Chapter 1

Introduction

1.1 Objective

The aim of this thesis is to develop and explore a general methodology for dealing with problems involving many nuisance parameters. Nuisance parameters, sometimes referred to as incidental parameters, are parameters that are needed in order to make a model realistic, but that are of no intrinsic interest in themselves. Neyman and Scott (1948) provides a fundamental introduction to nuisance parameter problems, and McCullagh and Nelder (1989, Chapter 7) provides a discussion of nuisance parameters in a practical setting.

Example 1 *Bernoulli matched pairs*

To illustrate the natural way in which nuisance parameters arise, consider the allocation of two treatments to matched pairs of individuals, so that one individual obtains treatment 1 say, and the other receives treatment 2. A binary response is observed on each individual. A common model for this situation is to postulate

$$logit(\pi_{i1}) = \lambda_i \text{ and } logit(\pi_{i2}) = \lambda_i + \psi,$$

where $\pi_{ij} = P(X_{ij} = 1)$, and X_{ij} is the binary random variable representing the outcome on the j^{th} individual in the i^{th} pair. Here, λ_i is a pair specific factor, needed to capture the fact that there is some commonality within the pair, but that the pairs themselves are heterogeneous. The parameter ψ is a measure of the difference

in effect between the two treatments. In this setting the λ_i are nuisance parameters and ψ is the parameter of interest.

Nuisance parameters need not be elements of $\mathbb{R}^p$; they may in fact be functions. In Cox's proportional hazards model (Cox, 1972) where the hazard function $h(t, \mathbf{x})$ is assumed to be of the form $h(t, \mathbf{x}) = \lambda_0(t)exp(\boldsymbol{\beta}^T\mathbf{x})$, the baseline hazard function, $\lambda_0(\cdot)$, is an unspecified function of t and plays the role of nuisance parameter. This is an example of a wide class of models known as *semi-parametric models*, in which the nuisance parameter may be infinite dimensional. However, nuisance parameters of a functional form will not be considered in this thesis.

Nuisance parameter problems need special attention because the presence of nuisance parameters in increasing numbers may lead to the breakdown of the classic optimality results of maximum likelihood theory, namely the consistency and efficiency of the maximum likelihood estimate for the parameter of interest. This thesis is an attempt to explore in detail a methodology that provides a partial repair to the theory of maximum likelihood.

1.2 Overview

The remainder of this chapter examines through examples the statistical consequences of models with many nuisance parameters and describes some of the available methodologies for dealing with them, for example Barndorff-Nielsen (1983), Cox and Reid (1987), McCullagh and Tibshirani (1990), DiCiccio and Efron (1992). The seminal Neyman-Scott problem is introduced, which will form a focus for much of the subsequent work. Traditional approaches to nuisance parameter problems are reviewed, as well as some more recent developments. The work of Small and

McLeish (1988) is especially relevant, since some of the core ideas used in this thesis can be found in their work. The thesis represents a considerable expansion of their results: here, a multivariate parameter of interest is explicitly considered, a "rectangular array" asymptotic embedding is developed, the investigation of constructive properties of the projective scores is presented, and an interpretative framework in terms of certain model curvatures for some of the key results is given. Furthermore, comparisons are made to several alternative methodologies.

Chapter 2 focuses on important work by Barankin (1949), Bahadur (1957) and Lindsay (1982, 1983) that is used as a foundation for the theory developed later. In particular, the optimality properties of conditional score functions are described and the space generated by the centered likelihood ratios is introduced. Notation is established, and the Bhattacharyya basis is discussed. The projective scores are defined and their finite-sample optimality properties established.

In Chapter 3 the Bhattacharyya functions are investigated. Some simple, yet fundamental identities are derived that provide an essential tool for the analysis of the projective score methodology. A notation for the Bhattacharyya basis functions is introduced that is taken from the work of McCullagh (1987), where it was used in the context of generalized cumulants. A class of matrices, which are referred to as *permutation invariant* are discussed. They arise naturally in the projective score methodology and display an inherent structure that allows major simplification. Essentially, the structure of the matrices allows for the construction of a simple Moore-Penrose generalized inverse. This simplicity has important practical consequences, as demonstrated in Chapter 7, where the projective score methodology is applied to the class of canonically parameterized Generalized Linear Models.

Chapter 4 contains a series of asymptotic results that establish why the

projective score methodology is a worthwhile undertaking. The key idea is that it enables bias in the estimating equations to be reduced, which is fundamental to obtaining consistent estimates.

