0.04	0.05	0.05

Table 2.12: Example 4 : Skewness and the standard error of the coefficients

This Example suggests that this link function ($\sqrt{}$) is similar to the log link which is the canonical link for Poisson errors.

Comments

1. The standardized measures of skewness used in statistical packages are variants of

$$K_3 = E(y - \mu)^3/\sigma^3.$$

Since SPSS calculates skewness using standardization, the scales of the observed ('skewness') and the expected ('expected') will be different in the plots for Figure 2.7, Figure 2.8 and Figure 2.9. The first two examples showed constant variance and so the effect of using the standardized skewness in place of the raw value should be minimal. For Examples 3 and 4 the variance of the regression coefficient does not appear to be constant. Since the raw form of the skewness gave similar results, a different subsampling could be considered. The reason for considering such a scheme is that the applicability of Equation (2.31) is governed by the appropriateness of the Taylor's expansion in Equation (2.30). Hence reducing or changing the interval over which the observed skewness is calculated could change the relation shown in the plot of skewness. The validity of the Taylor's expansion could also affect the stability of the variance for the estimated skewness.

2. The graphic has no obvious meaning for Normal errors, since $T_{ijk} = 0$ and so no relation would be expected in the plot, as the skewness would be zero.

3. Overall, the graph of skewness versus expected value for the non-canonical links showed a clear departure from a linear relation, whereas the graph for the log link (canonical) showed no such departure. This linearity is predicted by Equation (2.25) in the case of canonical links. The square root link showed similar results to the log link, presumably due to like behaviour of the square root function to the log function over the region defined by the design points.

2.15.10 Summary

This analysis of the exponential connection complements that of Kass (1984)[32], who considered α–connections in the form

$$\overset{\alpha}{\Gamma}_{abc} = \delta \ \overset{m}{\Gamma}_{abc} + (1-\delta) \ \overset{e}{\Gamma}_{abc}$$

where[33]

$$\overset{m}{\Gamma}_{abc} = \ \overset{e}{\Gamma}_{abc} + T_{abc}$$

and then showed that the α–connections are identical *iff* $T_{abc} = 0$. This is shown here by setting $T_{abc} = 0$ to give

$$\overset{\alpha}{\Gamma}_{abc} = \ \overset{e}{\Gamma}_{abc}$$

which is then independent of α.

An alternative form of the α–connection has been used, viz,

$$\overset{\alpha}{\Gamma}_{abc} = \ \overset{e}{\Gamma}_{abc} + \delta T_{abc}$$

and the conditions under which the exponential connection ($\overset{e}{\Gamma}_{abc}$) vanishes are investigated. For generalized linear models with canonical link function, the exponential (or 'Efron') connection in terms of the regression coefficients is shown to vanish. Conversely, it has been shown that for GLMs a vanishing exponential connection in terms of the regression coefficients gives the link function as canonical.

[32]Be aware of the *notational* differences.

[33]Note that $\delta = (1-\alpha)/2$.

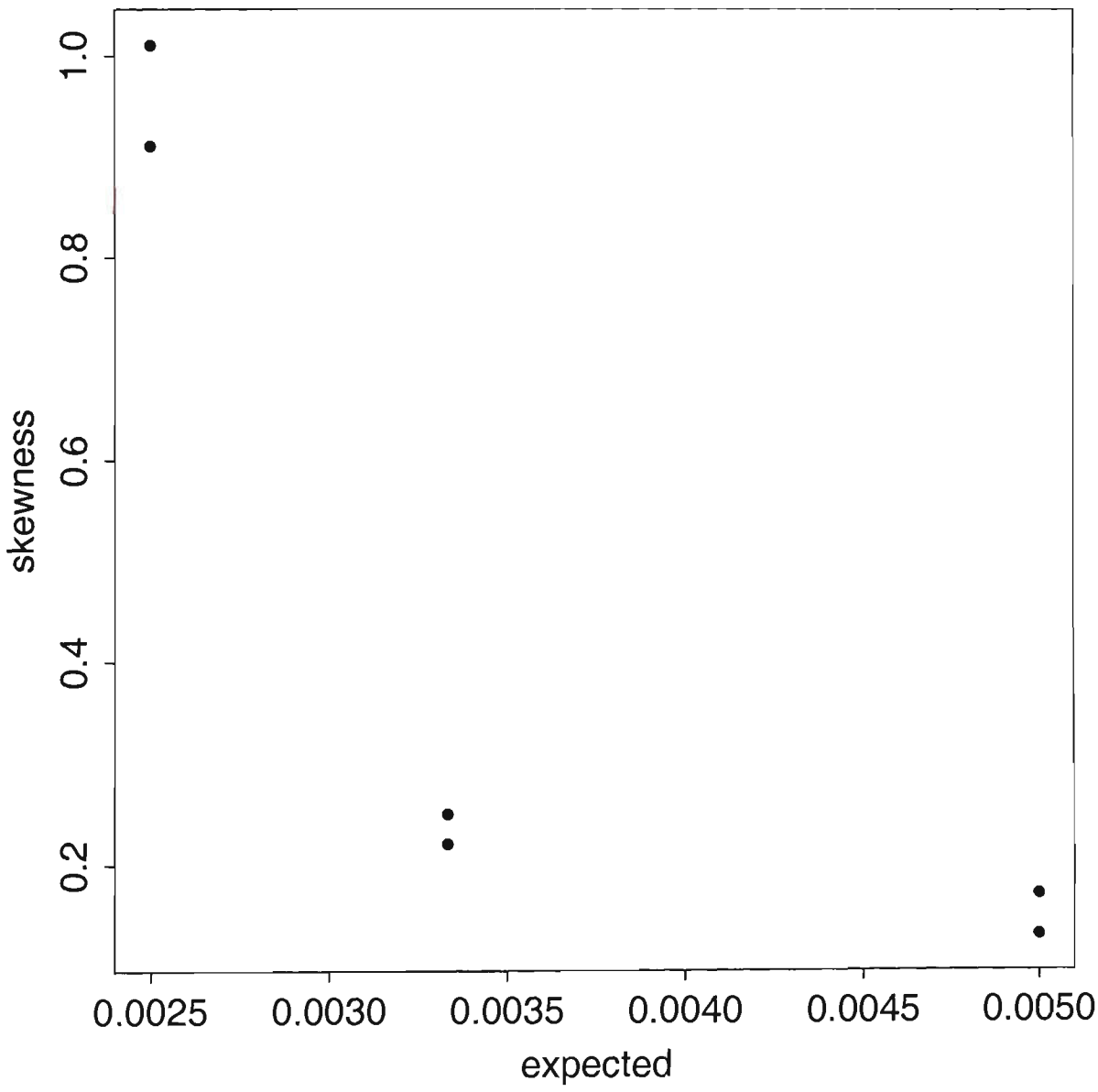

Figure 2.7: Example 1 : Reciprocal Link

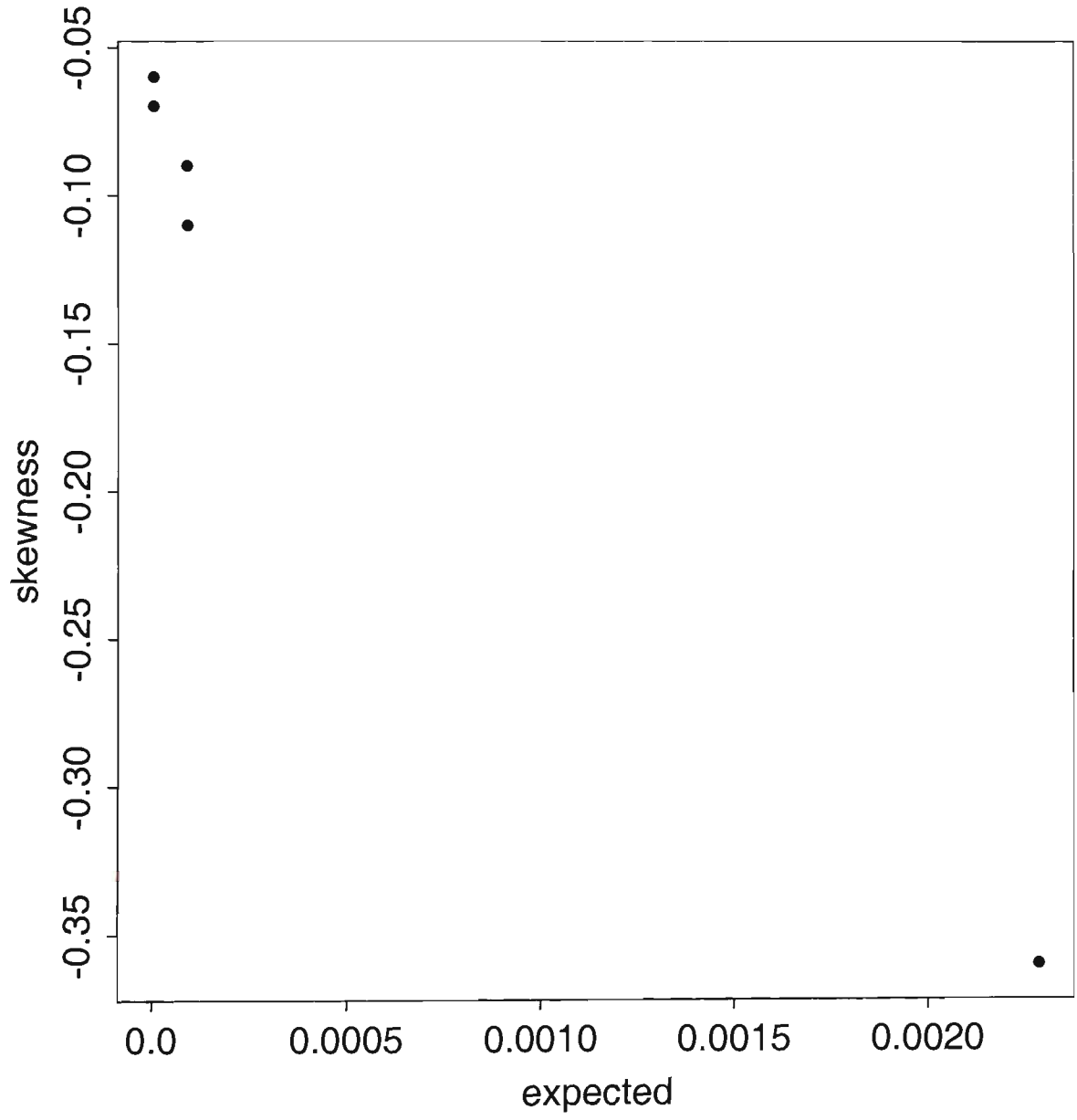

Figure 2.8: Example 3 : Log Link

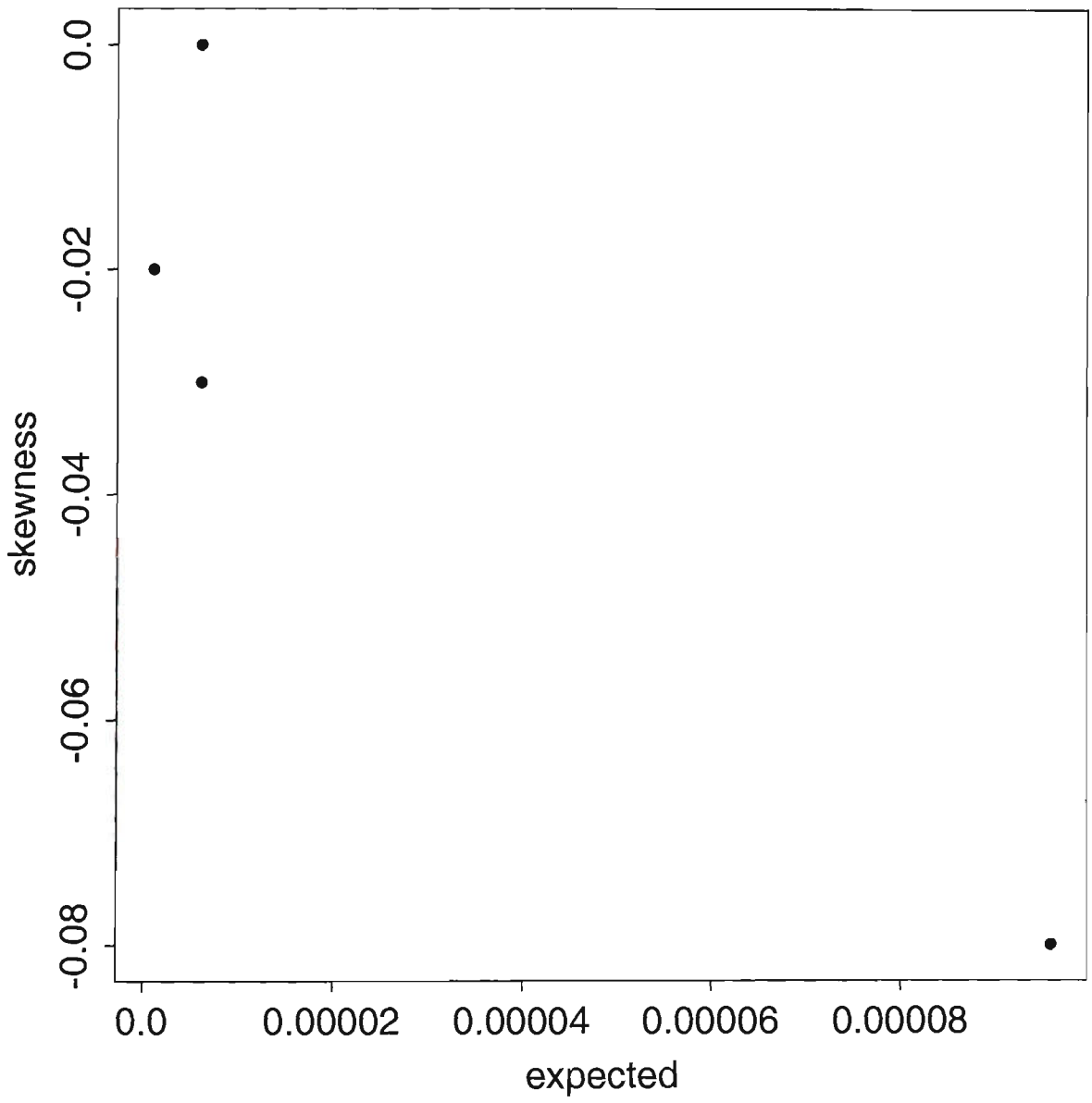

Figure 2.9: Example 4 : Square Root Link

Chapter 3

α–Curvatures

3.1 Introduction

This Chapter outlines the fundamentals of the generalization of curvature measures for exponential family error models, giving special attention to generalized linear models having a unit scale parameter, eg., those with error terms that are Bernoulli, Poisson or Exponential. This restriction is made simply to obviate a nuisance constant appearing in the relations used. The extension to models having scale parameter different to unity can be made by a simple rescaling of the natural parameter, as shown in Section 2.14.3.

The models considered address bivariate data $(\boldsymbol{X}_i, Y_i \ , \ i = 1 \ldots n)$. These models are of the type

$$Y_i = \mu_i + \varepsilon_i$$

where $\boldsymbol{\mu}$ is a deterministic function of the predictors $\boldsymbol{X}$, eg.,[1]

$$\mu_i = \boldsymbol{f}(X_{ij}\beta^j) \tag{3.1}$$

and $\boldsymbol{\varepsilon}$ is a disturbance describing the random behaviour of the response $\boldsymbol{Y}$.

For a generalized linear model (GLM), using the notation of McCullagh and

[1]The Einstein convention is used whereby a repeated index implies summation over that index.

Nelder (1989), the contribution to the log–likelihood for an observation is

$$\ln f(y;\theta) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi).$$

Models with unit scale parameter, $a(\phi) = 1$, correspond to the general form for an exponential family with log–likelihood

$$\ell = c(\boldsymbol{y}) + \theta^i y_i - \psi(\boldsymbol{\theta})$$

following Amari (1982a, 2.20, p362). The canonical (or natural) parameters $\boldsymbol{\theta}$ are related to the space of expectations $\boldsymbol{\mu}$ via

$$\mu_i = E(Y_i) = \partial_i \psi(\boldsymbol{\theta})$$

and the response Y_i is modelled as

$$Y_i = \mu_i + \varepsilon_i.$$

The deterministic component is μ_i and the disturbance is ε_i. In practice, the regression coefficients $\boldsymbol{u}$ that relate the expectation of the response Y_i to the predictors $\boldsymbol{X}_i$ are of interest, so $\boldsymbol{\theta}$ is a function of $\boldsymbol{u}$, viz

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{u}).$$

This explains the term *curved* exponential family, since even though the original form is exponential in terms of $\boldsymbol{\theta}$,[2] it may not be exponential in terms of $\boldsymbol{u}$. The dimension of the regression parameter space of $\boldsymbol{u}$ is less than that of the natural parameter space of $\boldsymbol{\theta}$. For a GLM in McCullagh and Nelder's (1989) notation

$$\boldsymbol{u} = \boldsymbol{\beta}$$

giving

$$\theta^i = \mathsf{f}(X_{ij}\beta^j)$$

using the notation of Appendix B.7. The restriction to GLMs with unit scale parameter avoids the notational inconvenience of the canonical parameter $\boldsymbol{\theta}$ having

[2]This implies that the form in the exponent is *linear* in θ.

a slightly different meaning in Amari's (1982a) notation to that of McCullagh and Nelder (1989). The restriction is made to save recurring constants appearing throughout the following discussion, and so $\boldsymbol{\theta} = a(\phi)\boldsymbol{\vartheta} = \boldsymbol{\vartheta}$ since $a(\phi) = 1$, from Section 2.14.1. This restriction to unit scale parameter is not critical, as similar results follow for an arbitrary valued scale parameter, as shown in Section 2.14.3.

3.1.1 Transformation Rule (Γ)

A reparameterization of the regression coefficients from $\boldsymbol{\beta}$ to $\boldsymbol{\mathcal{B}}$ implies a 1:1 transformation from $\boldsymbol{u} = (u^a)$ to $\boldsymbol{v} = (v^{a'})$. The α–connection with respect to the regression coefficients then transforms according to

$$\overset{\alpha}{\Gamma}_{a'b'c'} = B^a_{a'} B^b_{b'} B^c_{c'} \overset{\alpha}{\Gamma}_{abc} + B^a_{c'} \left(\partial_{a'} B^b_{b'}\right) g_{ab} \,, \tag{3.2}$$

where a' is associated with $\boldsymbol{\mathcal{B}}$, and a is associated with $\boldsymbol{\beta}$. So in general an affine connection is not a tensor, due to the presence of the second term. A general definition of tensors via coordinate transformation is given in Section 1.8.3. Note that the rule as stated by Amari (1982a, p364, 2.28), refers to natural parameters rather than regression coefficients, but the required relation is equivalent. An alternative treatment is given by Lovelock and Rund (1989, p79, 5.16). This transformation rule will be used later when examining properties of parameter–effects curvatures.

3.2 Curvatures

The imbedding $\boldsymbol{\theta}(\boldsymbol{u})$ defines a subspace $T_{\boldsymbol{u}}$ of the tangent space $T_{\boldsymbol{\theta}}$. This subspace is defined by the regression coefficients $\boldsymbol{u}$ and so is spanned by the vectors B^i_a. The curvature of a subspace is defined by the intrinsic change in the tangent (or normal) directions of the subspace. The tangent direction will generally be used.

3.2.1 Derivation

Following a similar argument to that used in the derivation of an α–connection in Chapter 2, the rate of change in the *tangent* direction from $B_b^i(\boldsymbol{u})$ at $\boldsymbol{u}$ to $B_b^i(\boldsymbol{u}+d\boldsymbol{u})$ at $\boldsymbol{u}+d\boldsymbol{u}$ is given by

$$\lim_{d\boldsymbol{u}\to 0} \frac{\boldsymbol{B}_b(\boldsymbol{u}+d\boldsymbol{u}) - \boldsymbol{B}_b(\boldsymbol{u})}{du^a} \rightsquigarrow \overset{\alpha}{\nabla}_a B_b$$

ie., the covariant derivative of the vector field $\boldsymbol{B}$.[3] This yields

$$\overset{\alpha}{H}{}_{ab}^{i}(\boldsymbol{u}) = \partial_a B_b^i(\boldsymbol{u}) + \overset{\alpha}{\Gamma}{}_{jk}^{i} B_a^j(\boldsymbol{u})\, B_b^k(\boldsymbol{u}) \tag{3.3}$$

as the definition of α–curvature. Note that the contravariant (upper index) version is described. A full derivation from first principles is given in Appendix C.1.

Note

An α–curvature $\overset{\alpha}{L}_{abi}$ can also be defined in the *normal* direction. This definition coincides with the covariant form of the tangential α–curvature[4] if

- $\alpha = 0$, i.e., when the information connection is used, or
- $T_{ijk} \overset{\text{def}}{=} E\left(\partial_i \ell \partial_j \ell \partial_k \ell\right) = 0$, i.e., errors are Gaussian $\Longleftrightarrow$ skewness tensor T_{ijk} is zero.

The L form is appropriate if expectation rather than canonical parameters are of interest, due to the *duality* between these two parameter spaces. The imbedded space of regression coefficients is the target. So it is of no real concern whether the tangential or normal form is used. For a full derivation, see Amari (1982a, p370, 4.9 and 4.10). Hereafter, all references to α–curvature will be to the contravariant or upper index form in the tangential direction, ie., $\overset{\alpha}{H}{}_{ab}^{i}$.

[3] Since nearby tangent spaces are compared, this derivative (curvature) must involve an affine connection (α–connection).

[4] That is, $\overset{\alpha}{L}_{abi} = \overset{\alpha}{H}_{abi} = \overset{\alpha}{H}{}_{ab}^{k} g_{ki}$.

3.2.2 Transformation Rule (H)

If the coordinate system is changed from $\boldsymbol{u} = (u^a)$ to $\boldsymbol{v} = (v^{a'})$ via reparameterization, then, α–curvature transforms according to

$$\overset{\alpha}{H}{}^{i}_{a'b'} = B^{a}_{a'} B^{b}_{b'} \overset{\alpha}{H}{}^{i}_{ab} + B^{i}_{b} \partial_{a'} B^{b}_{b'} \tag{3.4}$$

The derivation of this relation is given in Appendix C.2. The presence of the second term means that in general the α–curvature is not a tensor, since the transformation law for a (0,2) tensor is

$$\bar{S}_{hk} = B^{j}_{h} B^{l}_{k} S_{jl}$$

following Lovelock and Rund (1989, p60, 2.9), and Section 1.8.3. Special cases of this transformation rule will be used to identify intrinsic and parameter–effects curvatures.

3.3 Projections

In the case of nonlinear regression, Bates and Watts (1980) decomposed acceleration into components normal and tangential to the solution locus (expectation surface). For general exponential family error models, the analogue of this acceleration is α–curvature, which can similarly be decomposed into normal and tangential components. Below is a description of the projection of α–curvature onto these orthogonal subspaces. For brevity, the 'normal component of α–curvature' will be called the 'normal α–curvature', likewise the 'tangential component of α–curvature' will become 'tangential α–curvature'.

3.3.1 Normal Component

If the projection operator onto the normal subspace is N^{i}_{j} then the normal component of α–curvature is denoted by

$$\overset{\alpha}{\mathcal{N}}{}^{i}_{ab} = N^{i}_{j} \overset{\alpha}{H}{}^{j}_{ab}$$

where N^i_j is the projection operator onto the normal subspace of imbedded parameters ($\boldsymbol{\beta}$), ie, the regression coefficients.

The projection operator N^i_j is derived as[5]

$$N^i_j = \delta^i_j - P^i_j$$

where P^i_j is the projection operator onto the *tangential* subspace of imbedded parameters, ie., as per Amari (1990, p156), viz

$$P^i_j = g^{ab} B^i_b B^k_a g_{kj} \, .$$

Rewriting the operator as

$$P^i_j = \left(B^k_a g^{ab} B^i_b \right) g_{kj}$$

shows that the term inside the braces can be recognised as an ordinary projection; see Morgan (1993, p44) and Seber and Wild (1989, p683, A11.4.) Now

$$\overset{\alpha}{\mathcal{N}}{}^i_{ab} = N^i_j \overset{\alpha}{H}{}^j_{ab} = \left(\delta^i_j - P^i_j \right) \overset{\alpha}{H}{}^j_{ab}$$

$$= \overset{\alpha}{H}{}^i_{ab} - P^i_j \overset{\alpha}{H}{}^j_{ab} = \overset{\alpha}{H}{}^i_{ab} - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d$$

since

$$P^i_j \overset{\alpha}{H}{}^j_{ab} = \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d$$

from Section 3.3.3. Thus

$$\overset{\alpha}{\mathcal{N}}{}^i_{ab} = \partial_a B^i_b + \overset{\alpha}{\Gamma}{}^i_{jk} B^j_a B^k_b - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d \tag{3.5}$$

is the final general form for intrinsic curvature. This derivation agrees with the results of Amari (1990, p241), and Lovelock and Rund (1989, p269, 4.16).

Note

While this result is not new, it has not been described previously in detail. When the α–curvature is projected onto the normal subspace, the non–tensorial terms

[5]See Seber and Wild (1989, p691, B3.3).

vanish, leaving normal components of α–curvature that form a tensor. The vanishing of the non–tensorial terms in the transformation of α–curvature is shown in Appendix C.3. These normal components that form a tensor represent the intrinsic curvature of the imbedded subspace of regression coefficients. Following the definitions of Bates and Watts (1980), and the approach of Amari (1990, p156), the *scalar* measure of intrinsic curvature is derived in general. It will be shown that this measure is invariant under 1:1 transformations of the parameter space. The derivation of the scalar form of intrinsic curvature requires some fundamental results from Riemannian geometry.[6]

3.3.2 The Invariance of Intrinsic Curvature

This derivation deals with α–curvature, but the generic term 'curvature' will stand for α–curvature. Hence the use of the notation $\mathcal{N}$ will imply normal α–curvature $\overset{\alpha}{\mathcal{N}}$, that is the normal component of α–curvature.

Prelude

Define the curve $\theta^j = \theta^j(s)$ parametrically in terms of the arc length s. A line element on this curve is given by

$$ds^2 = g_{hj} d\theta^h d\theta^j \ .$$

The tangent vector $\theta'^j = d\theta^j/ds$ is a unit vector, since

$$g_{hj}\theta'^h\theta'^j = 1 \ .$$

Thus, covariant differentation yields

$$0 = \frac{D}{Ds}(g_{hj}\theta'^h\theta'^j) = 2g_{hj}\theta'^h\frac{D\theta'^j}{Ds}$$

[6]When the metric is positive definite, as here, the term *pseudo*–Riemannian is sometimes used.

showing that the vector $\dfrac{D\theta'^j}{Ds}$ is normal to the tangent vector θ'^j.[7] If the length of this normal vector is defined as

$$\left|\frac{D\theta'^j}{Ds}\right| = \frac{1}{\rho}$$

then ρ^{-1} can be interpreted as the *curvature of the curve.* (Lovelock and Rund, 1989, pp250–252.)

Derivation

If the parameterization is such that

$$\theta^j = \theta^j(\beta^a(s)) = \theta^j(u^a(s))$$

then

$$\frac{D\theta'^j}{Ds} = \mathcal{N}^j_{ab}u'^a u'^b + B^j_a\frac{Du'^a}{Ds}$$

where

$$u'^a = \frac{\partial\beta^a}{\partial s}\,.$$

The first term depends at each point P only on the coordinates β^a, and components u'^a of the unit tangent vector at P. This term is therefore identical for all curves of the tangent subspace which pass through P and have common tangent u'^a.

A *geodesic* Γ through P having the tangent u'^a is defined by $\dfrac{Du'^a}{Ds} = 0$, yielding

$$\left(\frac{D\theta'^j}{Ds}\right)_\Gamma = \mathcal{N}^j_{ab}u'^a u'^b.$$

From the definition of curvatures of curves, (Lovelock and Rund, 1989, p272) and (Stoker, 1969), the curvature of Γ is

$$\left(\frac{1}{\rho_\Gamma}\right)^2 = g_{jh}\left(\frac{D\theta'^j}{Ds}\right)_\Gamma\left(\frac{D\theta'^h}{Ds}\right)_\Gamma = g_{jh}\left(\mathcal{N}^j_{ab}u'^a u'^b\right)\left(\mathcal{N}^h_{eg}u'^e u'^g\right) \tag{3.6}$$

This quantity depends only on P and the direction u'^a at P. This 'normal curvature' is the scalar form of intrinsic curvature , (Lovelock and Rund, 1989, pp267–273).

[7]The operator D is defined by $\dfrac{DX^i}{D\theta^j} = \nabla_j X^i$, (Appendix C.1).

The invariance of scalar intrinsic curvature can now be demonstrated.

If the reparameterization is from $\boldsymbol{\beta}$ to $\boldsymbol{\mathcal{B}}$, viz, a to a' in terms of indices, then, in terms of $\boldsymbol{\mathcal{B}}$, intrinsic curvature becomes

$$\left(\frac{1}{\rho_{\Gamma'}}\right)^2 = g_{kl}\left(\mathcal{N}^k_{a'b'}u'^{a'}u'^{b'}\right)\left(\mathcal{N}^l_{e'g'}u'^{e'}u'^{g'}\right)$$

where $u'^{a'} = \dfrac{\partial \mathcal{B}^{a'}}{\partial s}$. The unit tangent vector can be written as

$$u'^{a'} = \frac{\partial u^{a'}}{\partial u^a}\frac{\partial u^a}{\partial s} = B^{a'}_a \frac{\partial u^a}{\partial s} = B^{a'}_a u'^a$$

From Appendix C.3, the tensorial law for the normal component of curvature gives

$$\mathcal{N}^k_{a'b'} = B^{a^*}_{a'} B^{b^*}_{b'} \mathcal{N}^k_{a^*b^*}$$

and so

$$\mathcal{N}^k_{a'b'}u'^a u'^b = B^{a^*}_{a'} B^{b^*}_{b'} \mathcal{N}^k_{a^*b^*}\left(B^{a'}_a B^{b'}_b u'^a u'^b\right)$$

but

$$B^{a^*}_{a'} B^{a'}_a = B^{a^*}_a$$

giving

$$\mathcal{N}^k_{a'b'}u'^{a'}u'^{b'} = \left(B^{a^*}_a B^{b^*}_b \mathcal{N}^k_{a^*b^*}\right)u'^a u'^b = \mathcal{N}^k_{ab}u'^a u'^b \,.$$

So

$$\left(\frac{1}{\rho_{\Gamma'}}\right)^2 = g_{kl}\left(\mathcal{N}^k_{ab}u'^a u'^b\right)\left(\mathcal{N}^l_{eg}u'^e u'^g\right) = \left(\frac{1}{\rho_{\Gamma}}\right)^2 .$$

Thus invariance is satisfied.