This bias reduction is critical to the results presented in Chapter 5, which is concerned with a type of asymptotics termed *rectangular array*. This is an asymptotic embedding relevant to Neyman-Scott type problems in which the number of observations in each strata, and the number of nuisance parameters are allowed to tend to infinity *simultaneously*. The important result obtained is that in the square array asymptotics it is possible to obtain consistent and *fully* efficient estimates of the parameter of interest, but in order to achieve this full efficiency the bias-reduced U_2 projective score function must be used.

Chapter 6 is concerned with *curvature*. By an appropriate reparameterization of the nuisance parameters, the limiting bias term in the U_1 projective score and the information correction between the U_1 and U_2 projective scores can be identified in terms of certain model curvatures. The curvatures are associated with the surface defined by the expectation of the score for the "least favorable family distribution", but measured in a *mean value* metric. Comparisons are made to the work of Efron (1985).

Chapter 7 is an application of the methodology to the class of canonically parameterized Generalized Linear Models. The aim is to demonstrate the feasibility of applying the projective score methodology to a challenging and realistic practical problem. This is successfully achieved and some numerical results are presented for a Poisson regression problem.

Finally, Chapter 8 details some further directions for the continuation of the line of research undertaken in this thesis.

1.3 Inferential consequences of models with nuisance parameters

A traditional proving ground for methodologies that attempt to deal with nuisance parameters is the example of many normal means introduced by Neyman and Scott (1948).

Example 2 *Neyman-Scott problem*

Let X_{ij}, for $i = 1, 2, \ldots, s$ and $j = 1, 2, \ldots, n$ be independent normally distributed random variables with common variance ψ but different means λ_i. The set up is that of a balanced One-way Analysis of Variance with s treatments or strata and n observations within each treatment. If the number of treatments is held fixed and the number of observations within each treatment goes to infinity, then the standard maximum likelihood theory (Wald, 1949) shows the *m.l.e.* of ψ to be both consistent and efficient. However, if the number of observations is held fixed but the number of treatments is allowed to go to infinity, then the *m.l.e.* of ψ is no longer consistent, in fact $\hat{\psi} \xrightarrow{p} \frac{n-1}{n}\psi$.

This example is similar to example 1 in the sense that the number of nuisance parameters increases with the strata sample size, and these nuisance parameters remain unknown even as the sample size approaches infinity.

The breakdown of maximum likelihood theory can be attributed to *substitution bias*, where the concept of substitution bias is illustrated as follows:

The score equations for the i^{th} stratum are

$$U_i(\psi, \lambda_i) = \frac{\partial}{\partial \psi} ln\, f(\psi, \lambda_i, \mathbf{x}_i) = -\frac{n}{2\,\psi} + \frac{1}{2\,\psi^2}\sum_{j=1}^{n}(x_{ij} - \lambda_i)^2, \tag{1.1}$$

$$V_i(\psi, \lambda_i) = \frac{\partial}{\partial \lambda_i} ln\, f(\psi, \lambda_i, \mathbf{x}_i) = \frac{1}{\psi}\sum_{j=1}^{n}(x_{ij} - \lambda_i). \tag{1.2}$$

Treating ψ as fixed, the solution to (1.2) is $\hat{\lambda}_{i\,\psi} = \overline{\mathbf{x}}_i$. If this estimate for λ_i is "substituted" back into (1.1), then the resulting *derivative of the profile log-likelihood*, $U_i(\psi, \hat{\lambda}_{i\,\psi})$, has the property that

$$E[U_i(\psi, \hat{\lambda}_{i\,\psi})] = E\left[-\frac{n}{2\,\psi} + \frac{1}{2\,\psi^2}\sum_{j=1}^{n}(x_{ij} - \hat{\lambda}_i)^2\right] = -\frac{1}{2\,\psi}, \tag{1.3}$$

which is non-zero. This is called substitution bias. McCullagh (1990, p.329) presents an argument illustrating why substitution bias exists in general, not just in this example.

A second example in Neyman and Scott (1948) illustrates a case where the maximum likelihood estimate is consistent, but fails to be efficient:

Example 3 *Unbalanced, heteroskedastic, one-way analysis of variance*

The statistical experiment consists of observing n_i normal random variables in the i^{th} stratum. All strata have a common mean μ, but the variances σ_i^2 differ between strata. Neyman and Scott establish the consistency of the maximum likelihood estimate of μ under certain mild regularity conditions, but also construct another consistent estimator of μ that is *uniformly more efficient* than the *m.l.e.* provided not all the n_i are equal. To see this consider the score equation for the parameter of interest, after the *m.l.e.*'s for the nuisance parameters have been substituted:

$$\sum_{i=1}^{p} \frac{n_i(\overline{x}_i - \mu)}{S_i^2 + (\overline{x}_i - \mu)^2}, \tag{1.4}$$

where $n_i S_i^2 = \sum_j (x_{ij} - \overline{x}_i)^2$. Neyman and Scott consider a more general version of (1.4) in which arbitrary weights w_i are allowed:

$$\sum_{i=1}^{p} w_i \frac{(\overline{x}_i - \mu)}{S_i^2 + (\overline{x}_i - \mu)^2}. \tag{1.5}$$

The maximum likelihood estimator of μ, solves (1.5) but with the weights $w_i = n_i$. The solution to (1.5) is shown to be consistent under quite mild regularity assumptions. Denoting the asymptotic variance of the *m.l.e.* as $\sigma^2_{\hat{\mu}}$ and the asymptotic variance of the root of (1.5) as σ^2_{α}, Neyman and Scott express the former in terms of the latter:

$$\sigma^2_{\hat{\mu}} = \sigma^2_{\alpha} + \frac{\sum_{i=1}^{p} \frac{n_i - 2}{\sigma_i^2} \left(\frac{n_i}{n_i - 2} - V \right)^2}{\left(\sum_{i=1}^{s} \frac{n_i}{\sigma_i^2} \right)^2},$$

where

$$V = \frac{\sum_{i=1}^{p} \frac{n_i - 2}{\sigma_i^2} \frac{n_i}{n_i - 2}}{\sum_{i=1}^{p} \frac{n_i - 2}{\sigma_i^2}}.$$

V is a "weighted sum of the ratios $\frac{n_i}{n_i - 2}$". Clearly $\sigma^2_{\hat{\mu}} \geq \sigma^2_{\alpha}$, which indicates the possible inefficiency of the *m.l.e.* Equality of the two asymptotic variances is only achieved when $\frac{n_i}{n_i - 2} = V$ for all i, which is equivalent to all the sample sizes n_i being constant. Neyman and Scott make an interesting remark:

> *The authors must confess that, to their mind, the possible lack of asymptotic efficiency of the maximum-likelihood estimates when they are consistent appears to be more surprising than the possibility of bias.*

1.4 Approaches for dealing with nuisance parameters

1.4.1 Traditional methods

The traditional approach to dealing with nuisance parameters is to reduce the problem to one in which the nuisance parameters have been eliminated, ideally with as little sacrifice of information as possible about the parameter of interest. In a limited, but often important set of problems, reduction to a relevant *conditional likelihood* or a *marginal likelihood* that is free of the nuisance parameter is possible.

For example, in the Bernoulli matched pairs (Example 1), a complete sufficient statistic for the nuisance parameter λ_i is $X_{i1} + X_{i2}$. The outcome of any pair must be of the form $(0,0)$, $(0,1)$, $(1,0)$ or $(1,1)$. Let a, b, c and d denote the number of pairs of each form that are observed respectively. With $X_{i1} + X_{i2}$ as the conditioning event, the conditional likelihood is free from λ_i, so that the nuisance parameter has been completely eliminated. The conditional log likelihood is given by

$$l_c(\psi) = c\psi - (b+c)\log(1+\exp\psi).$$

Note that it depends on the data only through the discordant pairs. The reduction to the conditional likelihood forms the basis for the construction of the well known McNemar's test (Agresti, 1990, p. 350).

As an illustration of a situation where reduction to a marginal likelihood is possible consider:

Example 4 *Circular normal problem*

The experiment consists of observing pairs of independent normal random variables Y_{i1} and Y_{i2}, where

$$Y_{i1} \sim N(\rho\cos\phi_i, 1) \qquad Y_{i2} \sim N(\rho\sin\phi_i, 1).$$

The data pair (y_{i1}, y_{i2}) constitutes an observation from a bivariate normal distribution, whose mean is located on a circle centered at the origin but of unknown radius. The radius parameter ρ is taken to be the parameter of interest, and the unknown angle ϕ_i is the nuisance parameter. This model is of great theoretical interest, but also has practical applications, see Anderson (1981).