Note

This *new* result is quite general. It is not required that the family be curved exponential; a parameter subspace derived from any likelihood function would suffice. The result is a generalization of the proof of invariance in the case of nonlinear regression, given by Seber and Wild (1989, B5, pp692-694), using a Taylor's series expansion, following a reparameterization. The *scalar* form of intrinsic curvature

as derived above is a generalization of that due to Amari (1990, p156), given in the nonlinear regression case. This intrinsic component corresponds to those effects that are unchanged by reparameterization of the model. The normal component will only change if the 'model' is changed. In the model formulation,

$$Y_i = \mu_i + \varepsilon_i \ , \ \mu_i = \boldsymbol{f}(X_{ij}\beta^j)$$

a change of intrinsic curvature can only occur if

- The deterministic function f is changed, eg, for a GLM, by changing the link function, or if
- the error distribution ε is changed, eg, from say Poisson to Negative Binomial.

These observations follow from inspection of the form of the normal component of α–curvature, viz,

$$\overset{\alpha}{\mathcal{N}}{}^i_{ab} = \partial_a B^i_b + \overset{\alpha}{\Gamma}{}^i_{jk} B^j_a B^k_b - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d$$

which shows that a change in the deterministic function f or the error distribution ε will affect intrinsic curvature, since the first term can be removed by choosing a canonical link in a GLM, and the second term is zero for Gaussian errors. The components of the last term are also affected by the deterministic function and the error distribution, in general since

$$\overset{\alpha}{\Gamma}_{abc} = (\partial_a B^i_b) B^j_c g_{ij} + B^i_a B^j_b B^k_c \ \overset{\alpha}{\Gamma}_{ijk} \ .$$

3.3.3 Tangential Component

The projection of α–curvature onto the tangent subspace $T_{\boldsymbol{u}}$ yields an affine connection. The projection operator onto $T_{\boldsymbol{u}}$ is

$$P^i_j = g^{ab} B^i_b B^k_a g_{kj} \ ,$$

following Amari (1990, p156) and Seber and Wild (1989, pp690–691). Applying the projection operator P^i_j to the α–curvature $\overset{\alpha}{H}{}^j_{ab}$ yields the tangential component

$$\overset{\alpha}{\mathcal{T}}{}^i_{ab} = P^i_j \overset{\alpha}{H}{}^j_{ab} = g^{cd} B^i_d B^k_c g_{kj} \overset{\alpha}{H}{}^j_{ab} = \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d .$$

Proof

The projection of α–curvature onto the tangent subspace gives

$$\overset{\alpha}{\mathcal{T}}{}^{i}_{ab} \overset{\text{def}}{=} P^i_j \overset{\alpha}{H}{}^{j}_{ab} = \left(\partial_a B^j_b + \overset{\alpha}{\Gamma}{}^{j}_{ik} B^i_a B^k_b\right) P^i_j$$

$$= \left(\partial_a B^j_b + \overset{\alpha}{\Gamma}{}^{j}_{ik} B^i_a B^k_b\right) g^{cd} B^i_d B^l_c g_{lj}$$

$$= \left[\left(\partial_a B^j_b\right) B^l_c g_{lj} + \overset{\alpha}{\Gamma}{}^{j}_{ik} B^i_a B^k_b B^l_c g_{lj}\right] g^{cd} B^i_d$$

$$= \left[\left(\partial_a B^j_b\right) B^l_c g_{lj} + \overset{\alpha}{\Gamma}{}^{j}_{ik} g_{lj} B^i_a B^k_b B^l_c\right] g^{cd} B^i_d$$

$$= \left[\left(\partial_a B^j_b\right) B^l_c g_{lj} + \overset{\alpha}{\Gamma}_{ikl} B^i_a B^k_b B^l_c\right] g^{cd} B^i_d,$$

hence

$$\overset{\alpha}{\mathcal{T}}{}^{i}_{ab} = \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d \tag{3.7}$$

in agreement with Amari (1990, p156, 5.26). The tangential component of α–curvature ($\overset{\alpha}{\mathcal{T}}{}^{i}_{ab}$) is thus related to the α–connection ($\overset{\alpha}{\Gamma}_{abc}$). This tangential component is the generalization of 'parameter–effects' curvature from the nonlinear regression case, as defined by Bates and Watts (1980). In the general case, this measure will correspond to effects due to reparameterization of the deterministic function in the model.

This known result has not previously been derived in detail.

3.3.4 Scalar Parameter–effects Curvature

From the derivation of scalar intrinsic curvature the **scalar** form of generalized *parameter–effects* curvature can be defined. Instead of choosing a geodesic to create

$$\frac{Du'^a}{Ds} = 0,$$

any arbitrary curve in the regression parameter subspace can be used. This defines another scalar curvature, called the 'geodesic curvature',[8]

$$\left(\frac{1}{\rho_{\mathrm{g}}}\right)^2 = g_{ab}\left(\frac{Du'^a}{Ds}\right)\left(\frac{Du'^b}{Ds}\right) , \tag{3.8}$$

ie, the 'parameter–effects' curvature, so called since it is dependent on the choice of curve in the tangent subspace. The term 'geodesic' can be applied to parameter–effects curvature, since this quantity can be zeroed by choosing the arbitrary curve as a geodesic. The term 'tangential' curvature is used synonymously with 'geodesic' curvature by Struik (1988, p74 and p127).

3.4 Decomposition

For the nonlinear regression model, Bates and Watts (1980) showed that the normal and tangential components represented intrinsic and parameter–effects curvature respectively. In general exponential family error models, α–curvature can similarly be decomposed into normal and tangential components which have analogous interpretations. Using the notation for the normal and tangential components of α–curvature, the decomposition becomes

$$\overset{\alpha}{\mathcal{N}}{}^i_{ab} + \overset{\alpha}{\mathcal{T}}{}^i_{ab} = \overset{\alpha}{H}{}^i_{ab} \tag{3.9}$$

in agreement with Amari (1990, p156, 5.27).[9] These components were described by Amari (1982a), but were not explicitly derived for the general case.

The model proposed by Wei (1994) shows a similar form of decomposition of curvature into tangential and normal components. Using the notation of Wei (1994, pp329–330) the acceleration (W) is related to U by $U = H^{\mathsf{T}}WH$ and

$$U = [Q][A^p] + [N][A^I] = [Q][Q^{\mathsf{T}}i][U] + [N][N^{\mathsf{T}}i][U]$$

$$= \{P_T\}[U] + \{P_N\}[U]$$

[8] See Lovelock and Rund (1989, p272, 4.29).

[9] Amari (1990, p154) describes the nonlinear regression case.

where P_T and P_N are projection operators onto the tangential and normal spaces respectively. This decomposition is in line with Bates and Watts (1980) and Amari (1982a).

3.4.1 Decomposition of Scalar Curvature

A corresponding decomposition of the *scalar* form of curvature also holds, which generalizes the scalar decomposition described in Seber and Wild (1989, p131). This decomposition of scalar curvature into the normal (intrinsic) and tangential (parameter–effects) is described below. Note that the term 'geodesic' is synonymous with tangential or parameter–effects.

The curve $\theta^j = \theta^j(s)$ is defined parametrically in terms of the arc length s. In terms of the operator D, the vector

$$\frac{D\theta'^j}{Ds}$$

is normal to the tangent vector θ'^j. The length of this normal vector

$$\left|\frac{D\theta'^j}{Ds}\right| = \frac{1}{\rho}$$

can be used to form *total* scalar curvature, viz

$$\left(\frac{1}{\rho}\right)^2 = g_{ij}\left(\frac{D\theta'^i}{Ds}\right)\left(\frac{D\theta'^j}{Ds}\right)$$

as the form of generalized scalar curvature. If the parameterization is such that

$$\theta^j = \theta^j(\beta^a(s)) = \theta^j(u^a(s))$$

then

$$\frac{D\theta'^j}{Ds} = \mathcal{N}^j_{ab}u'^a u'^b + B^j_a\frac{Du'^a}{Ds}$$

where

$$u'^a = \frac{\partial\beta^a}{\partial s}\,.$$

Total scalar curvature becomes

$$\left(\frac{1}{\rho}\right)^2 = g_{jh}\left(\frac{D\theta'^j}{Ds}\right)\left(\frac{D\theta'^h}{Ds}\right) = g_{jh}\left(\mathcal{N}^j_{ab}u'^a u'^b + B^j_a\frac{Du'^a}{Ds}\right)\left(\mathcal{N}^h_{eg}u'^e u'^g + B^h_e\frac{Du'^e}{Ds}\right)$$

$$= g_{jh}\left(\mathcal{N}^j_{ab}u'^a u'^b\right)\left(\mathcal{N}^h_{eg}u'^e u'^g\right)$$

$$+g_{jh}\mathcal{N}^j_{ab}u'^a u'^b B^h_e \frac{Du'^e}{Ds} + g_{jh}\mathcal{N}^j_{eg}u'^e u'^g B^j_a \frac{Du'^a}{Ds} + g_{jh}B^j_a B^h_e \frac{Du'^a}{Ds}\frac{Du'^e}{Ds}$$

Application of the Lemma (Appendix C.4) removes the middle terms, and since

$$g_{jh}B^j_a B^h_e = g_{ae}$$

then

$$\left(\frac{1}{\rho}\right)^2 = \left(\frac{1}{\rho_\Gamma}\right)^2 + g_{ae}\left(\frac{Du'^a}{Ds}\right)\left(\frac{Du'^e}{Ds}\right)$$

$$= \left(\frac{1}{\rho_\Gamma}\right)^2 + \left(\frac{1}{\rho_g}\right)^2.$$

This demonstrates the decomposition of total scalar curvature into scalar intrinsic and scalar parameter–effects curvature. The subscript Γ is associated with intrinsic curvature and the subscript g is associated with parameter–effects curvature. For formal derivations, see Lovelock and Rund (1989, pp267–273), and Struik (1988, p69).

3.5 Examples

The general results obtained for α–curvature and its components are now illustrated in specific situations.

3.5.1 Nonlinear Regression

The α–curvature degenerates to the curvature measures of Bates and Watts (1980), for the nonlinear regression model.

Considering the α–curvature

$$\overset{\alpha}{H}{}^i_{ab}(\boldsymbol{u}) = \partial_a B^i_b(\boldsymbol{u}) + \overset{\alpha}{\Gamma}{}^i_{jk}\, B^j_a(\boldsymbol{u})\, B^k_b(\boldsymbol{u})$$

$$= \partial_a B^i_b + \overset{\alpha}{\Gamma}_{jkn} g^{ni} B^k_b B^j_a.$$

But, all the α–connections with respect to the natural parameters are zero for the Gaussian distribution, so the α–curvature becomes

$$\overset{\alpha}{H}{}^{i}_{ab}(\boldsymbol{u}) = \partial_a B^i_b(\boldsymbol{u}).$$

For the nonlinear regression model,

$$Y_i = \mu_i + \varepsilon_i = \mathsf{f}(\boldsymbol{X}_i;\boldsymbol{\beta}) + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2)$$

with

$$B^i_b = \frac{\partial \theta^i}{\partial u^b} = \frac{\partial \mathsf{f}_i}{\partial \beta^b} = \mathsf{f}'_{ib}.$$

In the notation of Equation (3.1), $\mathsf{f} = \boldsymbol{f}$ for Normal errors since $\theta^i = \mu^i$. This gives the α–curvature as

$$\overset{\alpha}{H}{}^{i}_{ab} = \partial_a \mathsf{f}'_{ib} = \frac{\partial^2}{\partial\beta^a \partial\beta^b}\mathsf{f}\,(\boldsymbol{X}_i;\boldsymbol{\beta}) \ ,$$

ie., the acceleration term of Bates and Watts (1988); see Seber and Wild (1989, pp129–133). The decomposition of this term into the normal and tangential components for the nonlinear regression model is seen as a special case of the decomposition of α–curvature. The behaviour of these components as intrinsic and parameter–effects is simplified by the disappearance of any error effect due to the disturbance law, since $\overset{\alpha}{\Gamma}_{ijk} = 0$ for Gaussian errors.

The generalized curvature H is independent of α for the nonlinear regression model.

Intrinsic Curvature

Intrinsic curvature for nonlinear regression has already been covered in the description of the normal component of α–curvature. The invariance of intrinsic curvature (Section 3.3.2) has been established in the general case as an extension of the result given in Seber and Wild (1989, B5), for nonlinear regression. This intrinsic curvature is a measure of model departure from a *linear* response, since the normal component of α–curvature vanishes, ie.,

$$\overset{\alpha}{\mathcal{N}}{}^{i}_{ab} = \partial_a B^i_b - \overset{\alpha}{\Gamma}_{abc}\, g^{cd} B^i_d = 0$$

for a linear model. The linear model is defined by

$$\mu_i = \mathsf{f}(\boldsymbol{X}_i; \boldsymbol{\beta}) = \boldsymbol{X}_i^\top \boldsymbol{\beta} = X_{ij}\beta^j$$

giving

$$B_b^i = \frac{\partial \mu_i}{\partial \beta^b} = X_{ib} \longrightarrow \partial_a B_b^i = 0,$$

which in turn annihilates parameter–effects curvature, since

$$\overset{\alpha}{\Gamma}_{abc} = \left(\partial_a B_b^i\right) B_c^j g_{ij} = 0.$$

Hence, $\overset{\alpha}{\mathcal{N}}{}_{ab}^i = 0$, as stipulated.

Parameter–Effects Curvature

Since the errors are Gaussian, the error effect vanishes, viz, $\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = 0$ giving

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = (\partial_a B_b^i) B_c^j \delta_{ij}/\sigma^2$$

which implies that $\overset{\alpha}{\Gamma}_{abc}$ is independent of α. Since the tangential component of α–curvature is effectively the α–connection, this leads to the description of 'parameter–effects' by a *common* affine connection, since the above shows that all the α–connections are the same for Gaussian Errors. Each of a number of key values of α is associated with desirable properties of estimators, such as unbiasedness, minimum variance etc; see Amari (1990, p152) and Kass (1984). Since the parameter–effects curvature is independent of α, this implies that all of these properties can be satisfied by a *single* parameterization,[10] for a given model and data set. So this parameterization could produce an estimator which is simultaneously unbiased, has minimum variance and zero skewness as well as other properties as detailed in Amari (1990, p152).

3.5.2 Generalized Linear Models

Generalized Linear Models (GLMs) were defined by Nelder and Wedderburn (1972). For a GLM with unit scale parameter, ie. with $a(\phi) = 1$, the canonical or natural

[10]See Amari (1990, p156).

parameter is

$$\theta^i = \mathsf{f}\left(X_{ij}\beta^j\right) = \mathsf{f}_i$$

and the tangent vector becomes

$$B_a^i = \frac{\partial\theta^i}{\partial u^a} = \frac{\partial \mathsf{f}_i}{\partial\beta^a} = \frac{\partial \mathsf{f}_i}{\partial\eta_I}\frac{\partial\eta_I}{\partial\beta^a} = \frac{\partial \mathsf{f}_i(\eta_I)}{\partial\eta_I}X_{Ia}.$$

Furthermore,

$$\partial_a B_b^i = \frac{\partial B_b^i}{\partial u^a} = \frac{\partial^2 \mathsf{f}_i}{\partial\beta^a\partial\beta^b} = \frac{\partial}{\partial\beta^a}\left(\frac{\partial \mathsf{f}_i(\eta_I)}{\partial\eta_I}X_{Ib}\right) = \frac{\partial^2 \mathsf{f}_i(\eta_I)}{\partial\eta_J\partial\eta_I}X_{Ja}X_{Ib}\ ,$$

thus B_a^i and $\partial_a B_b^i$ do not involve β, but are functions of the linear predictor η. Since the α–connection in terms of the regression coefficients $\boldsymbol{\beta}$ will be

$$\overset{\alpha}{\Gamma}_{abc} = (\partial_a B_b^i)B_c^j g_{ij} + B_a^i B_b^j B_c^k \overset{\alpha}{\Gamma}_{ijk},$$

this implies that this α–connection can be written as a function of η without explicit reference to β, a result that will be used later in examining parameter–effects curvature. For a non–Gaussian GLM, the α–connections are distinct (Kass 1984), since in general $\overset{\alpha}{\Gamma}_{ijk}$ is non zero, in contrast to the common connection for Gaussian errors, as in nonlinear regression. Hence a separate treatment is required for each of the key values of α ($-1, -\frac{1}{3}, 0, \frac{1}{3}$, and 1) that are associated with special properties of the estimators. Kass (1984) suggests the use of the mixture ($\alpha = -1$) and exponential ($\alpha = 1$) connections in assessing curvature effects, but since the constant of combination is arbitrary (see Section 2.5), an infinity of combinations is possible. Hence a desired statistical property of the estimator can be directly associated with a specific value of α. For example, zero skewness is related to $\alpha = -1/3$, while $\alpha = -1$ corresponds to unbiasedness. An assessment of each set of curvatures is required for each value of α, depending on which statistical property of the estimator is of interest.

The decomposition into tangential and normal components will be used, as generalizations of the curvatures used in nonlinear regression. The normal (intrinsic) and tangential (parameter–effects) components of curvature for a GLM will be considered in turn.

Intrinisic Curvature

The *intrinsic* component of curvature, i.e.,

$$\overset{\alpha}{\mathcal{N}}{}^{i}_{ab} = \partial_a B^i_b + \overset{\alpha}{\Gamma}{}^{i}_{jk} B^j_a B^k_b - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d$$

measures departures from exponentiality. This means departure from canonicality in a GLM since

$$\partial_a B^i_b = 0$$

for a GLM with a canonical link defined by

$$\theta_i = \eta_i = \mathsf{f}_i \;\Rightarrow B^i_b = \frac{\partial \mathsf{f}_i}{\partial \eta_I} X_{Ib} = X_{ib}.$$

The choice of the link function as canonical also affects both remaining terms in the normal component of curvature.

Parameter–Effects Curvature

The general form of tangential (parameter–effects) curvature is given by

$$\overset{\alpha}{\mathcal{T}}{}^{i}_{ab} = \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d.$$

For a GLM, $\overset{\alpha}{\mathcal{T}}{}^{i}_{ab}$ can be shown to form a tensor with respect to i, since if the reparameterization is from u^a to $v^{a'}$ then

$$\overset{\alpha}{\mathcal{T}}{}^{i}_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{\mathcal{T}}{}^{i}_{ab} + \left(\partial_{a'} B^b_{b'}\right) P^i_b,$$

from Appendix C.5. Since, for a GLM $B^b_{b'}$ is a matrix of constants (Section 4.1), then

$$\partial_{a'} B^b_{b'} = 0.$$

Thus, for a GLM, the tangential component of curvature obeys the tensorial law for a (0,2) tensor, viz,

$$\overset{\alpha}{\mathcal{T}}{}^{i}_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{\mathcal{T}}{}^{i}_{ab}.$$

Consequently, the α–connection is a tensor for a GLM (Section 4.1), implying that the tangential component of α–curvature must also be a tensor.

However the scalar form of tangential curvature

$$\left(\frac{1}{\rho_{\mathrm{g}}}\right)^2$$

is not necessarily an invariant (Section 3.3.3), in general. The transformation to a new parameterization for a GLM will always be linear, inducing a change of scale rather than a distortion (Section 4.2).

Investigation of parameter–effects curvature for a GLM in Chapter 4 will use the fact that the tangential component of α–curvature forms as tensor for a GLM. The form of parameter–effects α–curvature used in Chapter 4 will be scalar α–curvature.

3.6 Generalized Nonlinear Models

3.6.1 Definition

These models can be extended from generalized linear models and nonlinear regression models by defining

$$\theta^i = \mathsf{f}(\boldsymbol{X}_i; \boldsymbol{\beta})$$

with

$$Y_i = \mu_i + \varepsilon_i$$

where ε_i is from a distribution belonging to the exponential family, defined by

$$\ell = c(\boldsymbol{y}) + \theta^i y_i - \psi(\boldsymbol{\theta}).$$

This is indeed a generalisation from GLMs with unit scale parameter, since

$$\mu_i = E(Y_i) = \partial_i \psi(\boldsymbol{\theta}) = \partial_i \psi \left\{\mathsf{f}(\boldsymbol{X}_i; \boldsymbol{\beta})\right\} \stackrel{\text{def}}{=} p(\boldsymbol{X}_i; \boldsymbol{\beta}) \tag{3.10}$$

giving $\partial_i \psi$ as the identity function for nonlinear regression. The parameter θ^i is the natural parameter in the exponential family model, while the function f relates

these natural parameters $\boldsymbol{\theta}$ to the predictors $\boldsymbol{X}$ and their corresponding regression coefficients $\boldsymbol{\beta}$. A natural function q suggests itself, such that

$$\mathsf{q}(\mu_i) = \theta^i = \mathsf{f}(\boldsymbol{X}; \boldsymbol{\beta})$$

ie,

$$\mathsf{q} \equiv (\partial\psi)^{-1}.$$

The interpretation of this function q is that $\mathsf{q}p$ is the scale on which local sufficiency for $\boldsymbol{\beta}$ is assured by linearization.

For example, with Poisson errors and the response defined by

$$\mu = p(\boldsymbol{X}; \boldsymbol{\beta}) = \frac{\beta_1 x}{\beta_2 + x},$$

the Michaelis–Menten model (Michaelis and Menten, 1913) as reported in Bates and Watts (1988, p33),

$$\theta = \ln(\mu) = \mathsf{q}p = \ln(\beta_1 x) - \ln(\beta_2 + x) = \mathsf{f}$$

gives the scale on which local sufficiency for $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ can be determined via linearization.

For Normal errors, q is the identity and so local sufficiency via linearization is obtained directly on the scale of fitted values μ without transformation.

For a GLM

$$\theta^i = \mathsf{q}(\mu_i) = \mathsf{f}(\boldsymbol{X_i}^\top \boldsymbol{\beta})$$

and so

$$\mathsf{f}^{-1}\mathsf{q}(\mu_i) = g(\mu_i) = \boldsymbol{X_i}^\top \boldsymbol{\beta} = \eta_i$$

where g is the link function. If the link g is canonical, then f is the identity, giving

$$\mathsf{f}(\boldsymbol{X_i}^\top \boldsymbol{\beta}) = \boldsymbol{X_i}^\top \boldsymbol{\beta} = \eta_i$$

which leads to

$$\theta^i = \eta_i$$

ε_i	$p(\mathsf{f})$	q
Normal	f	identity
Poisson	e^{f}	logarithm
Bernoulli	$e^{\mathsf{f}}/(1+e^{\mathsf{f}})$	logit
Exponential	f^{-1}	reciprocal

Table 3.1: The functions p and q for generalized nonlinear models.

as required for a canonical link. Table 3.1 shows p as a function of f, together with the description of the natural function q for generalized nonlinear models (GNMs) with unit scale parameter.

The restriction to a unit scale parameter is simply to eliminate messy constants from the theoretical discussions. Results in the case of a non–unit scale parameter are similar to the above, as may be shown using arguments analogous to those in Section 2.14.1.

Thus, a generalized nonlinear model (GNM) can be seen to be a generalization of a generalized linear model (GLM) and a nonlinear regression model.

Wei and Zhu (1997, p130) have described an equivalent class of models that, by including a scale parameter, subsume the GNMs defined here. The exclusion of a scale parameter in the definition of GNMs is merely a theoretical convenience rather than an insurmountable restriction, as has been demonstrated in Section 2.14.2 for general curved exponential families. Consequently, the criticism by Wei (1998, p16) of the apparent restriction to a unit scale parameter in the curved exponential family model seems unwarranted, given the analysis of Section 2.14.2, where such models were shown to be able to incorporate an arbitrary scale parameter by a simple redefinition of the canonical parameter.

3.6.2 Curvatures

The general form of α–curvature for GNMs is given by

$$\overset{\alpha}{H}{}^{i}_{ab} = \frac{\partial^2 \mathsf{f}_i}{\partial\beta^a \partial\beta^b} + \overset{\alpha}{\Gamma}{}^{i}_{jk}\, \mathsf{f}'_{ja}\mathsf{f}'_{kb}$$

This can be seen to reduce to that given for nonlinear regression, since the α–connection with respect to the natural parameters vanishes for Gaussian errors.

Intrinsic curvature remains invariant (by definition), but for GNMs tangential or parameter–effects curvature no longer forms a tensor since the reparameterization cannot be guaranteed to be always linear. So, not surprisingly, GNMs inherit both the features of GLMs and nonlinear regression, ie.,

- The choice of α is determined by the feature of the estimator that is of interest, and
- parameter–effects curvature will be a function of the chosen parameterization of the model.

So the model and its form will affect estimation, and key features of the estimator will have to be investigated separately, by using special values of α.

Generalized curvature measures have been defined for curved exponential families. A decomposition of these α–curvatures into normal (intrinsic) and tangential (parameter–effects) components has been described, generalizing the situation for nonlinear regression. These curvature measures have been examined for generalized linear models and generalized nonlinear models. In the general case, contributions to curvature come from the Error distribution and from the deterministic component in the model. The contribution from the Error disturbance is zero for Normal Errors, and the contribution from the deterministic component will be zero for a GLM with canonical link.[11] Given that key values of α are associated

[11] A canonical link implies exponentiality and hence sufficiency with respect to the regression coefficients for a GLM.

with specific features of the estimators as described in Kass (1984), and Hougaard (1982), *three* aspects need to be considered in general

- the Error distribution,
- the form of the deterministic response, eg., the type of link for a GLM, and
- the property of the estimator that is of interest, ie., the choice of the value of α.[12]

All three of these influence the components of curvature. 'Nonlinearity' can be interpreted in the case of a GLM as a departure from canonical link, ie., a departure from exponentiality in terms of the regression coefficients. For Normal Errors, since the canonical link is the identity, the term 'nonlinear' is precise. For a generalized nonlinear model, an extended interpretation of 'nonlinearity' is not so forthcoming. The term 'linear' can also refer to the form of the response function. For a GLM, this simply means that the model is some function of a linear model of the regression coefficients. The interpretation described here extends another feature of Normal error linear models. This is the sufficiency of the estimators of the regression coefficients resulting from the implied exponential form of the resulting likelihood. For general models, some term such as 'non–sufficient' (or 'non–exponential') in place of 'non–linear' should be used to avoid confusion when describing the above extension.

3.6.3 Note

The generalized nonlinear models (GNMs) described in Section 3.6 should be distinguished from those models described in McCullagh and Nelder (1989, p379) where nonlinear parameters in the covariates were introduced. Such mildly nonlinear models are *also* called 'generalized nonlinear models' in the software implementation of GENSTAT 5, Release 3.2 (GENSTAT 5, 1993). The following extract

[12]To quote Hougaard (1982),' ... if you choose one, you (may) miss the others'.

from Genstat News of May 12, 1995 defines these models : 'The regression section now caters for "generalized nonlinear models". These are models that include some nonlinear parameters, but are otherwise in the form of generalized linear models. Such models are fitted relatively efficiently by fitting a standard g.l.m. at each stage of an iterative search for optimum values of the nonlinear parameters. One example is the model for probit analysis with unknown control mortality.' The following description of the procedure given in McCullagh and Nelder (1989) uses slightly different notation to avoid conflicting with similar usage elsewhere in this thesis. The usual linear component βX is replaced by a nonlinear covariate $\beta G(X;\Theta)$, with Θ unknown. Choosing a trial value Θ_0 the function $G(X;\Theta)$ is replaced by

$$G(X;\Theta_0) + (\Theta - \Theta_0)\,[\partial G/\partial\Theta]_{\Theta_0}$$

and so $\beta G(X;\Theta)$ is replaced by

$$\beta U + \gamma V$$

where

$$U = G(X;\Theta_0), \quad V = \partial G/\partial\Theta_0$$

and

$$\gamma = \beta(\Theta - \Theta_0).$$

In the iterative process, the new value for Θ becomes

$$\Theta_1 = \Theta_0 + \widehat{\gamma}/\widehat{\beta}.$$

The method is best for at most a few nonlinear parameters, due to the possibility of correlations amongst the parameter estimates.

Generalized nonlinear models (GNMs) are fundamentally different to the 'generalized nonlinear models' described in Genstat News (1995, May 12), since the linear predictor in such 'generalized nonlinear models' contains linear and nonlinear covariates, while a generalized nonlinear model (GNM) contains an arbitrary nonlinear function of the predictors and parameters.

The extension of 'non–linearity' in GLMs can be defined in terms of the variance function, the link function as well as for terms in the linear predictor, as described in McCullagh and Nelder (1989, pp372–378).

Other extensions of GLMs have been proposed by Jorgensen (1983); these allow correlated errors and nonlinear models for the expectation. For this 'extended class of generalized linear models' (Jorgensen, 1983, p20) it is no longer assumed that the density belongs to the exponential family. Also, the GLM restriction of the expected response being a function of a linear combination of the predictors has been relaxed, leading to the expected response being fully nonlinear. An unappealing consequence of the subsequent nonlinearity is that the simple GLM method for initiating the iterative fitting procedure, ie, the data, has been lost and starting values are required in general. The same criticism applies to GNMs. The general concepts of α–connections and α–curvatures in the extended case defined by Jorgensen (1983) can be applied as noted by Kass (1984, p89), and so the exponential family is not the only case covered by the differential geometric approach outlined in Chapter 2. The results of Chapter 2 can thus be applied to *any* non–Gaussian model.

3.7 Expected and Observed Geometries

So far, all the analysis has been in terms of *expected* geometries. There is a corresponding *observed* geometry, involving observed rather than expected information and an auxiliary statistic, Barndorff–Nielsen (1987, p135). This observed geometry is endowed with a full set of connections and tensors that mirror those in the expected geometry. The terminology used by Barndorff–Nielsen to denote the observed quantities is a slash. Thus if $\overset{\alpha}{\Gamma}$ is the α–connection in the expected geometry, then the corresponding connection in the observed geometry is denoted by $\overset{\alpha}{\not\Gamma}$. In terms of natural parameters [a (k, k) exponential model in Barndorff–Nielsen's notation], the observed and expected properties coincide, since no auxiliary statistic is involved due to the sufficiency of the natural parameters. So for example,

$\overset{\alpha}{\mathrm{F}} = \overset{\alpha}{\Gamma}$, and $\dot{\jmath} = i$ (i is the expected information matrix, while $\dot{\jmath}$ is the observed information matrix).

For a curved exponential family [a (k, d) exponential model],[13] asymptotic expansion of observed quantities such as the information and skewness tensor gives the corresponding expected quantities as the first term. Higher order terms in the metric tensor can be shown to disappear for zero exponential curvature (Barndorff–Nielsen, 1987, p139). In terms of GLMs, this indicates that the expected and observed metric tensors coincide for canonical links. Thus, expected and observed information coincide for GLMs with canonical links, in line with the observations of McCullagh and Nelder (1989, p43) and Aitkin, Anderson, Francis and Hinde (1989, p326).

[13] The number of data points is k and the number of regression coefficients is d, for say a GLM.

Chapter 4

Applications

The theory of generalized curvature will now be applied to generalized linear models and allied models with the purpose of extending results from Normal error linear models.

4.1 Tensorial α–connections and GLMs

Under the reparameterization from u^a to $v^{a'}$, the α–connection transforms to :

$$\overset{\alpha}{\Gamma}_{a'b'c'} = B^a_{a'} B^b_{b'} B^c_{c'} \overset{\alpha}{\Gamma}_{abc} + B^a_{c'} \left(\partial_{a'} B^b_{b'} \right) g_{ab} \, .$$

For a GLM, the response $\boldsymbol{Y}$ is defined by

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} = h(\boldsymbol{X}\boldsymbol{\beta}).$$

If the intrinsic curvature is to be unchanged, the model must remain the same. This means that the link function g must be unchanged and so the inverse function h will be unchanged. However since $\boldsymbol{\mu} = h(\boldsymbol{X}\boldsymbol{\beta}) = h(\boldsymbol{\eta})$ then the only form of reparameterization for a GLM is a linear transformation. Thus the linear predictor can be written as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{x}\boldsymbol{\mathcal{B}}$$

where the reparameterization is from $\boldsymbol{u}$ to $\boldsymbol{v}$. This reparameterization

$$\boldsymbol{\beta} = \boldsymbol{u}, \quad \boldsymbol{\mathcal{B}} = \boldsymbol{v}$$

gives

$$\boldsymbol{X\beta} = \boldsymbol{x\mathcal{B}} \rightarrow \boldsymbol{X}^\top \boldsymbol{X\beta} = \boldsymbol{X}^\top \boldsymbol{x\mathcal{B}}$$

with

$$\boldsymbol{\beta} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{x\mathcal{B}}.$$

Thus

$$B^b_{b'} = \frac{\partial u^b}{\partial v^{b'}}$$

becomes

$$\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\mathcal{B}}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{x}.$$

So $B^b_{b'}$ will always be a matrix of constants for a GLM, giving

$$\partial_{a'} B^b_{b'} = 0,$$

which implies that

$$\overset{\alpha}{\Gamma}_{a'b'c'} = B^a_{a'} B^b_{b'} B^c_{c'} \overset{\alpha}{\Gamma}_{abc} \tag{4.1}$$

i.e., $\overset{\alpha}{\Gamma}_{abc}$ behaves as a (0,3) tensor for a GLM. Thus the α–connection is a tensor for a generalized linear model.

Note

A precondition for establishing invariance of a quantity is to show that the quantity is a tensor (Bishop and Goldberg, 1980, p85). The above result must hold before scalar parameter–effects curvature can be shown to be invariant for a GLM.

4.1.1 Example

For a GLM with an arbitrary Error distribution and a specified link function, the linear predictor is

$$\boldsymbol{\eta} = \boldsymbol{X\beta} = \boldsymbol{x\mathcal{B}}$$

with

$$\boldsymbol{\beta} = \boldsymbol{u}, \quad \boldsymbol{\mathcal{B}} = \boldsymbol{v}.$$

The 'One Way Analysis of Variance' is a procedure applied to a simple model where the predictor $\boldsymbol{X}$ represents categories or grouping. The example given has 3 levels, ie., there are 3 groups. The parameterization v corresponds to estimates of group means τ_1, τ_2 and τ_3, whereas the u form is the default chosen by the computer package GLIM, ie., the **corner–point parameterization**, (Dobson, 1993, p89). This parameterization estimates a base line ($\tau_1 \rightsquigarrow$ group mean 1), and departures from that base line, viz,

$$\tau_2 - \tau_1 \rightsquigarrow \text{mean}_2 - \text{mean}_1$$

and

$$\tau_3 - \tau_1 \rightsquigarrow \text{mean}_3 - \text{mean}_1.$$

Table 4.1 shows the correspondence between the two forms of the same model, ie., the u and v parameterizations.

PARAMETERIZATION					
Corner–point			Group means		
τ_1	$\tau_2 - \tau_1$	$\tau_3 - \tau_1$	τ_1	τ_2	τ_3
u_1	u_2	u_3	$v_{1'}$	$v_{2'}$	$v_{3'}$

Table 4.1: The corner—point and group means parameterizations.

The Jacobian of the transformation from v to u is

$$B^a_{a'} = \frac{\partial u^a}{\partial v^{a'}}$$

ie.,

$$B^a_{a'} = \begin{array}{c|ccc} & u_1 & u_2 & u_3 \\ \hline v_1 & 1 & -1 & -1 \\ v_2 & 0 & 1 & 0 \\ v_3 & 0 & 0 & 1 \end{array}$$

which is indeed a matrix of constants, as claimed. Hence

$$\partial_{a'} B^b_{b'} = 0$$

inducing the tensorial law for the α–connection [Equation (4.1)]. So the invariance of parameter–effects curvature can now be established, since tensorial behaviour is a precondition to the establishment of invariance.

4.2 Invariance of Parameter–Effects Curvature

It has been demonstrated in Section 4.1 that the α–connection with respect to the regression coefficients forms a tensor for generalized linear models. The decomposition of scalar curvature into scalar intrinsic and scalar geodesic curvature has been described in Section 3.4.1. A consequent result, that scalar parameter–effects is invariant for a GLM, follows.

4.2.1 Theorem

Scalar parameter–effects curvature is an invariant for a generalized linear model.

Proof

The scalar form of parameter–effects curvature[1] is

$$\left(\frac{1}{\rho_{\mathbf{g}}}\right)^2 = g_{ab}\left(\frac{Du'^a}{Ds}\right)\left(\frac{Du'^b}{Ds}\right)$$

[1] Also known as geodesic or tangential curvature.

with definitions and notation from Section 3.3.3. To establish invariance, consider a reparameterization from $\boldsymbol{\beta}$ to $\boldsymbol{\mathcal{B}}$, i.e., u^a to $v^{a'}$. In the new parameterization, scalar parameter–effects curvature becomes

$$\left(\frac{1}{\rho_{g'}}\right)^2 = g_{a'b'}\left(\frac{Dv'^{a'}}{Ds}\right)\left(\frac{Dv'^{b'}}{Ds}\right)$$

where $v'^{a'} = \dfrac{\partial \mathcal{B}^{a'}}{\partial s}$. It is required to show that

$$\left(\frac{1}{\rho_{g'}}\right)^2 = \left(\frac{1}{\rho_g}\right)^2$$

in order to establish parameter–effects curvature as invariant.

The covariant differential is defined by

$$DX^j = dX^j + \Gamma^j_{hk} X^h dx^k.$$

In terms of the original parameters $u^a = \beta^a$ this becomes

$$Du'^a = du'^a + \Gamma^a_{bc} u'^b du^c$$

and in terms of the transformed coefficients $v^{a'} = \mathcal{B}^{a'}$

$$Dv'^{a'} = dv'^{a'} + \Gamma^{a'}_{\mathrm{bc}} v'^{\mathrm{b}} dv^{\mathrm{c}}.$$

These equations convert to

$$\frac{Du'^a}{Ds} = \frac{d^2u^a}{ds^2} + \Gamma^a_{bc}\frac{du^b}{ds}\frac{du^c}{ds} = \frac{du'^a}{ds} + \Gamma^a_{bc} u'^b u'^c$$

and

$$\frac{Dv'^{a'}}{Ds} = \frac{dv'^{a'}}{ds} + \Gamma^{a'}_{\mathrm{bc}} v'^{\mathrm{b}} v'^{\mathrm{c}}$$

as per Lovelock and Rund (1989, p254, 2.21). From the $v^{a'}$ parameterization, the scalar parameter–effects curvature is calculated as

$$\left(\frac{1}{\rho_{g'}}\right)^2$$

in terms of the original parameterization u^a. The indices $a\,b\,c$, $a{\cdot}b{\cdot}c{\cdot}$ are associated with $u^a = \boldsymbol{\beta}^a$, while $a'\,b'\,c'$, $\mathrm{a\,b\,c}$ relate to $v^{a'} = \boldsymbol{\mathcal{B}}^{a'}$.

Scalar parameter–effects curvature is now

$$\left(\frac{1}{\rho_{g'}}\right)^2 = g_{a'b'}\left(\frac{Dv'^{a'}}{Ds}\right)\left(\frac{Dv'^{b'}}{Ds}\right) \tag{4.2}$$

which becomes

$$\left(\frac{1}{\rho_{g'}}\right)^2 = g_{a'b'}\left[\frac{dv'^{a'}}{ds} + \Gamma^{a'}_{bc}v'^{b}v'^{c}\right]\left[\frac{dv'^{b'}}{ds} + \Gamma^{b'}_{de}v'^{d}v'^{e}\right] \tag{4.3}$$

and, from Section 3.3.2,

$$v'^{a'} = B^{a'}_{a}u'^{a}.$$

So

$$\frac{dv'^{a'}}{ds} = \frac{d}{ds}\left(B^{a'}_{a}u'^{a}\right)$$

$$= \frac{d}{du^b}\left(B^{a'}_{a}u'^{a}\right)\frac{du^b}{ds}$$

$$= \left[\partial_b(B^{a'}_{a})u'^{a} + B^{a'}_{a\cdot}\frac{du'^{a\cdot}}{du^b}\right]u'^{b}.$$

Following Section 4.1,

$$B^{a'}_{a} = \frac{\partial v^{a'}}{\partial u^a}$$

and $B^{a'}_{a}$ is a matrix of constants for a GLM, yielding

$$\partial_b(B^{a'}_{a}) = 0,$$

and so

$$\frac{dv'^{a'}}{ds} = B^{a'}_{a\cdot}(\partial_b u'^{a\cdot})u'^{b}.$$

The scalar curvature becomes

$$\left(\frac{1}{\rho_{g'}}\right)^2 = g_{a'b'}\left[B^{a'}_{a\cdot}(\partial_b u'^{a\cdot})u'^{b} + \Gamma^{a'}_{bc}v'^{b}v'^{c}\right]\left[B^{b'}_{b\cdot}(\partial_d u'^{b\cdot})u'^{d} + \Gamma^{b'}_{de}v'^{d}v'^{e}\right].$$

Now

$$v'^{b} = B^{b}_{b\cdot}u'^{b\cdot}$$

and

$$v'^{c} = B^{c}_{c\cdot}u'^{c\cdot}.$$

From the Lemma on page 141

$$\Gamma^{a'}_{\mathrm{bc}} = B^{a'}_{a\cdot}\Gamma^{a\cdot}_{\mathrm{bc}}$$

yielding

$$\Gamma^{a'}_{\mathrm{bc}}v'^{\mathrm{b}}v'^{\mathrm{c}} = B^{a'}_{a\cdot}\Gamma^{a\cdot}_{\mathrm{bc}}B^{\mathrm{b}}_{b\cdot}u'^{b\cdot}B^{\mathrm{c}}_{c\cdot}u'^{c\cdot} = B^{a'}_{a\cdot}\left[B^{\mathrm{b}}_{b\cdot}B^{\mathrm{c}}_{c\cdot}\Gamma^{a\cdot}_{\mathrm{bc}}\right]u'^{b\cdot}u'^{c\cdot} = B^{a'}_{a\cdot}\Gamma^{a\cdot}_{b\cdot c\cdot}u'^{b\cdot}u'^{c\cdot},$$

and similarly for the second term $\dfrac{Dv'^{b'}}{Ds}$, in Equation (4.2). Thus Equation (4.3) becomes

$$\left(\frac{1}{\rho_{\mathrm{g}'}}\right)^2 = g_{a'b'}\left[B^{a'}_{a\cdot}(\partial_b u'^{a\cdot})u'^{b} + B^{a'}_{a\cdot}\Gamma^{a\cdot}_{b\cdot c\cdot}u'^{b\cdot}u'^{c\cdot}\right]\left[B^{b'}_{b\cdot}(\partial_d u'^{b\cdot})u'^{d} + B^{b'}_{b\cdot}\Gamma^{b\cdot}_{d\cdot e\cdot}u'^{d\cdot}u'^{e\cdot}\right]$$

$$= g_{a'b'}B^{a'}_{a\cdot}B^{b'}_{b\cdot}\left[\frac{\partial u'^{a\cdot}}{\partial s} + \Gamma^{a\cdot}_{b\cdot c\cdot}u'^{b\cdot}u'^{c\cdot}\right]\left[\frac{\partial u'^{b\cdot}}{\partial s} + \Gamma^{b\cdot}_{d\cdot e\cdot}u'^{d\cdot}u'^{e\cdot}\right]$$

$$= g_{a\cdot b\cdot}\left(\frac{Du'^{a\cdot}}{Ds}\right)\left(\frac{Du'^{b\cdot}}{Ds}\right) = \left(\frac{1}{\rho_{\mathrm{g}}}\right)^2$$

and invariance of parameter–effects curvature for a GLM is established.

Note

In the special case of Normal errors, the above result has already been demonstrated, albeit indirectly, by Seber and Wild (1989, pp139–141). Consider linear transformations of the parameters. The transformations are

$$\phi = R\theta \text{ or } \theta = K\phi$$

where ϕ is associated with tangential scalar curvature Γ^T_d and θ has tangential scalar curvature γ^T_h, in their notation. It is shown that

$$\Gamma^T_d = \gamma^T_h.$$

Even though these are relative curvatures, the same result follows for scalar absolute curvatures. Due to typographical errors in their derivation (p141), a short synopsis is given below, using their notation

$$\Gamma_d^T = \rho \frac{||\boldsymbol{d}'\widehat{\boldsymbol{G}}_{\cdot\cdot}^T \boldsymbol{d}||}{||\widehat{\boldsymbol{G}}_{\cdot} \boldsymbol{d}||^2}$$

$$= \rho \frac{||\ddot{\boldsymbol{\mu}}_d^T||}{||\dot{\boldsymbol{\mu}}_d||^2}$$

$$= \rho \frac{||\ddot{\boldsymbol{\mu}}_h^T||}{||\dot{\boldsymbol{\mu}}_h||^2}$$

$$= \rho \frac{||\boldsymbol{h}'\widehat{\boldsymbol{F}}_{\cdot\cdot}^T \boldsymbol{h}||}{||\widehat{\boldsymbol{F}}_{\cdot} \boldsymbol{h}||^2}$$

$$= \gamma_h^T,$$

and so invariance is established under a linear reparameterization. Note that this result applies to a GLM with Normal errors and general link function.

Lemma

$$\Gamma_{\mathrm{bc}}^{a'} = B_{a\cdot}^{a'} \Gamma_{\mathrm{bc}}^{a\cdot} \quad .$$

Proof

Since Γ_{abc} is a tensor for a GLM,

$$\Gamma_{\mathrm{bc}}^{a'} = \Gamma_{\mathrm{bcd}} g^{\mathrm{d}a'}$$

$$= B_{\mathrm{d}}^{a} \Gamma_{\mathrm{bc}a} g^{\mathrm{d}a'}$$

$$= B_{\mathrm{d}}^{a} \ \Gamma_{\mathrm{bc}}^{a\cdot} g_{aa\cdot} g^{\mathrm{d}a'}$$

$$= \ \Gamma_{\mathrm{bc}}^{a\cdot} g^{\mathrm{d}a'} g_{aa\cdot} B_{\mathrm{d}}^{a} \ .$$

From Lovelock and Rund (1989, p268, 4.9),

$$g^{\alpha\varepsilon} g_{hj} B^h_\varepsilon = B^\alpha_\varepsilon$$

and so

$$\Gamma^{a'}_{\mathrm{bc}} = \Gamma^{a\cdot}_{\mathrm{bc}} B^{a'}_{a\cdot} \ .$$

4.2.2 Short Form of Proof

A condensed form of proof using statistical arguments can be invoked if results from previous sections can be combined with known relations from other sources.

Outline

The parameter–effects curvature (tangential component) $\overset{\alpha}{T}{}^i_{ab}$ is given by

$$\overset{\alpha}{T}{}^i_{ab} = \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d \ .$$

For a GLM,

$$B^i_d = X_{id}$$

and the metric tensor g is a function of the Error distribution, since

$$g_{ab} = B^i_a B^j_b g_{ij} (= X_{ia} X_{jb} g_{ij})$$

from Amari (1982a, 4.5, p370), and from

$$g_{ij} = b''(\theta^I)\delta_{ij}$$

in Section 2.13, again with upper case indices being nonsum. The form of the α–connection with respect to $\boldsymbol{\beta}$ is given by Equation (2.27) of Section 2.15.7, viz

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{\beta}) = \tfrac{1-\alpha}{2} X_{Ia} X_{Jb} X_{Kc} T_{ijk}(\boldsymbol{\theta}) \left(\frac{\partial \mathsf{f}_i}{\partial \eta_I} \frac{\partial \mathsf{f}_j}{\partial \eta_J} \frac{\partial \mathsf{f}_k}{\partial \eta_K} \right) + X_{Ia} X_{Jb} X_{Kc} g_{jk} \left(\frac{\partial^2 \mathsf{f}_j}{\partial \eta_I \partial \eta_J} \frac{\partial \mathsf{f}_k}{\partial \eta_K} \right)$$

where

$$\theta^i = \mathsf{f}(X_{ij}\beta^j) = \mathsf{f}_i \ .$$

The form of these expressions for the metric tensor g and the α–connection $\overset{\alpha}{\Gamma}$ indicate that parameter–effects curvature is a function of the linear predictor η, the function f and associated derivatives such as $\partial \mathsf{f}/\partial\eta$. The regression coefficients β do not appear explicitly in any of the formulae for g and $\overset{\alpha}{\Gamma}$, indicating that changes in the parameterization of a GLM will leave these quantities unaltered and hence parameter–effects $\mathcal{T}$ will be invariant to parameter transformation as previously shown.

4.3 Exponential Curvature

In the case of the exponential connection ($\alpha = 1$), the 'exponential' curvature (or 1–curvature) has been defined by Amari (1990, p114), and Amari (1987, p31).

4.3.1 Preamble

Following Kass (1984), a table of key values of α can be constructed (Table 4.2). In Table 4.2 a symbol has been designated as an additional identifier for the statistical interpretation of particular values of α.

α Value	Interpretation	Symbol
-1	Mean value (or Mixture) connection	m
-1/3	Skewness reducing	s
0	Information (or metric) connection	i
1/3	$E\left(\partial^3\ell/\partial\psi^3\right) = 0$; 'Normal likelihood'	n
1	Exponential (or Efron) connection	e

Table 4.2: Alternative symbols for the key values of α.

It should be noted that all quantities that are derived from α–connections

import the statistical interpretation peculiar to the specific value of α. Indeed, as remarked by Kass (1984, p87), ("In all other exponential families (other than the Normal) there are many 'parameter–effects' arrays"), an entire suite of curvatures is so generated. Key values of α have special connotations, but the 'exponential' (or Efron) connection ($\alpha = 1$) is the one often (implicitly) invoked, eg., Efron (1975), since this connection measures departure from the exponential form of distribution. The other main connection of interest is the information (or metric) connection ($\alpha = 0$), being a measure of departure from constant (co)–variance, as per Amari (1990).

4.3.2 Canonical Links in GLMs

The invariance of scalar intrinsic (α) curvature was established in Section 3.3.2. The following theorem holds for a particular value of α, [$\alpha = 1$ (e)], corresponding to the exponential connection and to exponential curvature. The result quoted holds for a GLM with canonical link.

Theorem

The scalar form of exponential intrinsic curvature for a GLM is minimal when the link is canonical.

Proof

The scalar form of intrinsic curvature as defined in Equation (3.6) is

$$\left(\frac{1}{\rho_\Gamma}\right)^2 = \left(\overset{\alpha}{\mathcal{N}}{}^j_{ab}\, u'^a u'^b\right)\left(\overset{\alpha}{\mathcal{N}}{}^h_{eg}\, u'^e u'^g\right) g_{jh}$$

where $u'^a = \dfrac{\partial \beta^a}{\partial s}$, with s being the arc length. The Normal component of α–curvature is defined as

$$\overset{\alpha}{\mathcal{N}}{}^i_{ab} = \partial_a B^i_b + \overset{\alpha}{\Gamma}{}^i_{jk} B^j_a B^k_b - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d.$$

Consider the exponential (or Efron) connection, corresponding to $\alpha = 1$. The scalar form of intrinsic e–curvature (1–curvature) becomes

$$\left(\frac{1}{\rho_\Gamma}\right)^2 = \left(\overset{e}{\mathcal{N}}{}^j_{ab}\, u'^a u'^b\right)\left(\overset{e}{\mathcal{N}}{}^h_{eg}\, u'^e u'^g\right) g_{jh}$$

with

$$\overset{e}{\mathcal{N}}{}^i_{ab} = \partial_a B^i_b + \overset{e}{\Gamma}{}^i_{jk} B^j_a B^k_b - \overset{e}{\Gamma}_{abc} g^{cd} B^i_d.$$

As a GLM is considered, and $\alpha = e$, then

$$\overset{e}{\Gamma}_{ijk} = 0$$

since the model is then from the exponential family, and further

$$\overset{e}{\Gamma}_{abc} = 0$$

for the resulting curved exponential family, as already shown in Section 2.15.4.

Thus

$$\overset{e}{\mathcal{N}}{}^i_{ab} = \partial_a B^i_b,$$

but for a GLM with canonical link

$$B^i_b = \frac{\partial \theta^i}{\partial \beta^b} = X_{ib} \rightsquigarrow \partial_a B^i_b = 0.$$

This gives

$$\overset{e}{\mathcal{N}}{}^i_{ab} = 0$$

which in turn means that

$$\left(\frac{1}{\rho_\Gamma}\right)^2 = 0.$$

This is a minimum, since the scalar form is positive definite.

Comment

The above result can be considered a generalisation of the Normal error/linear model combination, where the scalar intrinsic curvature is minimal (zero) for the canonical link which is the identity function, ie., a linear model.

4.4 The exponential form of α–curvature

In general, α–curvature is

$$\overset{\alpha}{H}{}^{i}_{ab} = \partial_a B^i_b + B^j_a B^k_b \overset{\alpha}{\Gamma}{}^{i}_{jk} = \overset{e}{H}{}^{i}_{ab} + B^j_a B^k_b \overset{\alpha}{\Gamma}{}^{i}_{jk}$$

since $\overset{e}{H}{}^{i}_{ab} = \partial_a B^i_b$ as $\overset{e}{\Gamma}{}^{i}_{jk} = 0$ for an exponential family model. Thus the leading term in α–curvature is simply the exponential curvature.

Since

$$\Gamma^k_{ji} = \Gamma_{jim} g^{mk}$$

then

$$\Gamma^j_{jk} = \Gamma_{jkm} g^{mi}.$$

This gives

$$\overset{\alpha}{H}{}^{i}_{ab} = \overset{e}{H}{}^{i}_{ab} + B^j_a B^k_b \overset{\alpha}{\Gamma}_{jkm} g^{mi}$$

$$= \overset{e}{H}{}^{i}_{ab} + B^j_a B^k_b \frac{1-\alpha}{2} T_{jkm} g^{mi}$$

and finally

$$\overset{\alpha}{H}{}^{i}_{ab} = \overset{e}{H}{}^{i}_{ab} + B^j_a B^k_b \frac{1-\alpha}{2} T^i_{jk}$$

where T^i_{jk} is the skewness tensor (contravariant form).

The following observations can be made.

- If errors are Normal, then

$$\overset{\alpha}{H}{}^{i}_{ab} = \overset{e}{H}{}^{i}_{ab} = \partial_a B^i_b$$

since the skewness tensor T_{ijk} is then zero, as the distribution is symmetric (Kass, 1984).

- For a GLM,

$$\overset{e}{H}{}^{i}_{ab} = \partial_a B^i_b$$

and further

$$\overset{e}{H}{}^{i}_{ab} = 0$$

for a canonical link.

- If the link is canonical in a GLM, or if $\partial_a B^i_b = 0$

$$\overset{\alpha}{H}{}^{i}_{ab} = B^j_a B^k_b \frac{1-\alpha}{2} T^i_{jk}.$$

4.5 Generalized Nonlinear Models

The following result can be given for Generalized Nonlinear Models (GNMs).

Theorem

Zeroing the first term in α–curvature in a GNM implies that the GNM is a GLM (special case) *and* that the link is canonical.

Proof

The condition

$$\overset{e}{H}{}^{i}_{ab} = \partial_a B^i_b = 0$$

implies that B^i_b is constant since

$$B^i_b = \frac{\partial \theta^i}{\partial \beta^b}.$$

This in turn means that for a GNM

$$\frac{\partial \mathsf{f}_i(\boldsymbol{X};\boldsymbol{\beta})}{\partial \beta^b} = \text{constant}$$

which implies that

$$\theta^i = \mathsf{f}_i(\boldsymbol{X};\boldsymbol{\beta}) = \text{constant} \times \boldsymbol{\beta}$$

ie., a GLM with canonical link, since $\theta = \eta$, the linear predictor.

The result can be thus stated.

A GNM with zero exponential curvature must be a GLM with canonical link.

4.6 Bias and Covariance of Estimators

Various workers have addressed the question of bias and (co)–variance of estimators, especially for models belonging to the curved exponential family. Key authors include Box (1971), Efron (1975), Clarke (1980), Bates and Watts (1980), Amari (1982a, 1990) and Corderio and McCullagh (1991). The two issues are bias and (co)–variance of estimators.

The first order ($o(N^{-1})$) term in the bias can be removed by an adjustment as given in Amari (1982a, p381) and Efron (1975, p1214, remark 11). This adjustment is a function of the mixture connection and mixture curvature of the ancillary subspace. See Section 2.10 and for the one–dimensional and Section 2.11 for the multi–dimensional case. The second order ($o(N^{-2})$) terms in the squared error of this corrected estimator become sums of squares of three terms, as given in Amari (1990, p133, Theorem 5.4) and Amari (1982a, p381).[2] These three components are sums of squares of the

1. mixture connection

2. exponential curvature, and

[2]Note the difference in treatment by a factor of N, viz, (5.29) of p381 Amari (1982a) versus (5.4) of p131 and (5.11) of p133 Amari (1990).

3. mixture connection of the ancillary subspace.

The first term depends on the parameterization, and so, in theory, could be eliminated by choice of parameter. This is called naming curvature or parameter–effects curvature. The second term is related to the exponential curvature which is known from the previous section to disappear for models such as GLMs with canonical link, and so is model dependent. The last term will be zero if the estimator is the MLE.

So, it can be seen from this breakdown that curvatures and connections feature in the assessment of properties of estimators.

In particular, it should be noted that for a GLM the parameter–effects curvature is invariant, and so this component of squared error could not be removed by parameter transformation, since the only possible transformations in a GLM are linear.

4.7 Variance Stabilizing Link Function

The constant information scale is used in model checking, as for example in McCullagh and Nelder (1989, p398), where the variance stabilizing transformation is listed for each error distribution that defines a GLM. The constant information scale has another interpretation as producing a link function that is variance stabilizing. According to Kass and Smyth (1990), this link function is the most frequent choice after the canonical. Choosing such a link zeros the 0–connection[3] (a special case of Section 2.10.3). A zero 0–connection implies that the link is variance stabilizing, ie., the link corresponds to the transformations given on page 398 of McCullagh and Nelder(1989, ed 2). Table 4.3 reproduces these constant–information transformations (link functions).

This condition (zero 0–connection) can be manipulated to produce an imbedded application of the elementary formula for the variance of a transformed random variable as given in Section 2.10.3.

[3]Also called the Riemannian connection.

Function	Error
μ	Normal
$2\sqrt{\mu}$	Poisson
$2\sin^{-1}\sqrt{\mu}$	Binomial
$2\ln\mu$	Gamma
$-2/\sqrt{\mu}$	Inverse Gaussian

Table 4.3: Constant information link functions.

The functions given in Table 4.3 can be obtained from the change of variable relation for the transformation $Y = g(X)$, ie,

$$V(Y) = V\left(g(X)\right) \approx \left(g'(\mu)\right)^2 V(X)$$

via Taylor's theorem applied on

$$Y = g(X) = g(\mu) + g'(\mu)(X - \mu) + \ldots$$

to give

$$E\left[Y - g(\mu)\right]^2 = \left[g'(\mu)\right]^2 E(X - \mu)^2 + \ldots$$

ie,

$$V(Y) \approx \left(g'(\mu)\right)^2 V(X).$$

In terms of the 0–connection, the choice of a particular link function is equivalent to a 1:1 transformation of the natural parameters. Under such a transformation the 0–connection on the new scale ($\boldsymbol{\xi}$) in terms of the old ($\boldsymbol{\theta}$) is given by a transformation rule as given by Equation (3.2). What is required is a 0–connection for the fitted value scale ($\boldsymbol{\xi} \equiv \boldsymbol{\mu}$) rather than the regression coefficients as given in Equation (3.2). The formulae are similar, but the parameters addressed are

different. To this end the required transformation rule is closer in notation to Amari (1982a, p364, 2.28), and Appendix C.1.

In this Section, the new scale (due to the link function) will be denoted by indices $i'\ j'\ k'$ whereas the original scale of natural coordinates will be denoted by $i\ j\ k$ as usual.

The scale of fitted values is also called the space of expectations, as described in Section 2.12 where the space of expectations is shown to be dual to the space of natural parameters, hence the use of the notation $\boldsymbol{\xi}$ for the fitted values.