The joint density for any particular pair is given by

$$\frac{1}{2\pi}\exp\left(-\frac{1}{2}\left((y_{i1}-\rho\cos\phi)^2+(y_{i2}-\rho\sin\phi)^2\right)\right).$$

Marginal inference is possible in this problem since the distribution of $R_i^2 = Y_{i1}^2 + Y_{i2}^2$ does not depend on ϕ_i. R_i^2 has a non-central χ^2 distribution on 2 degrees of freedom, with non-centrality parameter ρ^2. The marginal likelihood is

$$l_m = \sum_{k=0}^{\infty} \frac{\exp\left(-\rho^2/2\right)\left(\frac{\rho^2}{2}\right)^k}{k!} \frac{(r_i^2)^k \exp\left(-r_i^2/2\right)}{k!\ 2^{k+1}},$$

and the absence of ϕ_i in this expression demonstrates the successful elimination of the nuisance parameter.

For the majority of problems it will not be possible to construct relevant likelihoods for inference that are free of the nuisance parameter and a more general approach is needed in dealing with complex situations. The next section describes some recent developments along these lines. Their motivation comes primarily from nuisance parameter problems in which standard asymptotic embeddings are considered. That is, problems where in the limit each nuisance parameter appears in the distribution of an infinite number of random variables, and there are a finite number of nuisance parameters.

It is important to note that the primary concern of this thesis are Neyman-Scott type problems where each nuisance parameter appears in the probability law of a finite number of random variables, but the parameter of interest appears in the probability laws of an infinite number of random variables. When Neyman and Scott refer to *incidental parameters* and *structural parameters* they mean precisely parameters appearing in the above context.

1.4.2 Recent developments

Standard maximum likelihood theory provides two fundamental results for hypothesis testing and confidence interval construction. Assume that the experiment comprises observing n independent identically distributed (*i.i.d.*) observations, where the underlying probability model has univariate parameters (ψ, λ). It is desired to test the hypothesis that $\psi = \psi_0$. Then the *likelihood ratio test statistic*

$$2\{l(\tilde{\psi}, \tilde{\lambda}) - l(\psi_0, \hat{\lambda}_{\psi_0})\} \tag{1.6}$$

has an asymptotic χ^2 distribution on 1 degree of freedom. Here, tilde and circumflex denote maximum likelihood estimates under the full and restricted models respectively, and l denotes the log-likelihood.

The relevant result for confidence interval construction is that the *m.l.e.* of ψ has a limiting behavior:

$$\sqrt{n}(\psi - \psi_0) \sim N(0, i^{\psi\psi}),$$

where $i^{\psi\psi}$ is the appropriate term in the inverse of the per observation *Fisher's Information Matrix*, defined as $E[\nabla l \nabla l^T]$. This distributional result can be used to form the approximate pivotal quantity

$$\frac{\hat{\psi} - \psi_0}{\sqrt{i^{\psi\psi}}}, \tag{1.7}$$

which has a limiting standard normal distribution. Confidence intervals for ψ can be constructed from this pivotal.

Many of the recent developments can be viewed as refinements of these classical results, in which the objective of the refinement procedure is to produce a test statistic or pivotal that is better approximated by the limiting distribution.

Barndorff-Nielsen (1983) presents a formula for the conditional distribution of the maximum likelihood estimator given a maximal ancillary statistic. The formula is used to obtain a *modified profile likelihood*

$$L^0(\psi; u) = \det \frac{\partial \hat{\lambda}}{\partial \lambda_\psi} \det\, n j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) L(\psi, \hat{\lambda}_\psi), \tag{1.8}$$

that it is hoped will alleviate some of the inferential difficulties associated with the standard *profile likelihood*, $L(\psi, \hat{\lambda}_\psi)$. Note the dependence in (1.8) on u, the ancillary statistic. Here the term $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ is the observed information per observation for λ, that is $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) = -\frac{1}{n}\nabla^2 l$, where l is the log-likelihood.