For a GLM with non canonical link, the information connection in terms of the new scale is

$$\overset{0}{\Gamma}_{i'j'k'}(\boldsymbol{\xi}) = B^i_{k'}\left(\partial_{i'}B^j_{j'}\right)g_{ij} + B^i_{i'}B^j_{j'}B^k_{k'}\ \overset{0}{\Gamma}_{ijk}$$

A slight modification of previous notation is required to incorporate the effect of the choice of link function. In the notation of McCullagh and Nelder (1989), the link function g is defined by

$$g(\mu) = \eta \overset{\text{def}}{=} g(\xi)$$

and so

$$B^i_{i'} = \frac{\partial \mathsf{f}_i}{\partial \xi_{i'}} = \frac{\partial \mathsf{f}_i}{\partial \eta_I}\frac{\partial \eta_I}{\partial \xi_{i'}} = \frac{\partial \mathsf{f}_i}{\partial \eta_I}G_{Ii'}.$$

Noting that

$$g(\mu) = \eta = g(\xi)$$

it follows that

$$\frac{\partial \eta_I}{\partial \xi_{i'}} = g'(\mu) = g'(\xi) \overset{\text{def}}{=} G_{Ii'}.$$

The information connection on the new scale is then

$$\overset{0}{\Gamma}_{i'j'k'}(\boldsymbol{\xi}) = \tfrac{1}{2}G_{Ii'}G_{Jj'}G_{Kk'}T_{ijk}(\boldsymbol{\theta})\left(\frac{\partial \mathsf{f}_i}{\partial \eta_I}\frac{\partial \mathsf{f}_j}{\partial \eta_J}\frac{\partial \mathsf{f}_k}{\partial \eta_K}\right) + G_{Ii'}G_{Jj'}G_{Kk'}g_{jk}\left(\frac{\partial^2 \mathsf{f}_j}{\partial \eta_I \partial \eta_J}\frac{\partial \mathsf{f}_k}{\partial \eta_K}\right),$$

and so the zeroing of this quantity can be examined to find the transformation which induces stable variance for a particular error distribution. For a GLM in general the following relations hold

$$g_{ij} = b''(\theta^I)\delta_{ij}$$

and

$$T_{ijk} = b'''(\theta^K)E_{ijk},$$

with an upper case index being nonsum.

It can be shown that the previous link functions given in Table 4.3 induce stable variance for each error distribution as shown. In order to verify these results, a table of $b(\theta)$ and its derivatives will be required. These are given in Table 4.4 in the notation of McCullagh and Nelder (1989).

ERROR	$b(\theta)$	$b'(\theta)$	$b''(\theta)$	$b'''(\theta)$
Normal	$\theta^2/2$	θ	1	0
Poisson	e^θ	e^θ	e^θ	e^θ
Binomial	$\ln(1+e^\theta)$	$e^\theta/(1+e^\theta)$	$e^\theta/(1+e^\theta)^2$	$e^\theta(1-e^\theta)/(1+e^\theta)^3$
Gamma	$-\ln(-\theta)$	$-1/\theta$	$1/\theta^2$	$-2/\theta^3$
Inverse Gaussian	$-(-2\theta)^{1/2}$	$(-2\theta)^{-1/2}$	$(-2\theta)^{-3/2}$	$3(-2\theta)^{-5/2}$

Table 4.4: The canonical parameter function $b(\theta)$ and its derivatives.

For each of the distributions given in Table 4.3 and Table 4.4 it can be shown that choice of the nominated link function will zero the information connection, due to cancellation of terms in the given 0–connection. For example, in the case of Normal errors, $T_{ijk}(\boldsymbol{\theta}) \propto b'''(\boldsymbol{\theta}) = 0$ giving

$$\overset{0}{\Gamma}_{i'j'k'}(\boldsymbol{\xi}) = G_{Ii'}G_{Jj'}G_{Kk'}g_{jk}\left(\frac{\partial^2 \mathsf{f}_j}{\partial\eta_I\partial\eta_J}\frac{\partial \mathsf{f}_k}{\partial\eta_K}\right).$$

The variance stabilizing link from Table 4.3 is the identity, and since the identity is the canonical link for normal errors,

$$\theta = \eta = \mu = \mathsf{f}$$

giving

$$\frac{\partial^2 \mathsf{f}_j}{\partial\eta_I\partial\eta_J} = 0.$$

This gives

$$\overset{0}{\Gamma}_{i'j'k'}(\boldsymbol{\xi}) = 0$$

as expected for the Normal distribution.

Marginally more involved derivations are needed for other error distributions.

General Case

Examining the general form of the 0–connection under a transformation will show the general conditions under which this connection can be zeroed, and so induce constant variance on the scale of fitted values.

Using the form of the skewness tensor and information metric for a GLM gives the information connection as

$$\overset{0}{\Gamma}_{i'j'k'}(\boldsymbol{\xi}) = B^i_{k'}\left(\partial_{i'}B^j_{j'}\right)g_{ij} + B^i_{i'}B^j_{j'}B^k_{k'}\ \overset{0}{\Gamma}_{ijk}(\boldsymbol{\theta})$$

to become

$$\overset{0}{\Gamma}_{i'j'k'}(\boldsymbol{\xi}) = B^i_{k'}\left(\partial_{i'}B^j_{j'}\right)\left[b''(\theta^I)\delta_{ij}\right] + B^i_{i'}B^j_{j'}B^k_{k'}\left[b'''(\theta^K)E_{ijk}\right]/2$$

since

$$\overset{0}{\Gamma}_{ijk} = \frac{1}{2}T_{ijk}.$$

This gives

$$\overset{0}{\Gamma}_{i'j'k'}(\boldsymbol{\xi}) \propto 2B^i_{k'}\left(\partial_{i'}B^j_{j'}\right)b''(\theta^I)\delta_{ij} + B^i_{i'}B^j_{j'}B^k_{k'}b'''(\theta^K)E_{ijk}$$

$$= \frac{\partial}{\partial\xi_{i'}}\left[g_{jk}B^j_{j'}B^k_{k'}\right] = \frac{\partial}{\partial\xi_{i'}}\left[g_{j'k'}\right].$$

So, if the 0–connection with respect to $\boldsymbol{\xi}$ vanishes, ie,

$$\overset{0}{\Gamma}_{i'j'k'}(\boldsymbol{\xi}) = 0$$

this implies that

$$g_{j'k'} = \text{constant}.$$

That is, a constant information metric with respect to $\boldsymbol{\xi}$ (the scale of fitted values) is implied by the choice of link function.

Proof

Starting with

$$\frac{\partial}{\partial \xi_{i'}}\left[g_{j'k'}\right] = \frac{\partial}{\partial \xi_{i'}}\left[g_{jk}B^j_{j'}B^k_{k'}\right]$$

$$= \frac{\partial}{\partial \xi_{i'}}\left[b''(\theta^J)\delta_{jk}B^j_{j'}B^k_{k'}\right]$$

$$= b'''(\theta^I)E_{ijk}B^i_{i'}B^j_{j'}B^k_{k'} + b''(\theta^J)\delta_{jk}\left(\partial_{i'}B^j_{j'}\right)B^k_{k'} + b''(\theta^J)\delta_{jk}B^j_{j'}\left(\partial_{i'}B^k_{k'}\right)$$

which becomes

$$= b'''(\theta^I)E_{ijk}B^i_{i'}B^j_{j'}B^k_{k'} + 2b''(\theta^J)\delta_{jk}\left(\partial_{i'}B^j_{j'}\right)B^k_{k'}$$

by permuting the indices j and k. Since the indices ijk are arbitrary this gives the required result.

4.7.1 Other Link Functions

In the previous section, the relation between the information connection (0–connection) and the choice of the link function as being variance stabilizing was investigated. There are other link functions that can be associated with key values of α and the corresponding α–connection. As expected, these link functions induce those properties associated with the particular value of α. For example, the case of $\alpha = 1/3$, produces a 'normal' likelihood by zeroing the expected third derivative of the log–likelihood. Following Aitkin, Anderson, Francis and Hinde (1989), the corresponding link function for the Binomial would be an incomplete beta function, while for the Poisson the 'normal' link function is the cube root, as described in McCullagh and Nelder (1989, p198). The link functions corresponding to each of the key values of α are given in Table 4.5, together with their corresponding property.

The signed constants are included merely for completeness with the table extracted from McCullagh and Nelder (1989, p398), as reproduced in Table 4.3. For Binomial errors, the link functions for $\alpha = \pm 1/3$ are given in terms of the incomplete beta function (Kendall and Buckland, 1971)

	Mean ($= \mu$)	Canonical	Constant Var.	'Normal' ℓ	Skewness $\downarrow$
α	- 1	1	0	1/3	- 1/3
δ	1	0	1/2	1/3	2/3
Normal	θ	$\theta = \mu$	μ	μ	μ
Poisson	e^{θ}	$\theta = \ln \mu$	$2\sqrt{\mu}$	$3\sqrt[3]{\mu}$	$\frac{3}{2}\mu^{2/3}$
Binomial	$e^{\theta}/(1+e^{\theta})$	$\theta = \ln[\mu/(1-\mu)]$	$2\sin^{-1}\sqrt{\mu}$	$I_{\mu}(\frac{1}{3},\frac{1}{3})$	$I_{\mu}(\frac{2}{3},\frac{2}{3})$
Gamma	$-1/\theta$	$\theta = -1/\mu$	$2\ln \mu$	$-3\mu^{-1/3}$	$-3\mu^{1/3}$
Inverse Gaussian	$(-2\theta)^{-1/2}$	$\theta = 1/2\mu^2$	$-2/\sqrt{\mu}$	$-1/\sqrt[4]{\mu}$	$\ln \mu$

Table 4.5: Link functions for key values of $\alpha(\delta)$.

$$I_{\mu}(a,b) = \int_0^{\mu=\pi} x^{a-1}(1-x)^{b-1}dx, \quad a,b>0, \ 0 \leq x \leq 1.$$

The special cases for GLMs shown in Table 4.5 have parallels in the choice of transformations for general models. Box and Cox (1964) introduced a family of transformations using the likelihood to best select transformations satisfying optimality criteria. Anscombe (1948) considered the choice of transformations for non–Normal distributions. The choices given in Table 4.5 show the possibilities for choice of link function against desirable properties. In practice, the data should decide which link functions are appropriate.

Note

For the special case $\alpha = 1$ ($\delta = 0$) the exponential (or Efron) connection recovers the canonical form with implied sufficiency. In this case

$$\overset{1}{\Gamma}_{i'j'k'} = G_{Ii'}G_{Jj'}G_{Kk'}g_{jk}\left(\frac{\partial^2 \mathsf{f}_j}{\partial\eta_I\partial\eta_J}\frac{\partial \mathsf{f}_k}{\partial\eta_K}\right)$$

and

$$\overset{1}{\Gamma}_{i'j'k'} = 0 \quad \rightsquigarrow g_{jk}\left(\frac{\partial^2 \mathsf{f}_j}{\partial\eta_I\partial\eta_J}\right) = 0$$

which in turn implies

$$L_{i'j'}L_{k'} = 0.$$

This condition holds if the family is exponential in the parameter, ie, canonical in the parameter represented by the parameterization. In this case the corresponding link function is the canonical link function for the error distribution involved.

Chapter 5

Extensions and Conclusion

In this chapter several allied problems are examined in detail to demonstrate the utility of generalized curvature measures in the statistical approach to data analysis. Finally an overview of the thesis highlights is presented with concluding remarks.

5.1 Extensions

Two related areas are studied to demonstrate the use of generalized curvature measures both directly and indirectly in analysing statistical problems. The use of these curvature measures will be non–technical in order to give a *raison d'être* for such measures without lengthy algebraic discourses.

5.1.1 Leverage in Nonlinear Regression

In the spirit of Bates and Watts (1981), a reformulation of the local Taylor's series approximation to the nonlinear regression problem converts it to a GLM using square root link and an offset. As for the Bates and Watts quadratic form, the GLM local approximation of the solution locus provides a better local approximation than linearization, although strange behaviour can occur far away from the final value, as shown in Figure 5.1. The rationale for this approach is in considering

estimates of leverage. The linearization and the GLM must give differing leverages since they are effectively different approximants to the true nonlinear model, as can be seen from the diagram (Figure 5.1) of the solution locus for Test Problem 1. Of course, the two methods will give identical results at the optimum, ie, the least squares solution. The point of including this GLM algorithm for general nonlinear regression is to demonstrate implicitly the effect of curvature on leverage by approximating the same nonlinear regression model by two different approximants, linearization and the GLM quadratic approximation. As the two approximants of the same nonlinear regression function have different intrinsic curvatures (zero for linearization and non–zero for the GLM with square root link), then the leverages must be different (see Test Problem 2). Correspondingly parameter–effects curvature would be zero for the linearization approximation, and non–zero but invariant for the GLM. Given the nature of the GLM approximant, the GLM parameter–effects would be expected to be closer to that of the nonlinear regression model than the nil estimate from linearization.

The Problem

The nonlinear regression problem can be stated as the estimation of $\boldsymbol{\theta}$ given data (Y_i, X_i), $i = 1 \ldots n$. The general form of model to be fitted to the data is

$$\boldsymbol{Y} = f(\boldsymbol{X}; \boldsymbol{\theta}) + \boldsymbol{\varepsilon} \ , \ \boldsymbol{\varepsilon} \sim NIID(0, \sigma^2)$$

For a linear model

$$f(\boldsymbol{X}; \boldsymbol{\theta}) = \boldsymbol{X\theta}$$

meaning that the derivative

$$\partial f / \partial \boldsymbol{\theta} = \boldsymbol{X}$$

is independent of $\boldsymbol{\theta}$. A nonlinear model is one where the derivative

$$\partial f / \partial \boldsymbol{\theta} = F(\boldsymbol{\theta})$$

depends on $\boldsymbol{\theta}$. A special case of a nonlinear model is a Generalized Linear Model (GLM) for which

$$f(\boldsymbol{X};\boldsymbol{\theta}) = f(\boldsymbol{X}\boldsymbol{\theta}).$$

Using the criterion of Least Squares, the minimum of the function

$$S = \sum [Y - f(\boldsymbol{X};\boldsymbol{\theta})]^2$$

corresponds to the solution of the normal equations

$$\frac{\partial S}{\partial \boldsymbol{\theta}} = 2\sum [Y - f(\boldsymbol{X};\boldsymbol{\theta})]\,[-\frac{\partial f}{\partial \boldsymbol{\theta}}] = 0\,.$$

Only for linear models do these equations have an analytic solution, so alternative methods of estimation are used. All of these methods attempt to solve the minimisation problem

$$\overset{\text{Min}}{\boldsymbol{\theta}} \sum [Y - f(\boldsymbol{X};\boldsymbol{\theta})]^2\,.$$

Two of these methods are described briefly, viz,

1. Linearization (Gauss-Newton), and
2. Newton-Raphson.

Linearization

The nonlinear function is approximated by a linear model using the Taylor's expansion on

$$\boldsymbol{Y} = f(\boldsymbol{X};\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

to give

$$f(\boldsymbol{X};\boldsymbol{\theta}) \approx f(\boldsymbol{X};\boldsymbol{\theta}_0) + \frac{\partial f}{\partial \boldsymbol{\theta}}^{\top} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \ldots$$

which becomes

$$\boldsymbol{Y} - f_0 \approx \frac{\partial f}{\partial \boldsymbol{\theta}_0}^{\top} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}.$$

Converting the problem to Ordinary Least Squares (OLS) yields

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Y} - f_0 \\ \boldsymbol{Z} &= \frac{\partial f}{\partial \boldsymbol{\theta}_0} \\ \boldsymbol{\beta} &= \boldsymbol{\theta} - \boldsymbol{\theta}_0 \end{aligned}$$

The fitting algorithm becomes

- Choose $\boldsymbol{\theta}_0$.
- Regress $\boldsymbol{y}$ on $\boldsymbol{Z}$ to get $\widehat{\boldsymbol{\beta}}$, i.e., $\underset{\boldsymbol{\beta}}{\text{Min}} \sum(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^2$.
- $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\beta}} + \boldsymbol{\theta}_0$
- Iterate until $\widehat{\boldsymbol{\beta}}$ is trivial.

Newton–Raphson Method

This quadratic method expands the function

$$S(\boldsymbol{\theta}) = \sum [Y - f(\boldsymbol{X}; \boldsymbol{\theta})]^2$$

in a Taylor's series

$$S(\boldsymbol{\theta}) = S(\boldsymbol{\theta}_0) + \frac{\partial S}{\partial \boldsymbol{\theta}_0}^{\top} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\top} \frac{\partial^2 S}{\partial \boldsymbol{\theta}_0^2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \ldots$$

$$\frac{\partial S}{\partial \boldsymbol{\theta}} = 0 \rightarrow \frac{\partial S}{\partial \boldsymbol{\theta}_0} + \frac{\partial^2 S}{\partial \boldsymbol{\theta}_0^2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) = 0$$

$$\boldsymbol{g} + \boldsymbol{H}\boldsymbol{\beta} = 0 \rightarrow \widehat{\boldsymbol{\beta}} = -\boldsymbol{H}^{-1}\boldsymbol{g}$$

In general

$$\frac{\partial^2 S}{\partial \boldsymbol{\theta}^2} = 2 \frac{\partial f}{\partial \boldsymbol{\theta}}^{\top} \frac{\partial f}{\partial \boldsymbol{\theta}} - 2(\boldsymbol{Y} - \boldsymbol{f})^{\top} \frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}$$

ie., the first term corresponds to linearization.

GLM Variant

The nonlinear function f in

$$\boldsymbol{Y} = f(\boldsymbol{X};\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

is approximated by

$$\boldsymbol{Y} \approx \boldsymbol{f}_0 + \frac{\partial f}{\partial \boldsymbol{\theta}_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}\frac{\partial^2 f}{\partial \boldsymbol{\theta}_0^2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^2$$

to become

$$\boldsymbol{y} = \boldsymbol{Y} - \boldsymbol{f}_0 = \boldsymbol{Z}\boldsymbol{\beta} + \tfrac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{H}\boldsymbol{\beta} + \ldots$$

where

$$\boldsymbol{H} = \frac{\partial^2 f}{\partial \boldsymbol{\theta}_0^2} = \boldsymbol{V}^\top \boldsymbol{V}$$

and

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$$

as for linearization. Completing the square gives

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{\beta}$$

$$= \boldsymbol{Z}\boldsymbol{\beta} + \left(\frac{\boldsymbol{V}}{\sqrt{2}}\boldsymbol{\beta}\right)^\top \left(\frac{\boldsymbol{V}}{\sqrt{2}}\boldsymbol{\beta}\right)$$

$$= \left(\mathcal{A}_0 + \frac{\boldsymbol{V}}{\sqrt{2}}\boldsymbol{\beta}\right)^\top \left(\mathcal{A}_0 + \frac{\boldsymbol{V}}{\sqrt{2}}\boldsymbol{\beta}\right) - \mathcal{A}_0{}^\top \mathcal{A}_0$$

where

$$\mathcal{A}_0 = \frac{\boldsymbol{Z}\boldsymbol{V}^{-1}}{\sqrt{2}}.$$

The GLM approximant is then

$$\boldsymbol{Y} - \boldsymbol{f}_0 + \mathcal{A}_0{}^\top \mathcal{A}_0 = \left(\mathcal{A}_0 + \frac{\boldsymbol{V}}{\sqrt{2}}\boldsymbol{\beta}\right)^\top \left(\mathcal{A}_0 + \frac{\boldsymbol{V}}{\sqrt{2}}\boldsymbol{\beta}\right)$$

ie., the response is $\boldsymbol{Y} - \boldsymbol{f}_0 + \mathcal{A}_0{}^\top \mathcal{A}_0$, and the predictor is $\frac{\boldsymbol{V}}{\sqrt{2}}$, with $\mathcal{A}_0$ an *offset* in GLIM parlance. The *link* function is the square root.

Test Problem 1

The data used in this test problem is from Table 1.1 as shown in Figure 1.1. The three GLIM outputs given in Appendix D.1 show

- a one–step implementation of the GLM variant,
- the iterative form of the GLM variant, and
- a linearization procedure (one–step) for comparison with the GLM variant.

The GLIM variant and linearization concur at the optimum value for the parameter estimate, as shown in Table 5.1.

	GLM	NLR
$\widehat{\theta}$	2.0537	2.0537
$SE(\widehat{\theta})$	0.1573	0.1575
Deviance	2.9334	2.9334
Residuals	(-1.652, 0.453)	

Table 5.1: Summary output : Test problem 1

Figure 5.1 shows the solution locus as the solid curve (with circles at $\theta = 0, 1, 2$), the tangent to the solution locus at $\theta_0 = 0$, and $\theta = 2$ being in the top right hand corner of the graph. The GLM approximant to the solution locus is also pivoted at $\theta_0 = 0$ and is shown by crosses (+), while the data are shown by the box(□). Globally, neither the tangent nor the GLM approximant capture the full nature of the solution locus. However, in the neighbourhood of the final estimate ($\theta = 2$), the GLM variant is closer to the solution locus than the tangent produced by the linearization method. The quadratic nature of the GLM approximant is clearly demonstrated.

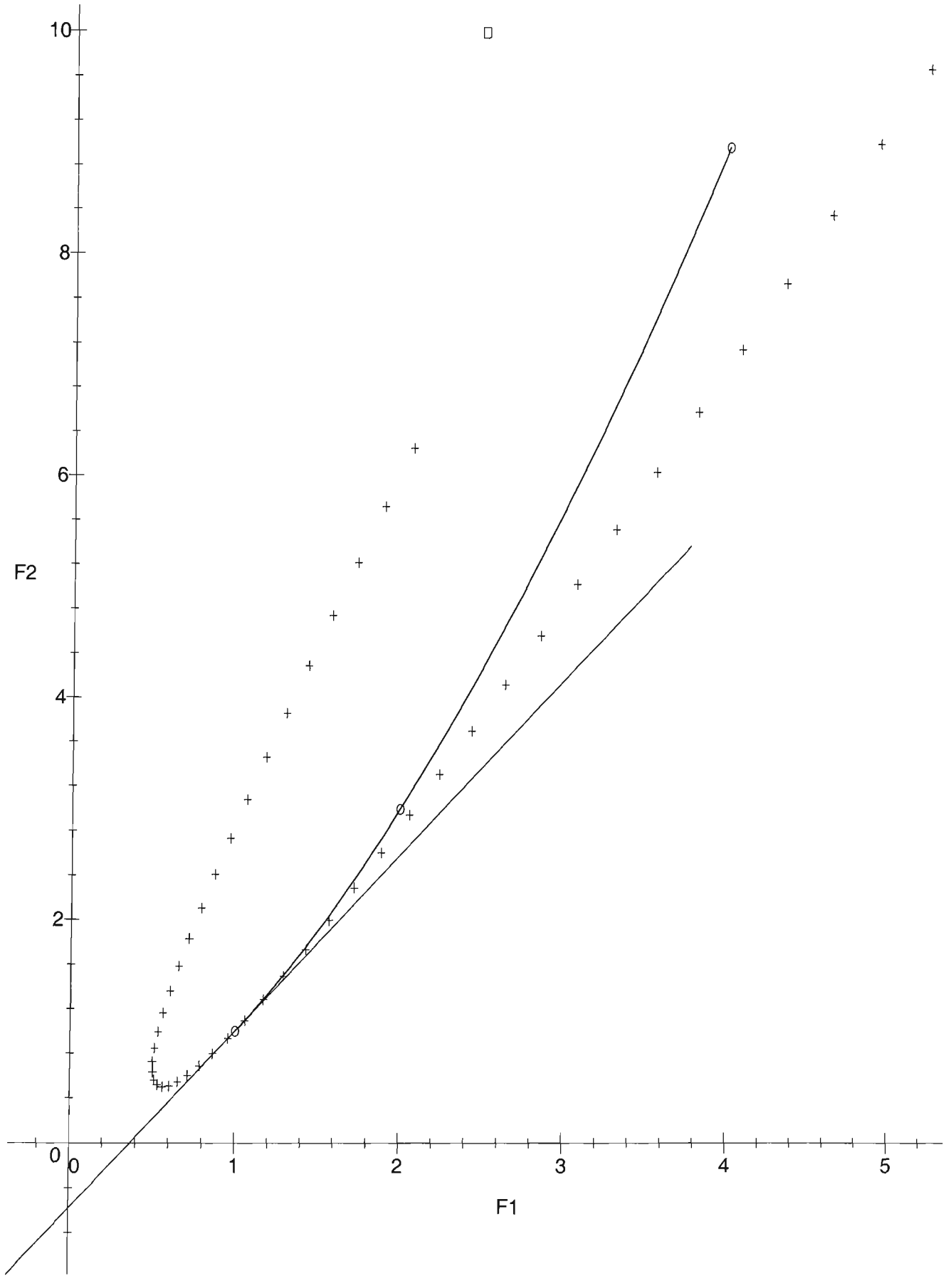

Figure 5.1: Solution Locus (solid curve), Tangent (line) and GLM approximant (crosses [+]). The data are shown by the box(□).

Test Problem 2

The following data are from Draper and Smith (1981, p517, Exercise B).

t	Y
0.5	0.96,0.91
1	0.86,0.79
2	0.63,0.62
4	0.48,0.42
8	0.17,0.21
16	0.03,0.05

Table 5.2: Data Set with replication

The model to be fitted is

$$E(Y) = e^{-\theta t}$$

assuming Normal errors. In the treatment that follows, the GLM variant on non-linear regression is contrasted with linearization, at parameter values away from the optimum (mle). In particular, *leverages* are obtained (by use of the GENSTAT 5 package), initially at the optimum parameter value, and then at parameter values two standard deviations away from the optimum value. The corresponding computer outputs are given in Appendix D.2.

The key results of these three sets of calculations are summarised in Table 5.3. The 'Interval' quoted in Table 5.3 is simply two standard errors.

Observed Estimate		
	GLM	Linearization
Upper	0.01618	0.01672
Lower	- 0.01623	- 0.01547
Interval = 0.01618 = 2 · 0.00809		

Standard Error		
	GLM	Linearization
Upper	0.00812	0.00873
Lower	0.00813	0.00746
SE(optimum) = 0.00809		

Table 5.3: Results summary : Data Set with replication

The corresponding leverages are reproduced in Table 5.4.

Optimum	+ 2 SEs		- 2 SEs	
GLM/NLR	GLM	NLR	GLM	NLR
0.012	0.012	0.014	0.012	0.011
0.012	0.012	0.014	0.012	0.011
0.040	0.040	0.045	0.040	0.035
0.040	0.040	0.045	0.040	0.035
0.105	0.105	0.114	0.106	0.095
0.105	0.105	0.114	0.106	0.095
0.183	0.184	0.187	0.184	0.177
0.183	0.184	0.187	0.184	0.177
0.140	0.139	0.126	0.139	0.154
0.140	0.139	0.126	0.139	0.154
0.020	0.019	0.014	0.019	0.029
0.020	0.019	0.014	0.019	0.029

Table 5.4: Leverages summary : Data set with replication

The following observations can now be made.

1. The leverages for the GLM variant at parameter values two standard deviations above and below the optimum are closer to their values at the optimum than are the corresponding leverages values for linearization, by inspection of the final column in each output.

2. The GLM regression estimates at the extremities are closer to the interval defined by two standard errors than are the corresponding regression estimates from linearization.

3. The standard errors for linearization at the extremities appear more variable than the GLM variant.

Given that the GLM variant is expected in theory to approximate the solution locus better than linearization, all these results are to be expected. These empirical results confirm the suggestion that the GLM variant will better approximate the nonlinear regression model than will a simple linear expansion.

5.1.2 Replication and Curvature

The question of replication in observational studies and experimental designs has been addressed by many workers, such as Draper and Smith (1981), Seber and Wild (1989) and Weisberg (1985). The statistical benefits of replication include increased precision of estimates, an independent measure of error, and the ability to test for interaction, as well as allowing for lack of fit tests. Bates and Watts (1980, p5, 2.2.1) considered the effect of replication on estimates of curvature for the nonlinear regression model. An r–fold replication reduces all curvatures by a factor of $\sqrt{1/r}$, following the arguments of Bates and Watts (1980) and Seber and Wild (1989). The explanation proceeds by observing that with replication the problem of fitting a model to the data reduces to fitting to the means, Seber and Wild (1989, p31)[1]. Thus the expected value for each replicate observation holds a carbon copy of the fitted values using the means as data. So having two replicates at each of three design points gives

$$\dot{\mu}_1 = (\dot{\mu}_1, \dot{\mu}_1)$$

[1]The replication is assumed to be the same at each design point. This restriction is not necessary, but it helps to simplify the discussion.

$$\dot{\mu}_2 = (\ddot{\mu}_2, \dot{\ddot{\mu}}_2)$$

$$\dot{\mu}_3 = (\ddot{\mu}_3, \dot{\ddot{\mu}}_3)$$

or

$$\dot{\boldsymbol{\mu}} = (\ddot{\boldsymbol{\mu}}, \dot{\ddot{\boldsymbol{\mu}}})$$

where the notation $\bar{\mu}$ refers to the fitted value obtained by regression using means rather than all the observations. In the notation of Seber and Wild (1989, p146, 4.2.5), curvature is

$$\gamma_h = \frac{||\dot{\boldsymbol{\mu}}||}{||\ddot{\boldsymbol{\mu}}||^2} = \frac{\sqrt{2}||\dot{\bar{\boldsymbol{\mu}}}||}{2||\ddot{\bar{\boldsymbol{\mu}}}||^2} = \frac{||\dot{\bar{\boldsymbol{\mu}}}||}{\sqrt{2}||\ddot{\bar{\boldsymbol{\mu}}}||^2} = \frac{\bar{\gamma}_h}{\sqrt{2}}$$

As stated earlier, under replication the estimation problem is tantamount to regressing on the means (Seber and Wild, 1989, p31), and the within replication variability is used to obtain an estimate of pure error. Switching to a regression based on the means induces a scale factor of $1/\sqrt{r}$ since $V(\bar{x}) = V(x)/r$, in agreement with the above analysis.

These developments are mirrored in the modifications due to Amari (1982a, p372, 4.3 and p376, 5.2) for the metric, affine connection and skewness tensor in the case of replicated observations in the general case. For example, the metric tensors are related by

$$N g_{ij} = \bar{g}_{ij}$$

where $\bar{g}_{ij}$ is the metric tensor for the problem cast in terms of fitting to the data from N replicate observations, and g_{ij} is the metric tensor for the variable on the original (single observation) scale.

Example

To demonstrate the effect of increased levels of replication, the following simple experiment was conducted. For the one–parameter model $E(Y) = \sqrt{x}$, the data points shown in Table 5.5 were chosen.

x	$E(Y)$
1	1
4	2
9	3

Table 5.5: Square root model – replication experiment

For sample sizes of 1, 2, 5 and 100 at each level of x, noise was generated from the Uniform distribution between -1 and 1. To emulate the effect of increasing replication, this noise was averaged and then added to the $E(Y)$ value. For each sample size, this procedure was repeated 100 times using different simulated data each time. A sample of generated data is shown in Table 5.6, being the last data set out of the 100 generated for each value of N. These are the data sets that are described in the series of plots that follow.

N	y_1	y_2	y_3
1	1.245 960	2.584 084	3.865 286
2	0.783 352	2.094 020	2.953 216
5	1.327 645	1.991 387	3.005 948
100	1.050 385	2.015 503	2.979 930

Table 5.6: 'Typical' data generated for the replication experiment

The function $E(Y) = x^{\theta}$ was fitted to each of these 100 data sets, assuming Normal errors. The average results for the 100 simulation are shown in Table 5.7.

In order to display the 'typical' results, the sum of squares was plotted against the parameter value. The plot was centred on the final estimate, using a width of

N	$\widehat{\theta}$ (average)	Standard Error
1	0.492 832	0.078 655
2	0.495 235	0.053 578
5	0.498 038	0.031 646
100	0.500 015	0.007 610

Table 5.7: Results(averages) for the simulation replication experiments

two standard errors on either side of the final estimate. These final estimates and their standard errors were obtained from fitting the model to the representative data given in Table 5.6. These estimates and standard errors (SE) are shown in Table 5.8 which shows the estimates and their corresponding standard errors for the last data set in each of the 100 simulations using increasing replication ($N = 1, 2, 5, 100$).

N	Estimate	SE
1	0.6245	0.0252
2	0.4991	0.0238
5	0.5003	0.0324
100	0.4983	0.0055

Table 5.8: Results for the 'typical' data

These plots are shown in Figures 5.2, 5.3, 5.4 and 5.5. For each graph the sum of squares function (SoS) based on the nonlinear model is shown by the solid line, while the sum of squares function based on the linear approximation centred

at the final estimate is shown by the curve using circles. The scale for the last graph($N = 100$) is different to the remainder($N = 1, 2, 5$).

The effect of increasing replication is twofold. The approximation of the sum of squares surface based on the linearization around the final parameter estimate improves with increasing replication, and the change in the sum of squares surface for both the 'true' value (based on the nonlinear function) and that based on the linear approximation decreases relatively with increasing replication. The second effect will correspond to decreasing curvatures, or to use the terminology of Ratkowsky (1983), being 'close to linear'. In expectation space, the corresponding solution locus for the model appears as a space curve, which can be displayed in 3 dimensions by using the means as expectation coordinates as in Figure 5.6. The solution locus as shown in Figure 5.6 is centred on the true value of $\theta = 0.5$ with the thin line corresponding to the range of parameter values from $-1/2$ and 1, while the thick line gives the confidence interval expected with replication ($N = 2$). No data point is shown since the solution locus is representative of all simulations described in Table 5.8. The 'close–to–linear' behaviour is clearly exhibited, in the straightness of the solution locus near the optimum.

General Notation

Following the developments in Seber and Wild (1989, pp30–32), the replication problem could be cast in terms of the means at each replicate point. However, as shown in Appendix D.3, the metric tensor and related functions such as the α–connection can be written as multiples of the corresponding functions in the single observation case. This result holds for the problem cast in terms of the replicate observations or in terms of the means, but the former tends to be used most often. Following Amari (1990), the notation $\bar{g}$ will be used to denote the metric tensor based on N replicates. The corresponding α–connection for the replicate observations will be denoted by $\overset{\alpha}{\bar{\Gamma}}$. This terminology is not to be confused with ($\bar{\Gamma} \equiv \overset{0}{\Gamma}$) which is used for the information (Riemannian) connection. See Lauritzen (1987). Some basic results for metric tensors, α–connections and other

quantities are given in Appendix D.3, which also contains an explanation of some results collected from various sources which elucidate the relations used later. Briefly, using the notation above and definitions from Appendix D.3,

$$\bar{g}_{ij} = E\partial_i\bar{\ell}(\boldsymbol{y};\boldsymbol{\theta})\partial_j\bar{\ell}(\boldsymbol{y};\boldsymbol{\theta}) = \sum E\partial_i\ell\partial_j\ell = Ng_{ij}$$

and

$$\overset{\alpha}{\bar{\Gamma}}_{ijk} = N\overset{\alpha}{\Gamma}_{ijk}$$

for the replicate data. So all subsequent discussions could proceed in terms of the single observation scale if desired.

Asymptotics

The form of the Central Limit Theorem given in Amari (1982a, p376, 5.2) shows the effect of increased levels of replication, since effectively for $N \equiv r$

$$\bar{g}_{ij} = Ng_{ij}.$$

In the limit, for the nonlinear regression model, high levels of replication mean that the nonlinear model behaves locally as a linear model as both intrinsic and parameter–effects will converge to zero. The local linear approximation to the nonlinear function will be excellent, with the means mapping out the deterministic component precisely, and the population error will be known without measurement error. These features are exploited in the lack of fit test, used not only for testing a linear model, but also for testing a nonlinear response function, (Seber and Wild, 1989, p32).

To study the general case, the imbedding theorem needs to be invoked. This theorem expresses the α-connection for the regression coefficients $\overset{\alpha}{\Gamma}_{abc}$ in terms of the α–connection of the natural parameters $\overset{\alpha}{\Gamma}_{ijk}$, namely

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = (\partial_a B^i_b)B^j_c g_{ij} + B^i_a B^j_b B^k_c \overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}(\boldsymbol{u})) .$$

Using the bar notation already described, this relation can now be written in terms of means rather than individual data points.

$$\bar{\overset{\alpha}{\Gamma}}_{abc} = (\partial_a B_b^i) B_c^j \bar{g}_{ij} + B_a^i B_b^j B_c^k \bar{\overset{\alpha}{\Gamma}}_{ijk} \quad . \tag{5.1}$$

Writing the results (5.2) of Amari (1982a) in the notation of this thesis, the Central Limit Theorem gives the distribution of the data $\bar{\boldsymbol{y}}$ as asymptotically normal,

$$\bar{\boldsymbol{y}} \sim N(\boldsymbol{\mu}, g_{ij}/r).$$

Examination of the imbedding theorem, Equation (5.1), now yields two distinct possibilities

(i) $\bar{\overset{\alpha}{\Gamma}}_{ijk} \to 0$, and/or

(ii) $\partial_a B_b^i \to 0$

as $r \to \infty$.

The first case (i) implies that, as all error distributions converge to the Normal under intensive replication, all such models will become nonlinear regression models, if only the error distribution is considered. This is due to $\bar{\overset{\alpha}{\Gamma}}_{ijk} = 0$ in the case of Normal errors. As the intrinsic curvature also goes to zero under high levels of replication, the nonlinear regression model will be locally well approximated by a linear model.

In the second case (ii), the exponential connection becomes

$$\bar{\overset{e}{\Gamma}}_{abc} = (\partial_a B_b^i) B_c^j \bar{g}_{ij} + B_a^i B_b^j B_c^k \bar{\overset{e}{\Gamma}}_{ijk}$$

but $\bar{\overset{e}{\Gamma}}_{ijk} = N \bar{\overset{e}{\Gamma}}_{ijk} = 0$ by definition for an exponential family model. Thus if $\partial_a B_b^i \to 0$ then $\bar{\overset{e}{\Gamma}}_{abc} \to 0$ implying that a generalized nonlinear model (GNM) will be a local exponential family model in terms of the regression coefficients, ie., a

GLM with canonical link. Sufficiency of the particular parameters follows as a local property in the limit[2].

The two conditions are connected by the Central Limit Theorem. The subsuming case is nonlinear regression. Furthermore, as $N \to \infty$ the means end up on the solution locus, and the linear model approximates the nonlinear model well in a local sense. In differential geometric terms, the expectation surface becomes 'locally Euclidean'. For the general case, ie, a GNM, this implies (in the notation of Section 3.6)

$$p(\boldsymbol{X};\boldsymbol{\beta}) \approx p_0(\boldsymbol{X};\boldsymbol{\beta}_0) + \frac{\partial p}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}_0} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

But the natural parameters are given by

$$\theta_i = \mathsf{f}(\boldsymbol{X};\boldsymbol{\beta})$$

and $p \rightsquigarrow \mathsf{f}$ for Normal errors, giving

$$B^i_b = \frac{\partial \theta^i}{\partial \beta^b} = \frac{\partial \mathsf{f}_i}{\partial \beta^b}.$$

The Taylor's expansion for f gives

$$\mathsf{f} = \mathsf{f}_0 + \frac{\partial \mathsf{f}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}_0} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

which means that

$$B^i_b = \text{constant}$$

leading to

$$\partial_a B^i_b = 0.$$

This means that the GNM will be locally approximated by a GLM with canonical link. However the same conditions that could cause this to happen would also mean that the GNM would become a nonlinear regression model in the limit.

[2]All that can be inferred for general likelihood models using (ii) alone is that local sufficiency holds for the (imbedded) regression parameters.

5.2 Overall Results

The following results are not entirely original but are derived in detail using procedures not presented elsewhere.

- The use of Bartlett's equations in the interpretation of α–connections.
- The projection of normal and tangential components of α–curvature.
- The decomposition of (scalar) α–curvature.
- Zero 1–connections and exponential families.
- Wedderburn's exponential form.
- For GLMs, an expanded table of link functions corresponding to transformation properties.

The following is a summary of points that are considered to be original.

- The α–connections in the multi–parameter case are interpreted, especially for α equal to zero, 1/3 and 1.
- The Exponential connection is zero iff the link function is canonical in a GLM.
- A test for canonical link adequacy in GLMs has been derived from the skewness tensor as imbedded in the α–connection for the regression coefficients.
- The invariance of intrinsic curvature is proved for the general case of non–Normal errors.
- Parameter-effects curvature is shown to be invariant for a GLM ; (long and short forms).
- The scalar form of exponential intrinsic curvature for a GLM is minimal when the link is canonical.

- A generalized nonlinear model (GNM) with zero exponential curvature is a GLM with canonical link.
- A zero information connection implies a variance stabilizing link in a GLM and conversely.
- An improved leverage estimator in nonlinear regression can be obtained via a GLM approximant to the nonlinear function.

5.2.1 Summary

Some of the results obtained using the methodology of differential geometry and tensor algebra represent known results in a new light. However this in itself could prove worthwhile as being a new way of viewing established relations, as stated in Kass(1989). For example, the statistical interpretation of α–connections joined the differential geometric methods of Amari (1982a) to the approach used by Bartlett (1953a). In the multi–parameter case discussion of the interpretation of these connections required the Bartlett notation to be modified and extended. Other key results that have been reported elsewhere are the projection of α–curvature(Amari, 1982a, p371), and its decomposition into normal and tangential components (Amari, 1990, p156). While these results have been quoted previously, a full derivation and explanation appears not to have been given, as has been done here. The decomposition of α–curvature has also been derived in scalar terms as well.

Considering the new results, the invariance of intrinsic α–curvature may be expected to hold, but it is claimed that the method of proof using differential geometric arguments with tensors has not previously been demonstrated. Likewise the invariance of parameter–effects curvature for a GLM is an expected result, but the proof using the methods of Section 3.3.3 is claimed as original. Certainly the invariance of intrinsic curvature in general and parameter–effects for a GLM are both generalizations of results in special cases (Normal errors and/or linear models), as shown in the body of the thesis.

A major thrust of the investigation centred on the curved exponential family and in particular generalized linear models as an important subclass of such families. Some of these new results obtained using the differential geometric approach are necessarily cast in terms of the apparatus peculiar to the consequent view of statistical distributions, such as α–connections. Thus the specialised results for the exponential connection and canonical link in GLMs may seem abstract, but proper interpretation of this result requires a full appreciation of the role of affine connections in statistical distributions. To this extent this result could be viewed as defining the role of the exponential connection in GLMs. These investigations have led to the formulation of a new test for canonical link adequacy in GLMs by employing the consequent relation for skewness of regression coefficients in GLMs, as determined from the α–connection and the imbedding theorem. Likewise, interpreting the minimality of scalar exponential intrinsic curvature for a GLM with canonical link requires an understanding of α–curvature. However, this result can also be seen as a generalization of the zero intrinsic curvature for linear models under Normal errors.

A new class of models, generalized nonlinear models (GNMs), has been defined as the generalization of the nonlinear regression model in the case of Normal errors. These models inherit features of GLMs and the nonlinear regression model. In fact GNMs and GLMs are related families of models since it has been shown that a GNM with zero exponential curvature is a GLM with canonical link.

Considering link functions other than the canonical in GLMs, the most popular link function after the canonical (Kass and Smyth, 1990) is the variance stabilizing link function. This constant information scale link implies a zero information connection and conversely. While this result is hardly surprising, again the development in terms of the α–connection requires the ability to manipulate and interpret these affine connections. This type of result could also be used as an interpretation of the α–connection itself in the case of a GLM. By considering other estimator properties such as 'normal' likelihood and skewness reduction, an expanded set of link functions has been determined, giving the user further options

for the choice of link function over the canonical and the variance stabilizing.

Finally the concept of curvature has been used in demonstrating an improved method of estimating leverages in nonlinear regression, and the effects of replication on curvature have been examined, with a view to investigating asymptotic behaviour, ie, the results of intense replication at the design points. The consequent behaviour confirms the expected results for increasing sample sizes at each replicated design point, in terms of local approximation by a linear model based on Normal errors.

Conclusion

This thesis has presented a generalization of curvature measures for non–normal error models by continued analogy with the nonlinear regression model. This generalization and its subsequent interpretations have been shown to reduce to the known results for Normal errors, where the differential geometry is Euclidean rather than Riemannian as in the general case. In addition, the generalization to non–Normal models of effects of curvature on model behaviour has been expounded. A particular class of curved exponential family models, generalized linear models, has been the special topic of consideration, with attention being given to the canonical link. As this link function is the non–Normal analogue of the linear model for Normal errors, it has been constantly been used as the reference point for investigating differential geometric quantities such as α–connections and α–curvatures, with a view to generating interpretations of such quantities in the general case. This strategy has proved most fruitful in investigating the behaviour of generalized nonlinear models and generalized α–curvature for such models. The 'theme' of contrasting the canonical link with the non–canonical link in GLMs has been employed in the development of an empirical test of link adequacy and has also resulted in the classification of alternative link functions to the canonical in GLMs. In essence, this thesis has extended the differential geometric approach from Normal to non–Normal error models, not only for generalized linear models but also for models having a general response function.

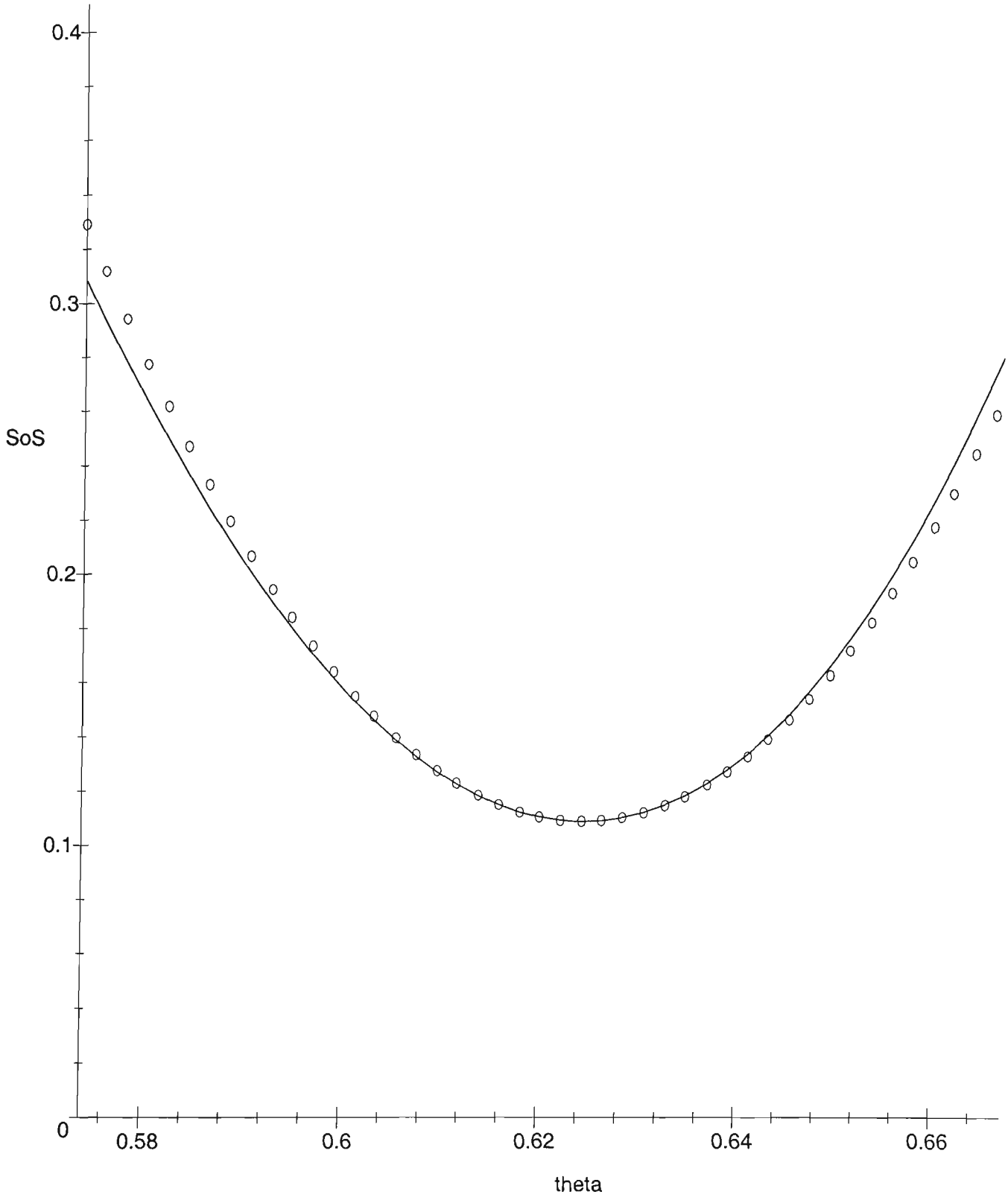

Figure 5.2: Sum of squares plotted against the parameter θ : $N = 1$

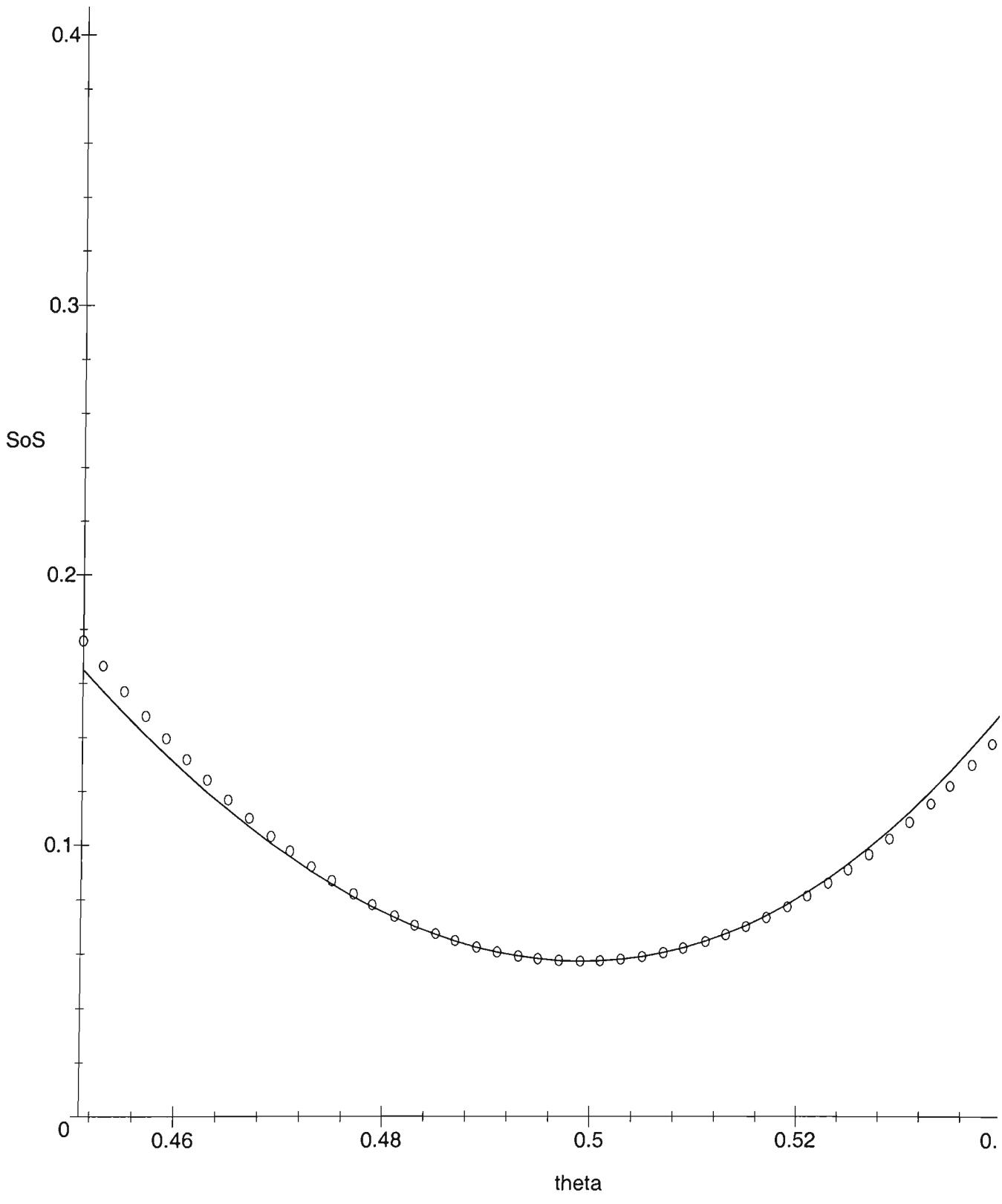

Figure 5.3: Sum of squares plotted against the parameter θ : $N = 2$

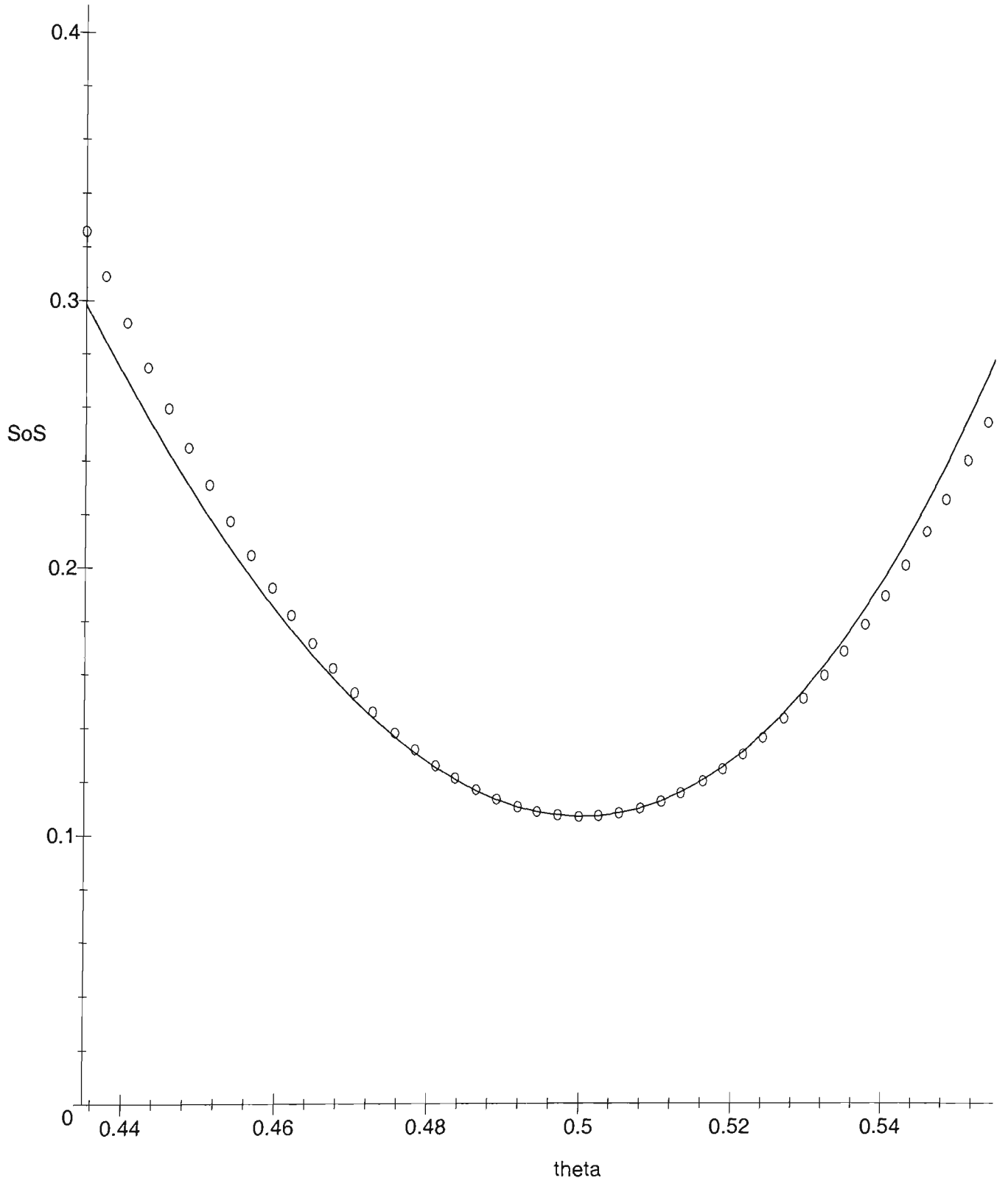

Figure 5.4: Sum of squares plotted against the parameter θ : $N = 5$

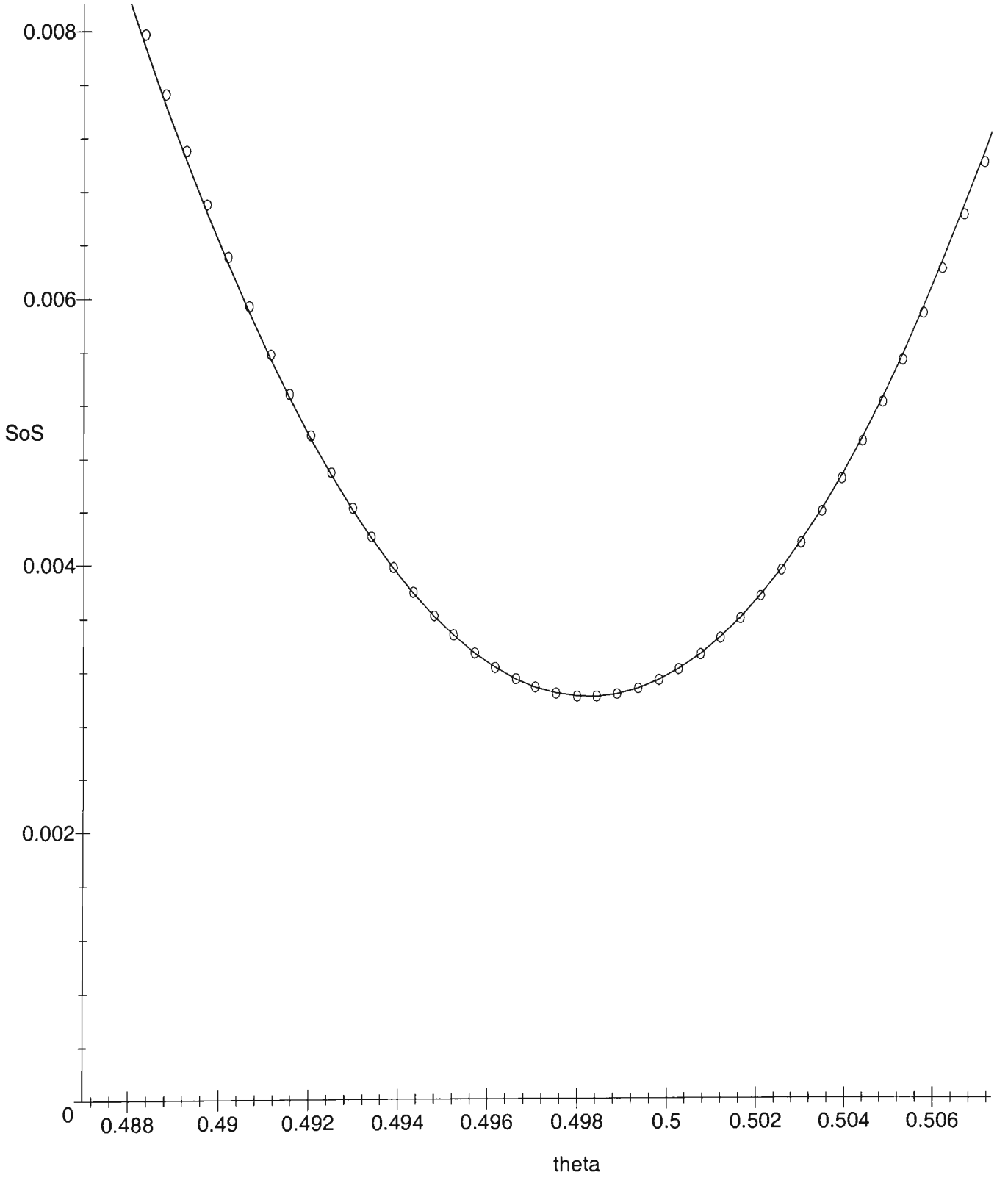

Figure 5.5: Sum of squares plotted against the parameter θ : $N = 100$

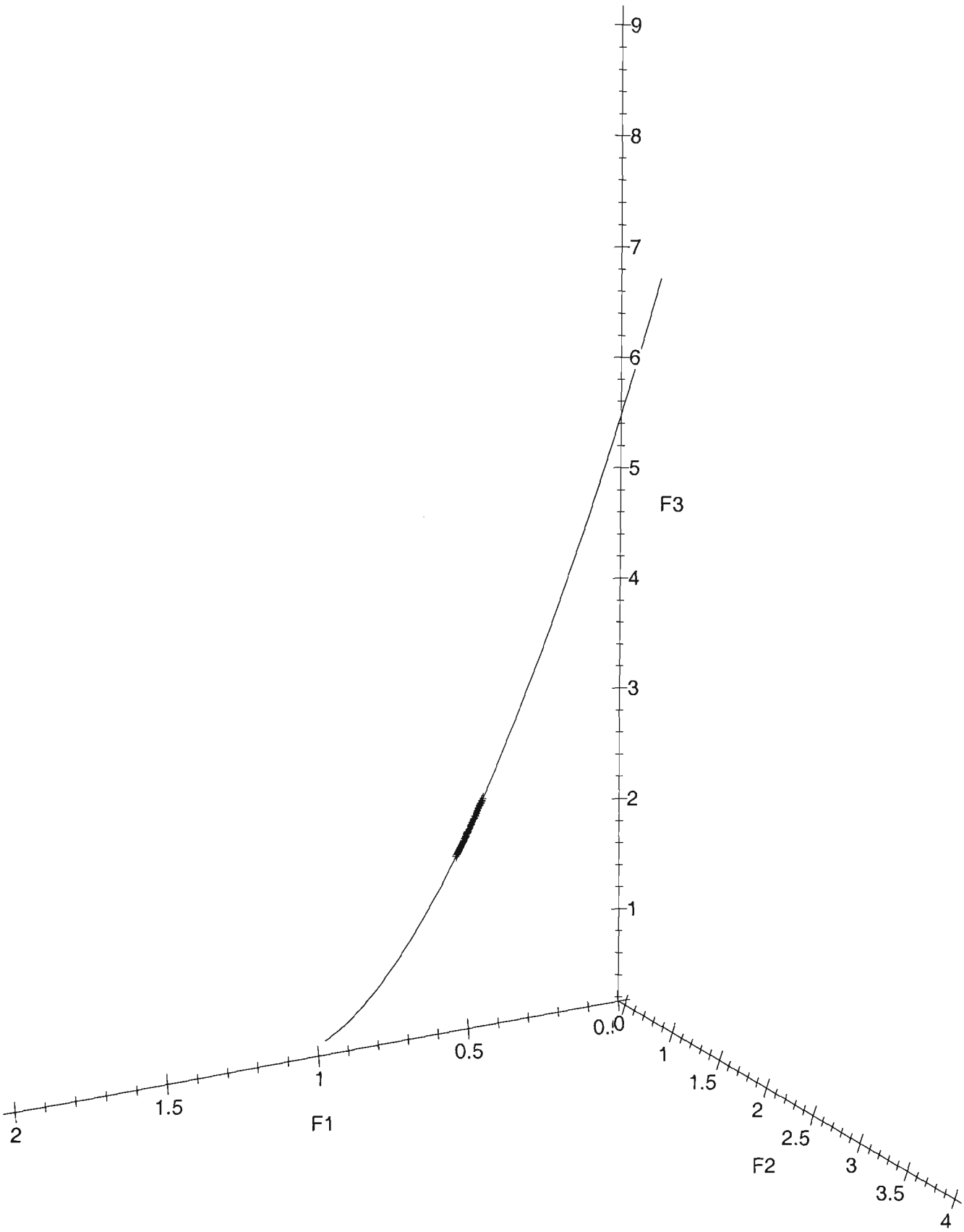

Figure 5.6: Solution locus : replication experiment

Appendix A

(Ch. 1)