The logarithm of the modified profile likelihood, denoted by l_{mp}, is given by

$$l_{mp}(\psi) = l(\psi, \hat{\lambda}_\psi) - \tfrac{1}{2}\log(\det\, n j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)) + \log(\det(d\hat{\lambda}/d\hat{\lambda}_\psi)), \tag{1.9}$$

where the first correction term corresponds to a variance stabilizing transform of the parameter λ and the second to the *"possible mathematical inelegance of the parameterization"*. An important property of (1.9) is that it is parameterization invariant in the sense of interest preserving transformations:

$$(\psi, \lambda) \to (\psi', \lambda') \qquad \text{where} \qquad \psi' = \psi \qquad \lambda' = g(\lambda, \psi)$$

and $g(\cdot\,, \psi)$ is invertible for all values of ψ. This is an intuitively desirable property because if nuisance parameters are genuinely not of interest, then any procedure that gives different results depending on the chosen parameterization is being unduly influenced by parameters, that are by definition, of no interest.

It is useful to consider situations under which the second of the corrections terms $\log(\det(d\hat{\lambda}/d\hat{\lambda}_\psi))$ will be zero since it can be a troublesome term to evaluate. It disappears in a number of situations including exponential family models of the

form

$$a(\eta)b(x)\exp(\eta^T t),$$

where $t = (u, v)$ and $\eta = (\eta_1, \eta_2)$ are partitions of the canonical statistic and parameters respectively. Denote the mean τ of t by $\tau = (\phi, \lambda)$. Then $\hat{\lambda}_{\eta_1} = \hat{\lambda} = v$. This situation corresponds to a "mixed" parameter, in the sense that the parameter of interest is a canonical parameter and the nuisance parameter is a mean parameter.

In some important instances the modified profile likelihood is equal to or approximates well an appropriate conditional or marginal likelihood. This provides a strong justification for its use. However, it should be noted that the the arguments are derived based on the existence of the ancillary statistic u (a statistic whose distribution depends on ψ alone). The identification of an ancillary statistic, the computation of the correction terms, and as mentioned previously particularly the computation of $d\hat{\lambda}/d\hat{\lambda}_\psi$ is a non-trivial matter, so much so that the methodology may be impossible to apply in a general model.

Another method for dealing with nuisance parameters, also proposed by Barndorff-Nielsen (1986), is a technique that adjusts the signed root of the log-likelihood ratio statistic, in order to obtain a higher order approximation to the normal distribution. This adjusted signed root can be used for setting confidence intervals for the parameter of interest, that have a higher order of accuracy than the standard ones, see (1.7). Following a similar notation to Barndorff-Nielsen, let $\mathcal{M}$ denote the model (ψ, λ), and $\mathcal{M}_{\psi_0}$ a sub-model of $\mathcal{M}$ obtained by fixing $\psi = \psi_0$. Maximum likelihood estimation under $\mathcal{M}$ and $\mathcal{M}_{\psi_0}$ is again denoted by tilde and circumflex respectively. The log-likelihood ratio statistic for testing $\mathcal{M}_{\psi_0}$ against $\mathcal{M}$ is

$$w = 2\left(l(\tilde{\psi}, \tilde{\lambda}) - l(\psi_0, \hat{\lambda}_{\psi_0})\right),$$

where l denotes the log-likelihood function. The signed log-likelihood ratio statistic is defined as

$$r = sgn(\tilde{\psi} - \psi_0)\sqrt{w}, \tag{1.10}$$

In the case where the dimension of ψ is 1, r has an asymptotic normal distribution with first 4 cumulants

$$[\alpha_n, \beta_n, 0, 0] \tag{1.11}$$

to order $O(n^{-1})$, the errors being of order $O(n^{-3/2})$. A standard normal distribution has the first 2 cumulants equal to 0 and 1 respectively, all other cumulants being zero. Under certain regularity conditions (McCullagh, 1987, see p.187) the order of α_n and $1 - \beta_n$ in (1.11) are $O(n^{-1/2})$ and $O(n^{-1})$.

The aim of Barndorff-Nielsen's approach is to correct r so that the modified version, r^*, has first 4 cumulants equal to $[0, 1, 0, 0]$ with errors of order $O(n^{-3/2})$. The correction to r is

$$r^* = r - \frac{\log \mathcal{K}}{r},$$

where the reader is referred to (Barndorff-Nielsen, 1986, (3.10) p.316) for the definition of $\mathcal{K}$.