A.1 The Hat Matrix for GLMs

The derivations below give the generalized forms of the leverage equation for GLMs which maps the data into the fitted values. Both forms are derived from the basic mapping equation for the working variate z from the GLIM algorithm, ie,

$$z = \eta + \left(\frac{d\eta}{d\mu}\right)(Y - \mu).$$

Since the predictors $\boldsymbol{X}$ are regressed onto $\boldsymbol{z}$ using weights $\boldsymbol{W}$, the equation involving the hat matrix for a GLM is

$$\boldsymbol{W}^{1/2}\hat{\boldsymbol{z}} = \boldsymbol{H_g}\boldsymbol{W}^{1/2}\boldsymbol{z}$$

where

$$\boldsymbol{H_g} = \boldsymbol{W}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{W}^{1/2}.$$

The scalar form for the weight function is

$$W = V^{-1}\left(\frac{d\mu}{d\eta}\right)^2.$$

A.1.1 Standardized Form

The working variate z (scalar) can be written as

$$z = \eta + W^{-1/2}V^{-1/2}(Y - \mu)$$

which becomes (in vector form)

$$\boldsymbol{W}^{1/2}(\boldsymbol{z}-\boldsymbol{\eta}) = \boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu}).$$

Pre–multiplying by $\boldsymbol{H}_g$ gives

$$\boldsymbol{H}_g\boldsymbol{W}^{1/2}(\boldsymbol{z}-\boldsymbol{\eta}) = \boldsymbol{H}_g\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu})$$

The LHS becomes

$$\boldsymbol{H}_g\boldsymbol{W}^{1/2}\boldsymbol{z} - \boldsymbol{H}_g\boldsymbol{W}^{1/2}\boldsymbol{\eta} = \boldsymbol{W}^{1/2}\hat{\boldsymbol{z}} - \boldsymbol{W}^{1/2}\boldsymbol{\eta}$$

using $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ and the leverage equation $\boldsymbol{W}^{1/2}\hat{\boldsymbol{z}} = \boldsymbol{H}_g\boldsymbol{W}^{1/2}\boldsymbol{z}$. This leads to

$$\boldsymbol{W}^{1/2}(\hat{\boldsymbol{z}}-\boldsymbol{\eta}) = \boldsymbol{H}_g\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu}).$$

But, the expected working variate is

$$\hat{\boldsymbol{z}} = \boldsymbol{\eta} + \boldsymbol{W}^{-1/2}\boldsymbol{V}^{-1/2}\left(\widehat{\boldsymbol{Y}}-\boldsymbol{\mu}\right)$$

since $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ are fixed for each iteration, and thus

$$\boldsymbol{W}^{1/2}(\hat{\boldsymbol{z}}-\boldsymbol{\eta}) = \boldsymbol{W}^{1/2}\boldsymbol{W}^{-1/2}\boldsymbol{V}^{-1/2}\left(\widehat{\boldsymbol{Y}}-\boldsymbol{\mu}\right) = \boldsymbol{H}_g\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu}).$$

Finally

$$\boldsymbol{V}^{-1/2}\left(\widehat{\boldsymbol{Y}}-\boldsymbol{\mu}\right) = \boldsymbol{V}^{-1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) = \boldsymbol{H}_g\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu})$$

as quoted.

A.1.2 Raw Form

Substituting the expansions for the working variate $\boldsymbol{z}$ in the leverage equation

$$\boldsymbol{W}^{1/2}\hat{\boldsymbol{z}} = \boldsymbol{H}_g\boldsymbol{W}^{1/2}\boldsymbol{z}$$

yields

$$\boldsymbol{W}^{1/2}\left(\boldsymbol{\eta} + \boldsymbol{W}^{-1/2}\boldsymbol{V}^{-1/2}(\widehat{\boldsymbol{Y}}-\boldsymbol{\mu})\right) = \boldsymbol{H}_g\boldsymbol{W}^{1/2}\left(\boldsymbol{\eta} + \boldsymbol{W}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu})\right).$$

Since $\boldsymbol{H}_g\boldsymbol{W}^{1/2}\boldsymbol{\eta} = \boldsymbol{W}^{1/2}\boldsymbol{\eta}$, this becomes

$$\boldsymbol{V}^{-1/2}\left(\widehat{\boldsymbol{Y}} - \boldsymbol{\mu}\right) = \boldsymbol{H}_g\boldsymbol{V}^{-1/2}\left(\boldsymbol{Y} - \boldsymbol{\mu}\right).$$

Since $\boldsymbol{V}^{-1/2} \propto \boldsymbol{W}^{1/2}$ then it follows that

$$\boldsymbol{V}^{-1/2}\boldsymbol{\eta} = \boldsymbol{H}_g\boldsymbol{V}^{-1/2}\boldsymbol{\eta}$$

leading to

$$\boldsymbol{V}^{-1/2}\boldsymbol{\mu} = \boldsymbol{H}_g\boldsymbol{V}^{-1/2}\boldsymbol{\mu}$$

due to the 1:1 correspondence between $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ via the link function. This leads to

$$\boldsymbol{V}^{-1/2}\widehat{\boldsymbol{Y}} = \boldsymbol{H}_g\boldsymbol{V}^{-1/2}\boldsymbol{Y}$$

which converts to

$$\widehat{\boldsymbol{Y}} = \boldsymbol{V}^{1/2}\boldsymbol{H}_g\boldsymbol{V}^{-1/2}\boldsymbol{Y} = \mathcal{H}\boldsymbol{Y}$$

as expected.

Appendix B

(Ch. 2)

B.1 Jeffreys' distance measure

The measure of the distance between two distributions at $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$ is given by

$$ds^2 = g_{ij} d\theta^i d\theta^j.$$

B.1.1 Preamble

A measure of the 'distance' between two distributions [due to Jeffreys(1961)] is reported in Barndorff–Neilsen, Cox and Reid (1986, pp86–87). The definition is given in Cox and Hinkley (1982, p130), and the explanation and derivation are in Cox and Hinkley (1978, pp51–52, problem 4.16). Here the term 'distance' is reported as not being quite accurate due to the metric in general not being Euclidean and the triangle inequality being violated. The derivation is given in terms of the notation of this thesis.

B.1.2 Derivation

The symmetric 'distance' measure is defined as

$$ds^2 = \int \ln\left(\frac{p(\boldsymbol{y};\boldsymbol{\theta}+d\boldsymbol{\theta})}{p(\boldsymbol{y};\boldsymbol{\theta})}\right)\left(p(\boldsymbol{y};\boldsymbol{\theta}+d\boldsymbol{\theta}) - p(\boldsymbol{y};\boldsymbol{\theta})\right) dy.$$

Now

$$p(\boldsymbol{y};\boldsymbol{\theta}+d\boldsymbol{\theta}) = p(\boldsymbol{y};\boldsymbol{\theta}) + \frac{\partial p(\boldsymbol{y};\boldsymbol{\theta})}{\partial \theta^i} d\theta^i + \dots$$

by a Taylor's expansion. So

$$\begin{aligned} ds^2 &= \int \ln\left(1 + \frac{1}{p}\frac{\partial p}{\partial \theta^i} d\theta^i + \dots\right)\left(\frac{\partial p}{\partial \theta^j} d\theta^j + \dots\right) dy \\ &= \int \frac{1}{p}\frac{\partial p}{\partial \theta^i} d\theta^i \frac{\partial p}{\partial \theta^j} d\theta^j dy + \dots \\ &= \int \frac{\partial \ln p}{\partial \theta^i}\frac{\partial \ln p}{\partial \theta^j} p dy d\theta^i d\theta^j \end{aligned}$$

Since

$$\ln p = \ell$$

then

$$ds^2 = d\theta^i d\theta^j E(\partial_i \ell \partial_j \ell) = g_{ij} d\theta^i d\theta^j.$$

B.2 Metric Tensor : alternative form

The alternative form for the information matrix (metric tensor) is

$$g_{ij} = -E_\theta(\partial_i \partial_j \ell).$$

B.2.1 Derivation

Now

$$p\partial_i \partial_j \ell = p\partial_i \left(\frac{\partial_j p}{p}\right)$$

since

$$\ell = \ln p.$$

Hence

$$\begin{aligned} p\partial_i \partial_j \ell &= p\left(\partial_i \partial_j p\right)\frac{1}{p} + p\left(-\frac{1}{p^2}\right)\partial_i p \partial_j p \\ &= \partial_i \partial_j p - p\frac{\partial_i p}{p}\frac{\partial_j p}{p} = \partial_i \partial_j p - p\partial_i \ell \partial_j \ell. \end{aligned}$$

Thus

$$\int p\partial_i\partial_j\ell = \int \partial_i\partial_j p - \int p\partial_i\partial_j\ell$$

giving

$$E(\partial_i\partial_j\ell) = \int \partial_i\partial_j p - g_{ij}.$$

Now

$$\int \partial_i\partial_j p = \int \partial_i(p\partial_j \ln p) = \partial_i \int p\partial_j\ell$$

but

$$\int p\partial_j\ell = 0$$

being the score statistic, so

$$\int \partial_i\partial_j p = 0$$

giving

$$E(\partial_i\partial_j\ell) = -g_{ij}.$$

B.3 Metric Tensor : results

The results below using the metric tensor are used throughout the thesis.

B.3.1 Metric tensor

$$g_{ij}g^{ik} = \delta^k_j$$

$$g^{ij} = g^{ik}g^{jm}g_{km}$$

B.3.2 Affine connection

$$\Gamma_{jim} = \Gamma^k_{ji}g_{km}$$

$$\Gamma^k_{ji} = \Gamma_{jim}g^{mk}$$

B.3.3 General tensors

Note the errors in Amari (1982a, p364 and p367). The correct tensor forms are

$$S^i_{jk} = S_{mjk} g^{ml}$$

and

$$S^{ij}_k = S^n_{lm} g^{li} g^{mi} g_{nk}$$

respectively. See Kay (1988, p55, sec. 5.4) for other examples.

B.3.4 Imbedding

The imbedding of regression coefficients $\boldsymbol{\beta}$ in the natural coordinates $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\beta}).$$

The metric tensor g_{ab} of the imbedded regression coefficients $\boldsymbol{\beta}$ in terms of the metric tensor g_{ij} of the natural parameters $\boldsymbol{\theta}$ is given by

$$g_{ab} = B^i_a B^j_b g_{ij}$$

where

$$B^i_a = \frac{\partial \theta^i}{\partial \beta^a}$$

and

$$B^j_b = \frac{\partial \theta^j}{\partial \beta^b}.$$

B.4 Riemann Christoffel Curvature Tensor

A space with an affine connection is *flat* when the Riemann Christoffel curvature tensor

$$R_{ijkl} = \left(\partial_i \ \Gamma^s_{jk} - \partial_j \ \Gamma^s_{ik}\right) g_{sl} + \ \Gamma_{irl} \ \Gamma^r_{jk} - \ \Gamma_{jrl} \ \Gamma^r_{ik}$$

vanishes identically. Then there exists an affine coordinate system such that

$$\Gamma_{ijk} = 0.$$

If the space is *not* curvature free ($R \neq 0$), no global affine coordinate system exists. At any point $\boldsymbol{\theta}_0$ however, there exists a coordinate system where the coefficients of the affine connection and its derivatives vanish, viz,

$$\Gamma_{ijk}(\boldsymbol{\theta}_0) = 0, \quad , \partial_m \Gamma_{ijk}(\boldsymbol{\theta}_0) = 0, \ \ldots$$

This creates a natural (local) coordinate system at $\boldsymbol{\theta}_0$, as described by Amari (1990, pp47–48).

Statistically, a model has an associated one parameter family of affine connections, ie., the α–connections. The corresponding Riemann Christoffel (RC) curvature tensor is $\overset{\alpha}{R}_{ijkl}$ with $\overset{\alpha}{\Gamma}$ replacing Γ in the previous definition of R, see Lauritzen (1987). The following statements are given as definitions of terms used in statistical applications.

- A statistical manifold S is α–flat when it is flat under the α–connection.
- When the manifold S is α–flat, there exists coordinates θ^i such that

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\theta}) = 0$$

  identically. The parameters $\boldsymbol{\theta}$ then form the α–affine coordinate system.

For a curved exponential family, the RC curvature tensor becomes

$$\overset{\alpha}{R}_{ijkl} = \frac{1-\alpha^2}{2} T_{km[i} T_{j]ln} g^{mn}$$

where the operation $[ij]$ is defined by[1]

$$T_{km[i} T_{j]ln} = \frac{T_{kmi} T_{jln} - T_{kmj} T_{iln}}{2}.$$

Thus the RC curvature tensor is a function of the skewness tensor, and since the α–connection is a function of the skewness tensor, then the α–flatness of the space is purely related to the solution of

$$\overset{\alpha}{\Gamma} = 0$$

as per Kass (1984, p87, 4).

[1]Note the typographical error in Amari (1982a, p365).

B.5 Exponential Families and 1–connections

The condition $\overset{1}{\Gamma}= 0$ does not necessarily imply that the parent family distribution is of exponential type.

one dimensional case

The condition $\overset{1}{\Gamma}_{\psi} = 0$ implies that $(L_1 L_2) = 0$, but since

$$L = \ell = \ln p(\boldsymbol{y};\theta) = \ln f(\boldsymbol{y};\theta)$$

this becomes

$$\int \frac{d^2\ell}{d\theta^2}\frac{d\ell}{d\theta} f\, dy = 0$$

which reduces to

$$\int \left(-\frac{1}{f}\left(\frac{df}{d\theta}\right)^2 + \frac{d^2 f}{d\theta^2}\right)\frac{d\ln f}{d\theta} dy = 0. \tag{B.1}$$

The condition

$$-\frac{1}{f}\left(\frac{df}{d\theta}\right)^2 + \frac{d^2 f}{d\theta^2} = 0$$

implies a canonical exponential family, but this condition is too stringent. The original implied condition, Equation (B.1), can be construed as

$$E\frac{d}{df}\left(\frac{d\ln f}{d\theta}\right)^2 = 0$$

which is satisfied by a general canonical exponential family, but there may be other solutions.

multidimensional case

The condition $\overset{1}{\Gamma}_{\mathsf{abc}} = 0$ implies that $(L_{\mathsf{ab}} L_{\mathsf{c}}) = 0$, which becomes

$$\int \frac{\partial\ell}{\partial\theta_{\mathsf{c}}}\frac{\partial^2\ell}{\partial\theta_{\mathsf{a}}\partial\theta_{\mathsf{b}}} f\, dy = 0$$

to give

$$\int \left(-\frac{1}{f}\frac{\partial f}{\partial\theta_{\mathsf{a}}}\frac{\partial f}{\partial\theta_{\mathsf{b}}} + \frac{\partial^2 f}{\partial\theta_{\mathsf{a}}\partial\theta_{\mathsf{b}}}\right)\frac{\partial\ln f}{\partial\theta_{\mathsf{c}}} dy = 0.$$

Imposing the constraint

$$-\frac{1}{f}\frac{\partial f}{\partial \theta_a}\frac{\partial f}{\partial \theta_b}+\frac{\partial^2 f}{\partial \theta_a \partial \theta_b}=0$$

as in the 1 D case implies a canonical exponential family, but in general there may be other solutions to the integration condition.

B.6 Wedderburn's Exponential Form

A description of the results reported in Hougaard (1982) for Wedderburn's exponential form is given. The notation used is that of Hougaard (1982). Wedderburn's one dimensional exponential family for the iid random variables $X_1, \ldots, X_n$ is

$$f(\boldsymbol{x};\theta) = e^{\theta t(\boldsymbol{x})}/\phi(\theta).$$

The parameterizations given by

$$\psi(\theta_1) = \int_{\theta_0}^{\theta_1} \left\{ \frac{d^2}{d\theta^2} \ln \phi(\theta) \right\}^{\delta} d\theta$$

are characterized by the key values of δ as given in Table B.1.

Value of δ	Induced Property
0	canonical parameter
1/3	'normal likelihood'
1/2	stable variance
2/3	zero asymptotic skewness
1	mean value parameter

Table B.1: Key values of δ.

1. The case $\delta = 0$ reproduces the canonical parameter θ via

$$\psi(\theta_1) = \theta_1 - \theta_0$$

which gives $\psi(\theta) = \theta$, ie., the natural (canonical) parameter θ.

So the transformation ψ is the identity.

2. The case $\delta = 1$ gives

$$\psi(\theta_1) = \left[\frac{d}{d\theta} \ln \phi(\theta)\right]_{\theta_0}^{\theta_1} = \ln' \phi(\theta_1) - \text{constant}$$

But

$$\ell = \theta t(\boldsymbol{x}) - \ln \phi(\theta)$$

and

$$\frac{d\ell}{d\theta} = t(\boldsymbol{x}) - \ln' \phi(\theta)$$

with

$$E\frac{d\ell}{d\theta} = 0$$

giving

$$Et(\boldsymbol{x}) = \ln' \phi(\theta) \stackrel{\text{def}}{=} \tau(\theta)$$

ie., the mean value parameter, since $\tau(\theta) = \psi(\theta)$.

3. The value $\delta = 1/2$ yields

$$\psi(\theta_1) = \int_{\theta_0}^{\theta_1} \left\{\frac{d^2}{d\theta^2} \ln \phi(\theta)\right\}^{1/2} d\theta.$$

From the likelihood

$$\frac{d^2\ell}{d\theta^2} = -\ln'' \phi(\theta)$$

and transformation $\psi(\theta)$ produces

$$\frac{d\ell}{d\psi} = \frac{d\ell}{d\theta}\frac{d\theta}{d\psi} = [t(\boldsymbol{x}) - \ln' \phi(\theta)]\frac{d\theta}{d\psi}$$

giving

$$\frac{d^2\ell}{d\psi^2} = -\ln'' \phi(\theta)\left(\frac{d\theta}{d\psi}\right)^2 + [t(\boldsymbol{x}) - \ln' \phi(\theta)]\frac{d}{d\theta}\left(\frac{d\theta}{d\psi}\right).$$

Since $V(\widehat{\psi}) = -Ed^2\ell/d\psi^2$, then a constant variance parameterization will induce

$$\text{constant} = \left(\frac{d\theta}{d\psi}\right)^2 \ln''\phi(\theta) - E(\frac{d\ell}{d\theta})\frac{d}{d\theta}\left(\frac{d\theta}{d\psi}\right).$$

For the score statistic

$$E\frac{d\ell}{d\theta} = 0$$

producing

$$\left(\frac{d\psi}{d\theta}\right)^2 \propto \ln''\phi(\theta).$$

In standardized form

$$\frac{d\psi}{d\theta} = \{\ln''\phi(\theta)\}^{1/2}$$

and finally

$$\psi(\theta) = \int \{\ln''\phi(\theta)\}^{1/2}\, d\theta.$$

4. For $\delta = 1/3$,

$$\frac{d^3\ell}{d\psi^3} = \frac{d}{d\psi}\left(\frac{d^2\ell}{d\psi^2}\right) = \frac{d}{d\theta}\left(\frac{d^2\ell}{d\theta^2}\left(\frac{d\theta}{d\psi}\right)^2 + \frac{d\ell}{d\theta}\frac{d^2\theta}{d\psi^2}\right)\frac{d\theta}{d\psi}$$

$$= \frac{d^3\ell}{d\theta^3}\left(\frac{d\theta}{d\psi}\right)^3 + 2\frac{d^2\ell}{d\theta^2}\left(\frac{d\theta}{d\psi}\right)^2\frac{d}{d\theta}(\frac{d\theta}{d\psi}) + \frac{d^2\ell}{d\theta^2}\frac{d^2\theta}{d\psi^2}\frac{d\theta}{d\psi} + \frac{d\ell}{d\theta}\frac{d}{d\theta}(\frac{d^2\theta}{d\psi^2})\frac{d\theta}{d\psi}$$

Setting the expected third derivative of the log likelihood to zero gives

$$0 = E\frac{d^3\ell}{d\psi^3} = -\ln'''\phi(\theta)\left(\frac{d\theta}{d\psi}\right)^3 - 2\ln''\phi(\theta)\left(\frac{d\theta}{d\psi}\right)^2\frac{d}{d\theta}(\frac{d\theta}{d\psi}) - \ln''\phi(\theta)\left[\frac{d}{d\psi}(\frac{d\theta}{d\psi})\right]\frac{d\theta}{d\psi} - 0.$$

So

$$0 = -\ln'''\phi(\theta)\left(\frac{d\theta}{d\psi}\right)^3 - 3\ln''\phi(\theta)\left(\frac{d\theta}{d\psi}\right)^2\frac{d}{d\theta}(\frac{d\theta}{d\psi})$$

$$= \frac{d}{d\theta}\left(\ln''(\frac{d\theta}{d\psi})^3\right).$$

This becomes

$$\ln''\phi(\theta)\left(\frac{d\theta}{d\psi}\right)^3 = \text{constant}$$

ie.,

$$\left(\frac{d\psi}{d\theta}\right)^3 \propto \ln''\phi(\theta)$$

which leads to

$$\frac{d\psi}{d\theta} = \{\ln'' \phi(\theta)\}^{1/3}.$$

This gives

$$\psi(\theta) = \int \{\ln'' \phi(\theta)\}^{1/3} \, d\theta.$$

Aitkin, Anderson, Francis and Hinde (1989, p327), give a parallel development for a GLM form of an exponential family with unit scale parameter. The transformation for the Binomial distribution is a function of the incomplete beta function, whereas the transformation for Poisson is the cube root.

5. Zero asymptotic skewness is produced for $\delta = 2/3$ via Wedderburn's equation

$$\psi(\theta_1) = \int_{\theta_0}^{\theta_1} \left\{ \frac{d^2}{d\theta^2} \ln \phi(\theta) \right\}^{2/3} d\theta.$$

This result can be verified by using the results of Hougaard (1982, p248) by setting

$$\psi = \psi(\beta) = \psi(\theta)$$

via

$$\theta(\beta) = \beta = \theta.$$

Note that $g = \psi$ and $\chi(\theta) = \ln \phi(\theta)$, using Hougaard's notation. The skewness for $\widehat{\psi}$ is now

$$E\left(\widehat{\psi} - E\widehat{\psi}\right)^3 = \ldots \left[g'(\beta) J^{-3} \left\{ -2 \frac{d^3\chi}{d\theta^3} \left(\frac{d\theta}{d\beta} \right) - 3 \left(\frac{d^2\theta}{d\beta^2} \right)' \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta} \right\} + 3g''(\beta) J^{-2} \right] + \ldots$$

Now $g = \psi$ and $\chi(\theta) = \ln \phi(\theta)$, with

$$J = -E \frac{d^2\ell}{d\theta^2} = \ln'' \phi = \frac{d^2\chi}{d\theta^2}.$$

So, zeroing the skewness gives

$$E\left(\widehat{\psi} - E\widehat{\psi}\right)^3 = 0$$

which produces

$$\psi'(\theta)J^{-3}\left\{-2\ln'''\phi - 3(0)\right\} + 3\psi'''(\theta)J^{-2} + \ldots = 0.$$

This reduces to

$$-2\psi'(\theta)\ln'''\phi + 3\psi''(\theta)\ln''\phi = 0$$

which can be written as

$$-2\left(\frac{d\psi}{d\theta}\right)^3 \ln'''\phi\,(\ln''\phi)^{-3} + 3\frac{d^2\psi}{d\theta^2}(\ln''\phi)^{-2}\left(\frac{d\psi}{d\theta}\right)^2 = 0.$$

This becomes

$$\frac{d}{d\theta}\left((\ln''\phi)^{-2}\left(\frac{d\psi}{d\theta}\right)^3\right) = 0$$

giving

$$(\ln''\phi)^{-2}\left(\frac{d\psi}{d\theta}\right)^3 = \text{constant}.$$

This produces

$$\left(\frac{d\psi}{d\theta}\right)^3 = (\ln''\phi)^2$$

to give

$$\frac{d\psi}{d\theta} = (\ln''\phi)^{2/3}$$

which converts to

$$\psi(\theta) = \int\left(\frac{d^2}{d\theta^2}\ln\phi(\theta)\right)^{2/3} d\theta.$$

This is Wedderburn's equation with $\delta = 2/3$.

Alternatively, the result can be determined directly by expanding

$$E\left(\widehat{\psi} - E\widehat{\psi}\right)^3$$

in terms of θ by use of Taylor's expansions, on $E\widehat{\psi}$.

B.7 GLM Notation

Notation is presented which defines the relation between the natural parameters and the imbedded regression parameters for a generalized linear model.

Using the notation of McCullagh and Nelder (1989) for a GLM

$$g(\mu_i) = \eta_i$$

with

$$\mu_i = h(\eta_i)$$

Now

$$E(Y_i) = b'(\theta_i) = \mu_i = h(\eta_i)$$

This relation can be defined as

$$d(\theta_i) \stackrel{\text{def}}{=} h(\eta_i)$$

So

$$\theta_i = d^{-1}[h(\eta_i)] \stackrel{\text{def}}{=} \mathsf{f}(\eta_i) = \mathsf{f}(\boldsymbol{X}_i{}^{\top}\boldsymbol{\beta}) = \mathsf{f}(\sum_j X_{ij}\beta_j) = \mathsf{f}(X_{ij}\beta^j)$$

If the canonical link is defined as c then $d \equiv c^{-1}$ so if the link is chosen as canonical then

$$\theta_i = d^{-1}[h(\eta_i)] = c[c^{-1}(\eta_i)] = \eta_i$$

as expected. In general, however

$$\theta_i = d^{-1}[h(\eta_i)] = c[h(\eta_i)] \neq \eta_i.$$

For example, d is

the identity function for the Normal distribution, and

the exponential function for the Poisson distribution.

B.8 Derivation of the Imbedding Theorem

The α–connection for the regression coefficients $\overset{\alpha}{\Gamma}_{abc}$ in terms of the α–connection for the natural parameters $\overset{\alpha}{\Gamma}_{ijk}$ is given by

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) \;=\; (\partial_a B_b^i)B_c^j g_{ij} + B_a^i B_b^j B_c^k \,\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\vartheta}(\boldsymbol{u}))$$

B.8.1 Proof

$$\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\vartheta}) = E(\partial_i\partial_j\ell\partial_k\ell) + \frac{1-\alpha}{2}E(\partial_i\ell\partial_j\ell\partial_k\ell)$$

and

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = E(\partial_a\partial_b\ell\partial_c\ell) + \frac{1-\alpha}{2}E(\partial_a\ell\partial_b\ell\partial_c\ell)$$

where

$$\partial_a\ell = B_a^i\partial_i\ell, \quad B_a^i = \frac{\partial\vartheta^i}{\partial u^a} \text{ and } \boldsymbol{\vartheta} = \boldsymbol{\vartheta}(\boldsymbol{u}).$$

Now

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = E\left[\partial_a(B_b^j\partial_j\ell)B_c^k\partial_k\ell\right] + \frac{1-\alpha}{2}E\left[B_a^i\partial_i\ell B_b^j\partial_j\ell B_c^k\partial_k\ell\right]$$

$$= E\left[B_b^jB_a^i\partial_i\partial_j\ell B_c^k\partial_k\ell + \partial_a(B_b^j)B_c^k\partial_j\ell\partial_k\ell\right] + B_a^iB_b^jB_c^k\frac{1-\alpha}{2}E(\partial_i\ell\partial_j\ell\partial_k\ell)$$

$$= \partial_a(B_b^j)B_c^kE(\partial_j\ell\partial_k\ell) + B_a^iB_b^jB_c^k\left[E(\partial_i\partial_j\ell\partial_k\ell) + \frac{1-\alpha}{2}E(\partial_i\ell\partial_j\ell\partial_k\ell)\right]$$

$$\overset{\alpha}{\Gamma}_{abc}(\boldsymbol{u}) = \partial_a(B_b^i)B_c^jE(\partial_i\ell\partial_j\ell) + B_a^iB_b^jB_c^k\,\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\vartheta}(\boldsymbol{u}))$$

$$= \partial_a(B_b^i)B_c^jg_{ij} + B_a^iB_b^jB_c^k\,\overset{\alpha}{\Gamma}_{ijk}(\boldsymbol{\vartheta}(\boldsymbol{u}))$$

as required.

B.9 Equivalence

The equivalence of Equation (A.2) of Kass (1984, p92) to (4.6) of Amari (1982a, p370) is now demonstrated.

Equation (A.2) of Kass (1984, p92) is

$$\overset{\alpha}{\Gamma}_{\theta} = \overset{\alpha}{\Gamma}_{\gamma}\left(\frac{d\gamma}{d\theta}\right)^3 + \left(\frac{d\gamma}{d\theta}\right)^{-1}i(\theta)\frac{d^2\gamma}{d\theta^2}.$$

Examining the transformation $\gamma \Rightarrow \theta$: and defining $B = d\gamma/d\theta$ yields

$$\overset{\alpha}{\Gamma}_{\theta} = \overset{\alpha}{\Gamma}_{\gamma}B^3 + (1/B)i(\theta)\partial_\theta B.$$

Since

$$i(\gamma) = g_\gamma = i(\theta)\left(\frac{d\theta}{d\gamma}\right)^2 = g_\theta B^{-2}$$

then

$$\overset{\alpha}{\Gamma}_\theta \;=\; (\partial_\theta B) B g_\gamma + B^3 \overset{\alpha}{\Gamma}_\gamma$$

in line with Amari (1982a, 4.6, p370).

Alternatively, the back transformation $\theta \Rightarrow \gamma$: yields

$$\left(\frac{d\gamma}{d\theta}\right)^3 \overset{\alpha}{\Gamma}_\gamma \;=\; \overset{\alpha}{\Gamma}_\theta - \left(\frac{d\theta}{d\gamma}\right) i(\theta) \frac{d^2\gamma}{d\theta^2}$$

and

$$\overset{\alpha}{\Gamma}_\gamma \;=\; \overset{\alpha}{\Gamma}_\theta \left(\frac{d\theta}{d\gamma}\right)^3 - i(\theta)\left(\frac{d\theta}{d\gamma}\right)\frac{d^2\gamma}{d\theta^2} \Big/ \left(\frac{d\gamma}{d\theta}\right)^3$$

Defining $B^* = d\theta/d\gamma$ gives

$$\overset{\alpha}{\Gamma}_\gamma = \overset{\alpha}{\Gamma}_\theta \, (B^*)^3 - g_\theta B^* \frac{d^2\gamma}{d\theta^2} \Big/ \left(\frac{d\gamma}{d\theta}\right)^3$$

But

$$\frac{d^2\gamma}{d\theta^2} = \frac{d}{d\theta}\left(\frac{d\gamma}{d\theta}\right) = \frac{d}{d\gamma}\left(\frac{d\gamma}{d\theta}\right)\frac{d\gamma}{d\theta} \;=\; \frac{d}{d\gamma}\left(\frac{d\theta}{d\gamma}\right)^{-1}\frac{d\gamma}{d\theta} = -\left(\frac{d\theta}{d\gamma}\right)^{-2}\frac{d^2\theta}{d\gamma^2}\frac{d\gamma}{d\theta}$$

This gives

$$\frac{d^2\gamma}{d\theta^2} = -\frac{d^2\theta}{d\gamma^2} \Big/ \left(\frac{d\theta}{d\gamma}\right)^3 \Longrightarrow -\frac{d^2\gamma}{d\theta^2} \Big/ \left(\frac{d\gamma}{d\theta}\right)^3 = \frac{d^2\theta}{d\gamma^2}$$

$$\therefore \;\; \overset{\alpha}{\Gamma}_\gamma \;=\; \overset{\alpha}{\Gamma}_\theta \, (B^*)^3 + g_\theta B^* \left(\frac{d^2\theta}{d\gamma^2}\right) = (\partial_\gamma B^*)\, B^* g_\theta + (B^*)^3 \overset{\alpha}{\Gamma}_\theta$$

in line with (4.6) of Amari (1982a, p370), ie., the 'imbedding theorem'.

Appendix C

(Ch. 3)

C.1 The derivation of α–curvature

The general definition of a covariant derivative with respect to a vector field $\boldsymbol{X}(\boldsymbol{\theta})$ is[1]

$$\nabla_j X^i = \frac{\partial X^i}{\partial \theta^j} + \Gamma^i_{jk} X^k(\boldsymbol{\theta}).$$

So

$$\nabla_j B^i_b(\boldsymbol{u}) = \frac{\partial B^i_b}{\partial \theta^j} + \Gamma^i_{jk} B^k_b$$

but $\partial/\partial u^a$ is required. Therefore, using the chain rule

$$B^j_a \overset{\alpha}{\nabla}_j B^i_b = \overset{\alpha}{H}{}^i_{ab} = B^j_a \left(\frac{\partial B^i_b}{\partial \theta^j} + B^k_b \overset{\alpha}{\Gamma}{}^i_{jk} \right) = \frac{\partial \theta^j}{\partial u^a} \frac{\partial B^i_b}{\partial \theta^j} + B^j_a B^k_b \overset{\alpha}{\Gamma}{}^i_{jk}.$$

$$\Longrightarrow \overset{\alpha}{H}{}^i_{ab} = \partial_a B^i_b + B^j_a B^k_b \overset{\alpha}{\Gamma}{}^i_{jk}.$$

Alternatively, using u^a directly,

$$\overset{\alpha}{H}{}^i_{ab} = \overset{\alpha}{\nabla}_a B^i_b = \frac{\partial B^i_b}{\partial u^a} + \overset{\alpha}{\Gamma}{}^i_{ak} B^k_b = \partial_a B^i_b + \overset{\alpha}{\Gamma}_{akn} g^{in} B^k_b$$

If an original coordinate system is taken as being indexed by l, k, n and the resultant system as being indexed by a, k, n, then the α–connection transforms according to equation (2.28) of Amari (1982a, p364) , viz,

$$\overset{\alpha}{\Gamma}_{akn} = B^l_a B^k_k B^n_n \overset{\alpha}{\Gamma}_{lkn} + B^l_k \left(\partial_a B^k_k \right) g_{lk}$$

[1] See Lovelock and Rund (1989, p76, 5.4).

$$\partial_a B^k_k = 0 \;\; \Rightarrow \overset{\alpha}{\Gamma}_{akn} = B^l_a \overset{\alpha}{\Gamma}_{lkn}.$$

This gives

$$\overset{\alpha}{H}{}^i_{ab} = \partial_a B^i_b + B^l_a \overset{\alpha}{\Gamma}_{lkn} g^{in} B^k_b = \partial_a B^i_b + B^l_a \overset{\alpha}{\Gamma}{}^i_{lk} B^k_b.$$
$$\Longrightarrow \overset{\alpha}{H}{}^i_{ab} = \partial_a B^i_b + B^j_a \overset{\alpha}{\Gamma}{}^i_{jk} B^k_b.$$

C.2 The transformation rule for α-curvature

If the coordinate system is changed from $\boldsymbol{u} = (u^a)$ to $\boldsymbol{v} = (v^{a'})$ then

$$\overset{\alpha}{H}{}^i_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{H}{}^i_{ab} + B^i_b \partial_{a'} B^b_{b'},$$

after Amari (1982a, p371, 4.4).

C.2.1 Proof

$$\overset{\alpha}{H}{}^i_{a'b'} \overset{\text{def}}{=} \partial_{a'} B^i_{b'} + \overset{\alpha}{\Gamma}{}^i_{jk} B^j_{a'} B^k_{b'}$$

$$= \frac{\partial B^i_{b'}}{\partial v^{a'}} + \overset{\alpha}{\Gamma}{}^i_{jk} \frac{\partial \theta^j}{\partial v^{a'}} \frac{\partial \theta^k}{\partial v^{b'}}$$

$$= \frac{\partial B^i_{b'}}{\partial u^a} \frac{\partial u^a}{\partial v^{a'}} + \Gamma^i_{jk} \frac{\partial \theta^j}{\partial u^a} \frac{\partial u^a}{\partial v^{a'}} \frac{\partial \theta^k}{\partial u^b} \frac{\partial u^b}{\partial v^{b'}}$$

$$= B^a_{a'} \frac{\partial}{\partial u^a} \left(\frac{\partial \theta^i}{\partial u^b} \frac{\partial u^b}{\partial v^{b'}} \right) + \Gamma^i_{jk} B^j_a B^k_b B^a_{a'} B^b_{b'}$$

$$\overset{\alpha}{H}{}^i_{a'b'} = B^a_{a'} \partial_a \left(B^i_b B^b_{b'} \right) + \Gamma^i_{jk} B^j_a B^k_b B^a_{a'} B^b_{b'}$$

$$= B^a_{a'} B^b_{b'} \partial_a B^i_b + B^a_{a'} B^b_{b'} \Gamma^i_{jk} B^j_a B^k_b + B^a_{a'} \partial_a \left(B^b_{b'} \right) B^i_b$$

$$\Rightarrow \overset{\alpha}{H}{}^i_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{H}{}^i_{ab} + B^i_b \partial_{a'} B^b_{b'}.$$

So $\overset{\alpha}{H}{}^{i}_{ab}$ is in general not a tensor, due to the presence of the second term.

Note that the transformation law for a (0,2) tensor is

$$\bar{S}_{hk} = B^j_h B^l_k S_{jl}$$

following Lovelock and Rund (1989, 2.9, p60).

C.3 Tensorial normal α–curvature

The transformation rule for α–curvature is

$$\overset{\alpha}{H}{}^{i}_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{H}{}^{i}_{ab} + B^i_b \partial_{a'} B^b_{b'} \ .$$

To prove this tensorial assertion, it is sufficient to show that under the projection N^i_j, the second term in the transformation rule, ie.,

$$B^i_b \partial_{a'} B^b_{b'}$$

vanishes.[2]

C.3.1 Proof

The normal projection of the alleged non–tensorial term is

$$N^i_j B^j_b \partial_{a'} B^b_{b'} = \left(\delta^i_j - P^i_j\right) B^j_b \partial_{a'} B^b_{b'}$$

The L.H.S. becomes

$$\left(\delta^i_j B^j_b - P^i_j B^j_b\right) \partial_{a'} B^b_{b'}$$

giving

$$\left(B^i_b - g^{\alpha\beta} B^i_\beta B^k_\alpha g_{kj} B^j_b\right) \partial_{a'} B^b_{b'}$$

The second term gives

$$B^i_\beta \left(g^{\alpha\beta} B^k_\alpha B^j_b g_{kj}\right)$$

[2]See Amari (1982a, p371).

The result from Lovelock and Rund (1989, p268, 4.10) can be stated as

$$g^{\alpha\beta}B^k_\alpha B^j_b g_{kj} = \delta^\beta_b.$$

Hence

$$P^i_j B^j_b \partial_{a'} B^b_{b'} = B^i_\beta \delta^\beta_b \partial_{a'} B^b_{b'} = B^i_b \partial_{a'} B^b_{b'}$$

This gives

$$N^i_j B^j_b \partial_{a'} B^b_{b'} = 0.$$

Hence, the normal components of α–curvature form a (0,2) tensor with respect to i, since

$$\left(\overset{\alpha}{N^i_j H^j_{a'b'}}\right) = N^i_j \left(B^a_{a'} B^b_{b'} \overset{\alpha}{H^j_{ab}}\right) = B^a_{a'} B^b_{b'} \left(\overset{\alpha}{N^i_j H^j_{ab}}\right),$$

to give the tensorial law

$$\overset{\alpha}{\mathcal{N}}{}^i_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{\mathcal{N}}{}^i_{ab}\ .$$

C.3.2 Alternative Derivation

Under the reparameterization defined by $u^a \rightarrow v^{a'}$, ie., $\boldsymbol{\beta} \rightarrow \boldsymbol{\mathcal{B}}$, the tangential components of α–curvature become

$$\overset{\alpha}{\mathcal{N}}{}^i_{a'b'} = \partial_{a'} B^i_{b'} + \overset{\alpha}{\Gamma}{}^i_{jk} B^j_{a'} B^k_{b'} - \overset{\alpha}{\Gamma}_{a'b'c} g^{cd} B^i_d$$

$$= B^a_{a'} \partial_a \left(B^i_b B^b_{b'}\right) + \overset{\alpha}{\Gamma}{}^i_{jk} B^j_a B^a_{a'} B^k_b B^b_{b'} - \left[B^a_{a'} B^b_{b'} \overset{\alpha}{\Gamma}_{abc} + B^a_c \left(\partial_{a'} B^b_{b'}\right) g_{ab}\right] g^{cd} B^i_d$$

since, under the transformation rule for an affine connection

$$\overset{\alpha}{\Gamma}_{a'b'c} = B^a_{a'} B^b_{b'} B^c_c \overset{\alpha}{\Gamma}_{abc} + B^a_c \left(\partial_{a'} B^b_{b'}\right) g_{ab}.$$

This gives $\overset{\alpha}{\mathcal{N}}{}^i_{a'b'}$ as

$$= B^a_{a'} \left(\partial_a B^i_b\right) B^b_{b'} + B^a_{a'} B^i_b \partial_a B^b_{a'} + B^a_{a'} B^b_{b'} \overset{\alpha}{\Gamma}{}^i_{jk} B^j_a B^k_b - \left[B^a_{a'} B^b_{b'} \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d + B^a_c \left(\partial_{a'} B^b_{b'}\right) g_{ab} B^i_d g^{cd}\right]$$

$$= B^a_{a'} B^b_{b'} \left(\partial_a B^i_b + \overset{\alpha}{\Gamma}{}^i_{jk} B^j_a B^k_b - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d \right) + B^a_{a'} B^i_b \partial_a B^b_{b'} - \left(\partial_a B^b_{b'} \right) B^a_{a'} \left(g^{dc} g_{ab} B^a_c \right) B^i_d \, .$$

From Lovelock and Rund (1989, p268, 4.9)

$$g^{\alpha\epsilon} g_{hj} B^h_\epsilon = B^\alpha_j$$

then

$$g^{dc} g_{ab} B^a_c = B^d_b$$

with

$$\overset{\alpha}{\mathcal{N}}{}^i_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{\mathcal{N}}{}^i_{ab} + B^a_{a'} B^i_b \partial_a B^b_{b'} - B^a_{a'} \left(\partial_a B^b_{b'} \right) B^d_b B^i_d$$

but

$$B^d_b B^i_d = B^i_b$$

giving

$$\overset{\alpha}{\mathcal{N}}{}^i_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{\mathcal{N}}{}^i_{ab} \, .$$

So the 'normal components with respect to i form a tensor ...', Amari (1982a, p371).

Note

This alternative derivation verifies the form of the projection operators.

C.4 Lemma

The vectors $\mathcal{N}^i$ normal to the imbedded tangent subspace T_u spanned by the vectors B^i_a satisfy

$$g_{hi} B^h_e \overset{\alpha}{\mathcal{N}}{}^i_{ab} = 0.$$

This is merely a restatement of the orthogonality results of Lovelock and Rund (1989, p270, 4.21), and Amari (1982a, p370, 4.7).

C.4.1 Proof

The normal component of α–curvature is

$$\overset{\alpha}{\mathcal{N}}{}^{i}_{ab} = \partial_a B^i_b + \overset{\alpha}{\Gamma}{}^{i}_{jk} B^j_a B^k_b - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d$$

So

$$\overset{\alpha}{\mathcal{N}}{}^{i}_{ab}\, g_{hi} B^h_e = \left(\partial_a B^i_b + \overset{\alpha}{\Gamma}{}^{i}_{jk} B^j_a B^k_b - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d\right) g_{hi} B^h_e$$

$$= \left(\partial_a B^i_b\right) g_{hi} B^h_e + \overset{\alpha}{\Gamma}{}^{i}_{jk} B^j_a B^k_b g_{hi} B^h_e - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d g_{hi} B^h_e$$

$$= \left(\partial_a B^i_b\right) B^h_e g_{hi} + B^j_a B^k_b B^h_e g_{hi} \overset{\alpha}{\Gamma}{}^{i}_{jk} - \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d B^h_e g_{hi}$$

$$= \left(\partial_a B^i_b\right) B^h_e g_{hi} + B^j_a B^k_b B^h_e \overset{\alpha}{\Gamma}_{jkh} - \overset{\alpha}{\Gamma}_{abc} \delta^c_e \,.$$

Thus,

$$g_{hi} B^h_e \overset{\alpha}{\mathcal{N}}{}^{i}_{ab} = \overset{\alpha}{\Gamma}_{abe} - \overset{\alpha}{\Gamma}_{abe} = 0.$$

This result reinforces the definition of $\overset{\alpha}{\mathcal{N}}{}^{i}_{ab}$ as being the Normal component of α–curvature. The fact that $\overset{\alpha}{\mathcal{N}}{}^{i}_{ab}$ as a vector with respect to i is normal to the tangent subspace T_u is also verified. This orthogonality is fundamental in establishing the decomposition of total scalar curvature.

C.5 Non–tensorial tangential α–curvature

The projection of α–curvature onto the tangent subspace gives the tangential component of α–curvature $\overset{\alpha}{\mathcal{T}}{}^{i}_{ab}$ as

$$\overset{\alpha}{\mathcal{T}}{}^{i}_{ab} \overset{\text{def}}{=} P^i_j \overset{\alpha}{H}{}^{j}_{ab} = \left(\partial_a B^j_b + \overset{\alpha}{\Gamma}{}^{j}_{ik} B^i_a B^k_b\right) P^i_j$$

to become

$$\overset{\alpha}{\mathcal{T}}{}^{i}_{ab} = \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d$$

in agreement with Amari (1990, p156, 5.26).

C.5.1 Proof

Under a reparameterization u^a to $v^{a'}$, the tangential component with respect to the new parameters $v^{a'}$ is

$$\overset{\alpha}{T}{}^{i}_{a'b'} = \overset{\alpha}{\Gamma}_{a'b'c} g^{cd} B^i_d.$$

The transformation rule for an affine connection is

$$\overset{\alpha}{\Gamma}_{a'b'c} = B^a_{a'} B^b_{b'} B^c_c \overset{\alpha}{\Gamma}_{abc} + B^a_c \left(\partial_{a'} B^b_{b'}\right) g_{ab},$$

from Appendix C.1 (alternative derivation), giving

$$\begin{aligned}\overset{\alpha}{T}{}^{i}_{a'b'} &= \left[B^a_{a'} B^b_{b'} B^c_c \overset{\alpha}{\Gamma}_{abc} + B^a_c \left(\partial_{a'} B^b_{b'}\right) g_{ab}\right] g^{cd} B^i_d \\ &= B^a_{a'} B^b_{b'} \overset{\alpha}{\Gamma}_{abc} g^{cd} B^i_d + \partial_{a'} B^b_{b'} \left(g^{cd} B^i_d B^a_c g_{ab}\right).\end{aligned}$$

Finally,

$$\overset{\alpha}{T}{}^{i}_{a'b'} = B^a_{a'} B^b_{b'} \overset{\alpha}{T}{}^{i}_{ab} + \left(\partial_{a'} B^b_{b'}\right) P^i_b,$$

using the projection operator from Section 3.3.1. Thus, in general, the tangential component is *not* a tensor, due to the presence of the second term. So the tangential component will be subject to changes under reparameterization. Hence the connotation 'parameter–effects' is justified.

Appendix D

(Ch. 5)

D.1 GLIM Output : Test Problem 1

GLM variant : one–step implementation

```
[o] GLIM 3.77 update1 (copyright)1985 Royal Statistical Society, London
[i] $units 2$
[i] $data x y
[i] $read
[i] 2 2.5
[i] 3 10.0
[i] $calc %t = 2.0537$
[i] $calc f0 = x**%t$
[i] $calc f1 = f0*%log(x)$
[i] $calc f2 = f1*%log(x)$
[i] $calc v = %sqrt(f2/2)$
[i] $calc a0 = f1/(2*v)$
[i] $calc q = y - f0 + a0*a0$
[i] $look q$
[o]         Q
[o]  1   0.4242
[o]  2   5.2265
[i] $yvar q$
[i] $link s$
[i] $offset a0$
[i] $fit v - 1$
[o] deviance = 2.9334 at cycle  3
[o]     d.f. = 1
[o]
[i] $dis erm$
```

```
[o]          estimate        s.e.     parameter
[o]     1   5.650e-06      0.1573     V
[o]     scale parameter taken as  2.933
[o]
[o]   unit  observed    fitted   residual
[o]      1    0.4242    2.0759     -1.652
[o]      2    5.2265    4.7735      0.453
[o]
[o] Current model:
[o]
[o]   number of units is  2
[o]
[o]   y-variate  Q
[o]   weight     *
[o]   offset     A0
[o]
[o]   probability distribution is NORMAL
[o]               link function is SQUARE ROOT
[o]             scale parameter is to be estimated by the mean deviance
[o]
[o]   terms =  V
```

GLM variant : iterative form

```
[o] GLIM 3.77 update1 (copyright)1985 Royal Statistical Society, London
[i] $units 2$
[i] $data x y$
[i] $read
[i] 2 2.5
[i] 3 10.0
[i] $accuracy 6$
[i] $calc %a=1$
[i] $calc %i=0$
[i] $calc %t = 1.0$
[i] $macro fit
[i] $calc f0 = x**%t$
[i] $calc f1 = f0*%log(x)$
[i] $calc f2 = f1*%log(x)$
[i] $calc v = %sqrt(f2/2)$
[i] $calc a0 = f1/(2*v)$
[i] $calc q = y - f0 + a0*a0$
[i] $yvar q$
[i] $link s$
[i] $offset a0$
```

```
[i] $fit v - 1$
[i] $extract %pe$
[i] $calc %a=%gt(%pe*%pe,0.0000001)$
[i] $calc %t=%pe + %t$
[i] $look %t$
[i] $calc %i=%i+1$
[i] $endmac
[i] $while %a fit$
[o] deviance = 3.605425 at cycle  3
[o]     d.f. = 1
[o]
[o]    2.17870
[o] deviance = 2.930771 at cycle  3
[o]     d.f. = 1
[o]
[o]    2.05302
[o] deviance = 2.933361 at cycle  3
[o]     d.f. = 1
[o]
[o]    2.05371
[o] deviance = 2.933361 at cycle  3
[o]     d.f. = 1
[o]
[o]    2.05371
[i] $look %t$
[o]    2.05371
[i] $calc fv=x**%t$
[i] $calc res=y-fv$
[i] $look x y fv res$
[o]         X         Y         FV        RES
[o]  1   2.00000    2.50000   4.15171  -1.651710
[o]  2   3.00000   10.00000   9.54699   0.453007
[i] $look %i$
[o]    4.00000
```

Linearization method

```
[o] GLIM 3.77 update1 (copyright)1985 Royal Statistical Society, London
[i] $units 2$
[i] $data x y
[i] $read
[i] 2 2.5
[i] 3 10.0
[i] $calc %t = 2.0537$
```

```
[i] $calc f0 = x**%t$
[i] $calc f1 = f0*%log(x)$
[i] $calc p = y - f0 $
[i] $yvar p$
[i] $fit f1 - 1$
[o] deviance = 2.9334
[o]     d.f. = 1
[o]
[i] $dis erm$
[o]          estimate        s.e.     parameter
[o]     1  -1.023e-05      0.1575     F1
[o]     scale parameter taken as  2.933
[o]
[o]   unit  observed    fitted   residual
[o]      1   -1.6517   -0.0000     -1.652
[o]      2    0.4531   -0.0001      0.453
[o]
[o] Current model:
[o]
[o]   number of units is  2
[o]
[o]   y-variate  P
[o]   weight     *
[o]   offset     *
[o]
[o]   probability distribution is NORMAL
[o]               link function is IDENTITY
[o]             scale parameter is to be estimated by the mean deviance
[o]
[o]   terms =  F1
[o]
[i] $calc fv=x**%t$
[i] $calc res=y-fv$
[i] $look x y fv res$
[o]        X       Y       FV      RES
[o]  1   2.000    2.500   4.152  -1.6517
[o]  2   3.000   10.000   9.547   0.4531
```

D.2 GENSTAT Output : Test Problem 2

Optimum value for the parameter θ.

```
Genstat 5  Release 3.2  (IBM-PC 80386/DOS)          02 April 1998 22:54:50
Copyright 1995, Lawes Agricultural Trust (Rothamsted Experimental Station)

       1  JOB "ds"
       2  UNITS [NVALUES=12]
       3  READ  time, y

        Identifier    Minimum       Mean    Maximum     Values    Missing
              time      0.500      5.250     16.000         12          0
                 y     0.0300     0.5108     0.9600         12          0

      17  CALC t = -0.2069
      18  CALC f0 = exp(time*t)
      19  CALC f1 = f0*(time)
      20  CALC f2 = f1*time
      21  CALC v = sqrt(f2/2)
      22  CALC a0 = f1/(2*v)
      23  CALC q = y - f0 + a0*a0
      24  MODEL  [OFFSET=a0 ; LINK=squareroot] q
      25  FIT [CONSTANT=omit; PRINT=model,summary,estimates,fittedvalues] v

25.....................................................................


*****     Regression Analysis *****

     Response variate: q
        Link function: Square root
      Offset variate: a0
        Fitted terms: v


***     Summary of analysis ***

                d.f.         s.s.         m.s.      v.r.
Regression         1            *            *
Residual          11      0.01202     0.001093
Total             12      0.00814     0.000678
```

```
Residual      variance exceeds variance of Y variate
Standard      error of observations is estimated to be 0.0331
*     MESSAGE: The following units have high leverage:
                   7        0.183
                   8        0.183


***      Estimates of regression coefficients ***

                      estimate          s.e.       t(11)
v                     -0.00001       0.00809        0.00


***      Fitted values and residuals ***

                                                  Standardized
             Unit     Response Fitted value residual Leverage
              1       0.5091       0.4509     1.77    0.012
              2       0.4591       0.4509     0.25    0.012
              3       0.4534       0.4065     1.45    0.040
              4       0.3834       0.4065    -0.71    0.040
              5       0.2994       0.3305    -0.99    0.105
              6       0.2894       0.3305    -1.31    0.105
              7       0.2615       0.2185     1.44    0.183
              8       0.2015       0.2185    -0.57    0.183
              9       0.0745       0.0955    -0.69    0.140
             10       0.1145       0.0955     0.62    0.140
             11       0.0117       0.0182    -0.20    0.020
             12       0.0317       0.0182     0.41    0.020

Mean                  0.2574       0.2534     0.12    0.083

      26  CALC p = y - f0
      27  MODEL p
      28  FIT [CONSTANT=omit; PRINT=model,summary,estimates,fittedvalues] f1

28.....................................................................


*****      Regression Analysis *****

    Response variate: p
        Fitted terms: f1
```

*** Summary of analysis ***

	d.f.	s.s.	m.s.	v.r.
Regression	1	0.00000	0.000000	0.00
Residual	11	0.01202	0.001093	
Total	12	0.01202	0.001002	

Residual variance exceeds variance of Y variate
Standard error of observations is estimated to be 0.0331
* MESSAGE: The following units have high leverage:

7	0.183
8	0.183

*** Estimates of regression coefficients ***

	estimate	s.e.	t(11)
f1	-0.00001	0.00809	0.00

*** Fitted values and residuals ***

Unit	Response	Fitted value	Standardized residual	Leverage
1	0.0583	0.0000	1.77	0.012
2	0.0083	0.0000	0.25	0.012
3	0.0469	0.0000	1.45	0.040
4	-0.0231	0.0000	-0.71	0.040
5	-0.0311	0.0000	-0.99	0.105
6	-0.0411	0.0000	-1.31	0.105
7	0.0429	0.0000	1.44	0.183
8	-0.0171	0.0000	-0.57	0.183
9	-0.0211	0.0000	-0.69	0.140
10	0.0189	0.0000	0.62	0.140
11	-0.0065	0.0000	-0.20	0.020
12	0.0135	0.0000	0.41	0.020
Mean	0.0041	0.0000	0.12	0.083

Two standard deviations above the optimum value.

Genstat 5 Release 3.2 (IBM-PC 80386/DOS) 02 April 1998 22:56:20

```
 1  JOB "ds"
 2  UNITS [NVALUES=12]
 3  READ  time, y
```

Identifier	Minimum	Mean	Maximum	Values	Missing
time	0.500	5.250	16.000	12	0
y	0.0300	0.5108	0.9600	12	0

```
17  CALC t = -0.22308
18  CALC f0 = exp(time*t)
19  CALC f1 = f0*(time)
20  CALC f2 = f1*time
21  CALC v = sqrt(f2/2)
22  CALC a0 = f1/(2*v)
23  CALC q = y - f0 + a0*a0
24  MODEL  [OFFSET=a0 ; LINK=squareroot] q
25  FIT [CONSTANT=omit; PRINT=model,summary,estimates,fittedvalues] v

25.....................................................................

*****     Regression Analysis *****

     Response variate: q
        Link function: Square root
      Offset variate: a0
        Fitted terms: v


***     Summary of analysis ***

                 d.f.         s.s.         m.s.      v.r.
Regression          1            *            *
Residual           11      0.01202     0.001093
Total              12      0.00822     0.000685

Residual     variance exceeds variance of Y variate
Standard     error of observations is estimated to be 0.0331
*     MESSAGE: The following units have high leverage:
                      7       0.184
                      8       0.184
```

```
***     Estimates of regression coefficients ***

                    estimate         s.e.     t(11)
v                    0.01618      0.00812      1.99


***     Fitted values and residuals ***

                                                Standardized
          Unit     Response Fitted value residual Leverage
             1       0.5128       0.4545     1.77    0.012
             2       0.4628       0.4545     0.25    0.012
             3       0.4600       0.4131     1.45    0.040
             4       0.3900       0.4131    -0.71    0.040
             5       0.3100       0.3411    -1.00    0.105
             6       0.3000       0.3411    -1.32    0.105
             7       0.2751       0.2322     1.44    0.184
             8       0.2151       0.2322    -0.57    0.184
             9       0.0861       0.1071    -0.68    0.139
            10       0.1261       0.1071     0.62    0.139
            11       0.0159       0.0223    -0.20    0.019
            12       0.0359       0.0223     0.42    0.019

Mean                 0.2658       0.2617     0.12    0.083

  26  CALC p = y - f0
  27  MODEL p
  28  FIT [CONSTANT=omit; PRINT=model,summary,estimates,fittedvalues] f1

28.....................................................................


***** Regression Analysis *****

 Response variate: p
     Fitted terms: f1


*** Summary of analysis ***

              d.f.         s.s.         m.s.      v.r.
Regression       1      0.00401     0.004009      3.66
Residual        11      0.01204     0.001094
```

```
Total            12       0.01605      0.001337

Residual variance exceeds variance of Y variate
Standard error of observations is estimated to be 0.0331
* MESSAGE: The following units have high leverage:
                    7       0.187
              8       0.187


*** Estimates of regression coefficients ***

                 estimate         s.e.      t(11)
f1                0.01672      0.00873       1.91


*** Fitted values and residuals ***
```

Unit	Response	Fitted value	Standardized residual	Leverage
1	0.0655	0.0075	1.77	0.014
2	0.0155	0.0075	0.25	0.014
3	0.0599	0.0134	1.44	0.045
4	-0.0101	0.0134	-0.72	0.045
5	-0.0101	0.0214	-1.01	0.114
6	-0.0201	0.0214	-1.33	0.114
7	0.0703	0.0274	1.44	0.187
8	0.0103	0.0274	-0.57	0.187
9	0.0021	0.0225	-0.66	0.126
10	0.0421	0.0225	0.64	0.126
11	0.0018	0.0075	-0.17	0.014
12	0.0218	0.0075	0.43	0.014
Mean	0.0208	0.0166	0.12	0.083

Two standard deviations below the optimum value.

```
Genstat 5  Release 3.2  (IBM-PC 80386/DOS)          30 March 1998 21:32:55
Copyright 1995, Lawes Agricultural Trust (Rothamsted Experimental Station)

  1  JOB "ds"
  2  UNITS [NVALUES=12]
  3  READ  time, y
```

```
    Identifier   Minimum      Mean    Maximum     Values   Missing
          time     0.500     5.250     16.000         12         0
             y    0.0300    0.5108     0.9600         12         0

  17  CALC t = -0.19072
  18  CALC f0 = exp(time*t)
  19  CALC f1 = f0*(time)
  20  CALC f2 = f1*time
  21  CALC v = sqrt(f2/2)
  22  CALC a0 = f1/(2*v)
  23  CALC q = y - f0 + a0*a0
  24  MODEL  [OFFSET=a0 ; LINK=squareroot] q
  25  FIT [CONSTANT=omit; PRINT=model,summary,estimates,fittedvalues] v

 25.....................................................................


***** Regression Analysis *****

 Response variate: q
    Link function: Square root
   Offset variate: a0
     Fitted terms: v


*** Summary of analysis ***

              d.f.         s.s.         m.s.      v.r.
Regression       1            *            *
Residual        11      0.01202     0.001092
Total           12      0.01069     0.000890

Percentage variance accounted for 20.8
Standard error of observations is estimated to be 0.0331
* MESSAGE: The following units have high leverage:
                  7       0.184
                  8       0.184


*** Estimates of regression coefficients ***

                 estimate         s.e.     t(11)
v                -0.01623      0.00813     -2.00
```

```
*** Fitted values and residuals ***

                                          Standardized
      Unit     Response Fitted value residual Leverage
         1       0.5055       0.4472     1.77    0.012
         2       0.4555       0.4472     0.25    0.012
         3       0.4468       0.3999     1.45    0.040
         4       0.3768       0.3999    -0.71    0.040
         5       0.2886       0.3196    -0.99    0.106
         6       0.2786       0.3196    -1.31    0.106
         7       0.2468       0.2039     1.44    0.184
         8       0.1868       0.2039    -0.57    0.184
         9       0.0613       0.0823    -0.69    0.139
        10       0.1013       0.0823     0.62    0.139
        11       0.0064       0.0130    -0.20    0.019
        12       0.0264       0.0130     0.41    0.019

Mean             0.2484       0.2443     0.12    0.083

  26  CALC p = y - f0
  27  MODEL p
  28  FIT [CONSTANT=omit; PRINT=model,summary,estimates,fittedvalues] f1

28.....................................................................


***** Regression Analysis *****

 Response variate: p
     Fitted terms: f1


*** Summary of analysis ***

              d.f.         s.s.         m.s.      v.r.
Regression       1      0.00471     0.004705      4.30
Residual        11      0.01204     0.001095
Total           12      0.01675     0.001396

Percentage variance accounted for 16.2
Standard error of observations is estimated to be 0.0331
* MESSAGE: The following units have high leverage:
                  7       0.177
```

8 0.177

*** Estimates of regression coefficients ***

	estimate	s.e.	t(11)
f1	-0.01547	0.00746	-2.07

*** Fitted values and residuals ***

Unit	Response	Fitted value	Standardized residual	Leverage
1	0.0510	-0.0070	1.76	0.011
2	0.0010	-0.0070	0.24	0.011
3	0.0336	-0.0128	1.43	0.035
4	-0.0364	-0.0128	-0.73	0.035
5	-0.0529	-0.0211	-1.01	0.095
6	-0.0629	-0.0211	-1.33	0.095
7	0.0137	-0.0289	1.42	0.177
8	-0.0463	-0.0289	-0.58	0.177
9	-0.0475	-0.0269	-0.68	0.154
10	-0.0075	-0.0269	0.64	0.154
11	-0.0173	-0.0117	-0.17	0.029
12	0.0027	-0.0117	0.44	0.029
Mean	-0.0141	-0.0181	0.12	0.083

D.3 Replication results

This section reports some basic results for differential geometric quantities such as the metric tensor and α–connection for replicated data, as used in Section 5.1.2. The sources for each results are shown at the start of each subsection. These sources do not always present a detailed argument, so the derivations given show the required working.

D.3.1 Introduction

Let $\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}$ be N independent observations from the same distribution, ie, replicates. The quantity $\underset{k}{\boldsymbol{y}}$ stands for the kth replicate of the response $\boldsymbol{y}$ where $\boldsymbol{y}^\top = (y_1, y_2, \ldots, y_n)$. Thus, if there are N replicates, the subscript n stands for the dimensionality of the response variable $\boldsymbol{y}$, while N in simply the number of replications. Then the joint pdf of $\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}$ is

$$\bar{p}(\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}; \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\underset{i}{\boldsymbol{y}})$$

to give the log–likelihood

$$\bar{\ell} \stackrel{\text{def}}{=} \ln \bar{p} = \sum_{i=1}^{N} \ell(\underset{i}{\boldsymbol{y}})$$

and so

$$\partial_j \bar{\ell} = \partial_j \bar{\ell}(\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}) = \sum_{i=1}^{N} \partial_j \ell(\underset{i}{\boldsymbol{y}}; \boldsymbol{\theta}).$$

D.3.2 Metric tensor

(Amari, 1990, p115)

The metric tensor based on the N replicate observations is

$$\bar{g}_{ij} \stackrel{\text{def}}{=} E(\partial_i \bar{\ell} \partial_j \bar{\ell}) = E\left[\sum_k \partial_i \ell(\underset{k}{\boldsymbol{y}}) \sum_l \partial_j \ell(\underset{l}{\boldsymbol{y}})\right]$$

ie,

$$= E\left[\sum_{k=1}^{N} \partial_i \ell(\underset{k}{\boldsymbol{y}}) \{\partial_j \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_j \ell(\underset{N}{\boldsymbol{y}})\}\right] = E\left[\partial_i \ell(\underset{1}{\boldsymbol{y}}) \partial_j \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_i \ell(\underset{N}{\boldsymbol{y}}) \partial_j \ell(\underset{N}{\boldsymbol{y}})\right]$$

since $\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}$ are independent. The metric tensor becomes

$$\bar{g}_{ij} = \sum_{k=1}^{N} \partial_i \ell(\underset{k}{\boldsymbol{y}}) \partial_j \ell(\underset{k}{\boldsymbol{y}}) = N E \partial_i \ell \partial_j \ell = N g_{ij}$$

since $\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}$ are identically distributed.

D.3.3 α–connection

(Amari, 1990, p116)

The α–connection using the replicate data is

$$\bar{\overset{\alpha}{\Gamma}}_{ijk} \overset{\text{def}}{=} E\left[\partial_i \partial_j \bar{\ell} \partial_k \bar{\ell} + \frac{1-\alpha}{2} \partial_i \bar{\ell} \partial_j \bar{\ell} \partial_k \bar{\ell}\right]$$

$$= E\left[\partial_i \left(\sum_l \partial_j \ell(\underset{l}{\boldsymbol{y}})\right)\left(\sum_m \partial_k \ell(\underset{m}{\boldsymbol{y}})\right) + \frac{1-\alpha}{2}\left(\sum_p \partial_i \ell(\underset{p}{\boldsymbol{y}})\right)\left(\sum_q \partial_j \ell(\underset{q}{\boldsymbol{y}})\right)\left(\sum_r \partial_k \ell(\underset{r}{\boldsymbol{y}})\right)\right]$$

$$= E\left[(\partial_i \partial_j \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_i \partial_j \ell(\underset{N}{\boldsymbol{y}})(\partial_k \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_k \ell(\underset{N}{\boldsymbol{y}})\right.$$

$$\left. + \frac{1-\alpha}{2}(\partial_i \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_i \ell(\underset{N}{\boldsymbol{y}})(\partial_j \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_j \ell(\underset{N}{\boldsymbol{y}})(\partial_k \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_k \ell(\underset{N}{\boldsymbol{y}})\right]$$

$$= E\left[\left(\partial_i \partial_j \ell(\underset{1}{\boldsymbol{y}}) \partial_k \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_i \partial_j \ell(\underset{N}{\boldsymbol{y}}) \partial_k \ell(\underset{N}{\boldsymbol{y}})\right)\right.$$

$$\left. + \frac{1-\alpha}{2}\left(\partial_i \ell(\underset{1}{\boldsymbol{y}}) \partial_j \ell(\underset{1}{\boldsymbol{y}}) \partial_k \ell(\underset{1}{\boldsymbol{y}}) + \ldots + \partial_i \ell(\underset{N}{\boldsymbol{y}}) \partial_j \ell(\underset{N}{\boldsymbol{y}}) \partial_k \ell(\underset{N}{\boldsymbol{y}})\right)\right]$$

since $\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}$ are independent. Thus

$$\bar{\overset{\alpha}{\Gamma}}_{ijk} = \sum_{m=1}^{N}\left[E \partial_i \partial_j \ell \partial_k \ell + \frac{1-\alpha}{2} E \partial_i \ell \partial_j \ell \partial_k \ell\right] = N \overset{\alpha}{\Gamma}_{ijk}$$

since $\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}$ are identically distributed.

Quote

The following quote (Amari, 1990, p116) sums up these results.

> 'This shows that the metric tensor and the α–connection based on N independent observations are N times those based on one observation. Hence the two geometric structures are similar, and it is not necessary to study them separately.'

D.3.4 Exponential family

For replicates $\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}$ independently and identically distributed with pdf $p(\boldsymbol{y}; \boldsymbol{\theta})$, their joint pdf is

$$\bar{p}(\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}; \boldsymbol{\theta}) = \prod_{k=1}^{N} p(\underset{k}{\boldsymbol{y}}).$$

The log–likelihood becomes

$$\bar{\ell}(\underset{1}{\boldsymbol{y}}, \ldots, \underset{N}{\boldsymbol{y}}) = \ln \bar{p} = \sum_{i=k}^{N} \ell(\underset{k}{\boldsymbol{y}}; \boldsymbol{\theta}).$$

The general exponential family has pdf

$$p(\boldsymbol{y}; \boldsymbol{\theta}) = \theta^i y_i - \Psi(\boldsymbol{\theta}) + c(\boldsymbol{y})$$

which gives

$$\bar{\ell} = \sum_{k=1}^{N} \left(\theta^i \underset{k}{y_i} \Psi(\boldsymbol{\theta}) + c(\boldsymbol{y})\right) = N\left(\theta^i \bar{y}_i - \Psi(\boldsymbol{\theta})\right) + \sum_{k=1}^{N} c(\underset{k}{\boldsymbol{y}})$$

where

$$\bar{y}_i = \left(\underset{1}{y_i} + \underset{2}{y_i} + \ldots + \underset{N}{y_i}\right)/N.$$

Note that $\underset{k}{y_i}$ is the kth replicate of the ith component of the response $\boldsymbol{y}$ where $\boldsymbol{y}^\top = (y_1, y_2, \ldots, y_n)$. If $c(\underset{k}{\boldsymbol{y}}) =$ constant (possibly zero), then

$$\bar{\ell} = N\left(\theta^i \bar{y}_i - \Psi(\boldsymbol{\theta})\right)$$

since then c can be absorbed into the definition of Ψ. An example where this occurs would be the full representation of the Normal distribution, as given in Section 2.4.3. Only in such cases can the result of Amari (1990, p116), viz,

$$\bar{\ell} = N\ell(\bar{y})$$

be invoked. For such models, it can be said that the log–likelihood based on N replicates is N times the log–likelihood based on their mean. For the more familiar exponential family models such as GLMs where the scale parameter is taken as a constant, this relation does not hold, due to the presence of the term $c(\boldsymbol{y})$.

Since then $\bar{\ell}$ and $N\ell(\bar{\boldsymbol{y}})$ differ only by a constant in general then

$$\partial_i \bar{\ell} = N\partial_i \ell(\bar{\boldsymbol{y}})$$

and so

$$\partial_i \ell(\bar{\boldsymbol{y}}) = \bar{y}_i - \Psi'(\boldsymbol{\theta}) = 0$$

to give

$$E\bar{y}_i = \Psi'(\boldsymbol{\theta})$$

as for the single observation case.

Metric tensor

As in Section D.3.2, the metric tensor for replicate data from general exponential families is

$$\bar{g}_{ij} = E\partial_i \bar{\ell}\partial_j \bar{\ell} = N^2 E\left(\bar{y}_i - \Psi'(\boldsymbol{\theta})\right)\left(\bar{y}_j - \Psi'(\boldsymbol{\theta})\right) = N^2 Cov(\bar{y}_i, \bar{y}_j) = Ng_{ij} = N\partial_i\partial_j\Psi.$$

This gives

$$Cov(\bar{y}_i, \bar{y}_j) = \frac{g_{ij}}{N} = \frac{\partial_i\partial_j\Psi}{N}.$$

α–connection

The α–connection for an exponential family model becomes

$$\bar{\overset{\alpha}{\Gamma}}_{ijk} = N\,\overset{\alpha}{\Gamma}_{ijk} = \frac{1-\alpha}{2} NE\partial_i\ell\partial_j\ell\partial_k\ell$$

since $\partial_i\ell\partial_j\ell = -\partial_i\partial_j\Psi$ and $E\partial_k\ell = 0$ being the score statistic. Thus

$$\bar{\overset{\alpha}{\Gamma}}_{ijk} = \frac{1-\alpha}{2} NT_{ijk} = \frac{1-\alpha}{2} N\partial_i\partial_j\partial_k\Psi(\boldsymbol{\theta}).$$

D.3.5 Curved exponential family

The imbedded regression coefficients $\boldsymbol{\beta}$ are defined by

$$\boldsymbol{u} = \boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta}) = \boldsymbol{u}(\boldsymbol{\theta}),$$

so the joint pdf for the replicate data becomes

$$\bar{q}(\boldsymbol{y};\boldsymbol{u}) = p(\underset{1}{y};\boldsymbol{\theta}(\boldsymbol{u})).p(\underset{2}{y};\boldsymbol{\theta}(\boldsymbol{u}))\ldots p(\underset{N}{y};\boldsymbol{\theta}(\boldsymbol{u})).$$

The same caveat applies to curved exponential families as in Section D.3.4, namely that the form of likelihood described in Amari (1987, p39) and Amari (1990, p117), whereby

$$\bar{q} = [p(\bar{\boldsymbol{y}};\boldsymbol{\theta}(\boldsymbol{u})]^N = e^{N(\theta^i\bar{y}_i - \Psi(\boldsymbol{\theta}))}$$

only applies to exponential family models for which $c(\boldsymbol{y}) = 0$ in the pdf

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \theta^i y_i - \Psi(\boldsymbol{\theta}) + c(\boldsymbol{y}).$$

For either form, the differential geometric quantities in terms of the regression coefficients can be obtained for the replicate data.

Metric tensor

The metric tensor for the replicates is

$$\bar{g}_{ab} = E\left(\partial_a\bar{\ell}\partial_b\bar{\ell}\right) = E\left(B^i_a\partial_i\bar{\ell}B^j_b\partial_j\bar{\ell}\right) = B^i_aB^j_bE(\partial_i\bar{\ell}\partial_j\bar{\ell})$$

$$\bar{g}_{ab} = B^i_aB^j_b\bar{g}_{ij} = B^i_aB^j_bNg_{ij} = NB^i_aB^j_bg_{ij} = Ng_{ab}$$

α–connection

The α–connection for the regression coefficients based on the replicate data is

$$\begin{aligned}\bar{\Gamma}^{\alpha}_{abc} &= \partial_a(B^i_b)B^j_c\bar{g}_{ij} + B^i_aB^j_bB^k_c\,\bar{\Gamma}^{\alpha}_{ijk}\\ &= N\left(\partial_a(B^i_b)B^j_cg_{ij} + B^i_aB^j_bB^k_c\,\Gamma^{\alpha}_{ijk}\right)\end{aligned}$$

to give

$$\bar{\Gamma}^{\alpha}_{abc} = N\,\Gamma^{\alpha}_{ijk}\ .$$

Thus, the quote at the end of Section D.3.3 applies equally to curved exponential families.

References

1. Aitkin M., Anderson D., Francis B. and Hinde J., (1989), *Statistical Modelling in GLIM*, Clarendon Press, Oxford.

2. Amari S., (1982a), *Differential geometry of curved exponential families – curvatures and information loss*, Ann. Statist., Vol. 10, No. 2, pp357 – 385.

3. Amari S., (1982b), *Geometrical theory of asymptotic ancillarity and conditional inference*, Biometrika, **69**, 1, pp1-17.

4. Amari S., Barndorff–Nielsen O.E., Kass R.E., Lauritzen S.L. and Rao C.R., (1987), *Differential Geometry in Statistical Inference*, Int. Math. Stat., Lecture Notes–Monograph Series, Vol. 10, Hayward, Calif.

5. Amari S., (1990), *Differential–Geometric Methods in Statistics*, Lecture Notes in Statistics, No. 28, Springer–Verlag, Berlin.

6. Anscombe F.J., (1948), *The transformation of Poisson, Binomial and Negative–binomial data*, Biometrika, **35**, pp246–254.

7. Atkinson A.C., (1980), *Discussion of the paper by D.M. Bates and D.G. Watts*, J. R. Statist. Soc. B, **42**, p21.

8. Barndorff–Nielsen O.E., Cox D.R. and Reid N., (1986), *The Role of Differential Geometry in Statistical Theory*, Int. Statist. Rev., **54**, pp83–96.

9. Barndorff–Nielsen O.E., *in* Amari S., Barndorff–Nielsen O.E., Kass R.E., Lauritzen S.L. and Rao C.R., (1987), *Differential Geometry in Statistical Inference*, Int. Math. Stat., Lecture Notes–Monograph Series, Vol. 10, Hayward, Calif.

10. Bartlett M.S., (1953a), *Approximate Confidence Intervals*, Biometrika, **40**, pp12–19.

11. Bartlett M.S., (1953b), *Approximate Confidence Intervals . II. More than one unknown parameter*, Biometrika, **40**, pp306–317.

12. Bates D.M. and Watts D.G., (1980), *Relative Curvature Measures of Non-linearity*, J. R. Stat. Soc. B, **42**, pp1–25.

13. Bates D.M. and Watts D.G., (1981), *Parameter transformations for improved approximate confidence regions in nonlinear least squares*, Ann. Statist., Vol. 9, No. 6, pp1152-1167.

14. Bates D.M. and Watts D.G., (1988), *Nonlinear Regression Analysis and Its Applications*, Wiley, New York.

15. Beale E.M.L., (1960), *Confidence Regions in non–linear estimation*, J. R. Statist. Soc. B, **22**, pp41–88.

16. Box G.E.P. and Cox D.R., (1964), *An analysis of transformations*, J. R. Stat. Soc. B, **26**, pp211–243.

17. Box M.J., (1971), *Bias in non–linear estimation*, J. R. Stat. Soc. B, **32**, pp171–201.

18. Bishop R.L. and Goldberg S.I., (1980), *Tensor Analysis on Manifolds*, Dover, New York.

19. Clarke G.P.Y., (1980), *Moments of the least–squares estimators in a non–linear regression model*, J. R. Statist. Soc. B., **42**, pp227–237.

20. Cook R.D., (1977), *Detection of influential observations in linear regression*, Technometrics, **19**, pp15–18.

21. Cook R.D. and Tsai C.L., (1985), *Residuals in nonlinear regression*, Biometrika, **72**, pp23–29.

22. Cook R.D. and Weisberg S., (1986), *Residuals and Influence in Regression*, Chapman and Hall, London.

23. Cordeiro G.M., (1983), *Improved Likelihood Ratio Statistics for Generalized Linear Models*, J. R. Statist. Soc. B, **45**, No. 3, pp404–413.

24. Cordeiro G.M. and McCullagh P., (1991), *Bias correction in Generalized Linear Models*, J. R. Statist. Soc. B, **53**, No. 3, pp629–643.

25. Cox D.R. and Hinkley D.V., (1978), *Problems and Solutions in Theoretical Statistics*, Chapman and Hall, London.

26. Cox D.R and Hinkley D.V. (1982), *Theoretical Statistics*, Chapman and Hall, London.

27. Dawid A.P. (1975), *Discussion of the paper by B. Efron*, Ann. Statist., **3**, pp1231–1234.

28. DiCiccio T.J., (1984), *On parameter transformations and interval estimation*, Biometrika, **71** (3), pp477–485.

29. Dobson A.J., (1993), *An Introduction to Generalized Linear Models*, Chapman and Hall, London.

30. Draper N. and Smith H., (1981), *Applied Regression Analysis*, 2nd ed., Wiley, New York.

31. Efron B., (1975) *Defining the curvature of a statistical problem (with applications to second order efficiency)* (with discussion), Ann. Statist., **3**, pp1189–1242.

32. Efron B., (1978), *The geometry of exponential families*, Ann. Statist., **6**, pp362–376.

33. Ellem B.A., (1992a), *The Statistical Interpretation of α-connections*, Technical Report No. 9201, School of Information Technology, Charles Sturt University–Mitchell, May 1992, pp1–7.

34. Ellem B.A., (1992b), *The Differential Geometry of One Dimensional Generalized Linear Models*, Technical Report No. 9202, School of Information Technology, Charles Sturt University–Mitchell, December 1992, pp1–15.

35. Ellem B.A., (1994), *The Exponential Connection and Canonical Links for Generalized Linear Models*, Technical Report No. 9401, School of Information Technology, Charles Sturt University–Mitchell, June 1994, pp1–11.

36. Ellem B.A., (1995), *Curvature Measures for Non–Normal Error Models*, Department of Applied Statistics, University of Wollongong, Preprint No. 1/95.

37. GENSTAT 5, (1993), Release 3, Reference manual, Genstat 5 Committee, Clarendon Press, Oxford.

38. Green P.J., (1984), *Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives* (with discussion), J. R. Stat. Soc. B, **46**, pp149–192.

39. Hartley H.O., (1961), *The modified Gauss–Newton method for the fitting of nonlinear regression functions by least squares*, Technometrics, **3**, pp269–280.

40. Hougaard P., (1982), *Parameterizations of Nonlinear Models*, J. R. Stat. Soc. B, **44**, No. 2, pp244–252.

41. Jeffreys H., (1961), *Theory of Probability*, 3rd. ed., Clarendon Press, Oxford.

42. Jorgennsen B., (1983), *Maximum likelihood estimation and large–sample inference dor generalized linear and nonlinear regression models*, Biometrika, **70**, pp19–28.

43. Kass R.E., (1980), *The Riemann Structure of Model Spaces : A geometrical approach to inference*, PhD dissertation, University of Chicago.

44. Kass R.E., (1984), *Canonical Parameterizations and Zero Parameter-Effects Curvature*, J. R. Stat. Soc. B, **46**, No. 1, pp86–92.

45. Kass R.E., (1989), *The Geometry of Asymptotic Inference*, Statistical Science, Vol. 4, No. 3, pp188–234.

46. Kass R.E. and Smyth G.K., (1990), *The Rate of Convergence of Fisher Scoring : A Geometric Interpretation*, Technical Report, Department of Mathematics, University of Queensland.

47. Kay D.C., (1988), *Theory and Problems of Tensor Calculus*, Schaum's Outline Series, McGraw–Hill, New York.

48. Kendall M.G. and Buckland W.R., (1971), *A Dictionary of Statistical Terms*, 3rd Ed., Oliver and Boyd, Edinburgh.

49. Kincaid D. and Cheney W., (1991), *Numerical Analysis*, Brooks/Cole, Pacific Grove.

50. Kreyszig E., (1991), *Differential Geometry*, Dover, New York.

51. Lauritzen S.L., *in* Amari S., Barndorff–Nielsen O.E., Kass R.E., Lauritzen S.L. and Rao C.R., (1987), *Differential Geometry in Statistical Inference*, Int. Math. Stat., Lecture Notes–Monograph Series, Vol. 10, Hayward, Calif.

52. Lovelock D. and Rund H., (1989), *Tensors, Differential Forms and Variational Principles*, Dover, New York.

53. Lowry R.K. and Morton R., (1983), *An asymmetry measure for estimators in nonlinear regression models*, Proc. 44th Session International Statistical Institute, Madrid, Contributed Papers, Vol 1, pp351–354.

54. McCullagh P., (1980), *Discussion of the paper by D.M. Bates and D.G. Watts*, J. R. Statist. Soc. B, **42**, p22.

55. McCullagh P., (1987), *Tensor Methods in Statistics*, Chapman and Hall, London.

56. McCullagh P. and Nelder J.A., (1983), *Generalized Linear Models*, Chapman and Hall, London.

57. McCullagh P. and Nelder J.A., (1989), *Generalized Linear Models*, 2nd Ed., Chapman and Hall, London.

58. Michaelis L. and Menten M.L., (1913), *Kinetik der Invertinwirkung*, Biochemsche Zeitschrift, **49**, p333.

59. Morgan F., (1993), *Riemannian Geometry ; A Beginners's Guide*, Jones and Bartlett, Boston.

60. Murray M.K. and Rice J.W., (1993), *Differential Geometry and Statistics*, Chapman and Hall, London.

61. NAG (Numerical Algorithms Group), (1985), *The GLIM System Release 3.77 Manual*, (ed. C.D. Payne), NAG, Oxford.

62. Nelder J.A. and Wedderburn R.W.M., (1972), *Generalized Linear Models*, J.R. Statist. Soc. A, **135**, pp370–384.

63. Norusis M.J., (1993), SPSS for Windows, Base System User's Guide, Release 6.0, SPSS Inc., Chicago.

64. Pregibon D., (1980), *Goodness of Link Tests for Generalized Linear Models*, Applied Statistics, **29**, No. 1, pp15–24.

65. Ratkowsky D.A., (1983), *Nonlinear Regression Modeling: A Unified Practical Approach*, Marcel Dekker, New York.

66. Reeds J., (1975), *Discussion of the paper by B. Efron*, Ann. Statist., **3**, pp1234–1238.

67. Reid N., (1980), *Discussion of the paper by D.M. Bates and D.G. Watts*, J. R. Statist. Soc. B, **42**, p20.

68. Ross G.J.S., (1980a), *Discussion of the paper by D.M. Bates and D.G. Watts*, J. R. Statist. Soc. B, **42**, pp19-20.

69. Ross G.J.S., (1980b), *Uses of non–linear transformations in non–linear optimisation problems*, In : COMPSTAT 1980, Wien : Physica–Verlag, pp381–388.

70. Ross G.J.S., (1990), *Nonlinear Estimation*, Springer–Verlag Series in Statistics, New York.

71. Sadler D.R., (1975), *Numerical Methods for Nonlinear Regression*, University of Queensland Press, St. Lucia, Queensland.

72. Seber G.A.F. and Wild C.J., (1989), *Nonlinear Regression*, Wiley, New York.

73. Smyth G.K., (1987), *Curvature and Convergence*, Proceedings of the Statistical Computing Section, American Statistical Society, pp278–283.

74. Spain B., (1960), *Tensor Calculus*, Oliver and Boyd, Edinburgh.

75. Spiegel M.R., (1990), *Vector Analysis*, Schaum's Outline Series, McGraw–Hill, New York.

76. Stoker J.J., (1969), *Differential Geometry*, Wiley–Interscience, New York.

77. Struik D.J., (1988), *Lectures on Classical Differential Geometry*, 2nd Ed., Dover, New York.

78. Vos P.W., (1987), *Dual Geometries and their application to Generalized Linear Models*, PhD dissertation, University of Chicago.

79. Wei B.C., (1994), *On confidence regions of embedded models in regular parametric families (a geometric approach)*, Austral. J. Statist., **36** (3), pp327–338.

80. Wei B.C. and Zhu H.T., (1997), *Some second order asymptotics in exponential family non–linear regression models (a geometric approach)*, Austral. J. Statist., **39** (2), pp129–148.

81. Wei B.C., (1998), *Exponential Family Nonlinear Models*, Lecture Notes in Statistics, No. 130, Springer, Singapore.

82. Weisberg S., (1985), *Applied Linear Regression*, 2nd ed., Wiley, New York.