Cox and Reid (1987) suggest a method similar to the modified profile likelihood of Barndorff-Nielsen. The motivation is as follows: the generalized likelihood ratio statistic $w = 2\left(l(\tilde{\psi}, \tilde{\lambda}) - l(\psi_0, \hat{\lambda}_{\psi_0})\right)$ for models in which a complete sufficient statistic exists for λ at each fixed ψ_0, can be constructed from the conditional distribution of the observations given the complete sufficient statistic. If this conditional distribution is free from λ even when the model is incorrectly specified, *i.e.* $\psi \neq \psi_0$, then the problem has been successfully reduced to one only involving ψ. This approach is not always available, but does work in exponential family problems with

ψ a component of the canonical parameter.

In full exponential family problems the standard analysis is to condition on the components of the sufficient statistic that correspond to the nuisance parameters. These are precisely the maximum likelihood estimates of the mean value parameters which are orthogonal to the canonical parameters. This motivates the general approach of using an orthogonal parameterization for the problem and using the conditional distribution of the data given the *m.l.e.* for the nuisance parameter, $\hat{\lambda}_{\psi_0}$. An *orthogonal parameterization* is defined as one in which the off-diagonal elements of the Fisher's Information matrix are all 0, or equivalently one in which all the score functions are uncorrelated. By using an orthogonal parameterization the dependence of $\hat{\lambda}_\psi$ on ψ is reduced in the sense that it varies only slowly with ψ. More precisely $\hat{\lambda}_\psi - \hat{\lambda} = O_p(1/n)$ rather than the usual $O_p(1/\sqrt{n})$.

The conditional density given $\hat{\lambda}_{\psi_0}$ is

$$f_Y(y;\psi,\lambda)J(\hat{\lambda}_{\psi_0})/f_{\Lambda_0},$$

where $J(\hat{\lambda}_{\psi_0})$ is the Jacobian of the transformation from y to $(\hat{\lambda}_{\psi_0}, h)$ and h is any convenient function of the data. The density f_{Λ_0} denotes the marginal distribution of the *m.l.e.* of λ with ψ fixed at ψ_0. This likelihood leads to a conditional profile likelihood ratio statistic of

$$2\left[\sup_{\psi}\{l_Y(\psi,\hat{\lambda}_{\psi_0}) - l_{\Lambda_0}(\psi,\hat{\lambda}_{\psi_0})\} - \{l_Y(\psi_0,\hat{\lambda}_{\psi_0}) - l_{\Lambda_0}(\psi_0,\hat{\lambda}_{\psi_0})\}\right]. \tag{1.12}$$

The above expression has the important property of being invariant under 1 to 1 transformations of λ. The second term in (1.12) presents some difficulty in that the density is being evaluated at ψ whereas the conditioning value of ψ was ψ_0. Cox and Reid suggest a way out by conditioning the first event at ψ rather than ψ_0 leading to what is referred to as a *double conditioning argument.* The use of Barndorff-Nielsen's

formula for the distribution of the *m.l.e.* (Barndorff-Nielsen, 1983) simplifies the resulting expression so that the approximate conditional modified likelihood ratio statistic is

$$\sup_{\psi}[l_Y(\psi,\hat{\lambda}_\psi)-\tfrac{1}{2}\log\det\{n\,j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)\}]-[l_Y(\psi_0,\hat{\lambda}_{\psi_0})-\tfrac{1}{2}\log\det\{n\,j_{\lambda\lambda}(\psi_0,\hat{\lambda}_{\psi_0})\}], \tag{1.13}$$

where $j_{\lambda\lambda}$ is the per observation observed information matrix for the nuisance parameters.

It follows that the modification to the usual profile log-likelihood is

$$l_{cp}=l_Y(\psi,\hat{\lambda}_\psi)-\tfrac{1}{2}\log\det\{n\,j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)\}. \tag{1.14}$$

Note the similarity between this correction and Barndorff-Nielsen's correction (1.9). Obviously, when the term $\log(\det(d\hat{\lambda}/d\hat{\lambda}_\psi))$ in (1.9) is zero the two corrections are identical.

Some important aspects of the methodology are:

1. It is only appropriate for a univariate parameter of interest.

2. The parameter of interest and the nuisance parameter are required to be orthogonal.

3. The approximations to (1.12) are not invariant under transformations of the nuisance parameter.

Even though it is in theory always possible to achieve an orthogonal parameterization by solving a set of partial differential equations, this will not necessarily be an easy task which may well discourage the use of the methodology; see, for example, Lesperance (1989, p.22).

The work by DiCiccio and Efron (1992) builds on earlier work by Efron (1985, 1987) in which the goal was to improve bootstrap confidence intervals. (Efron, 1985) introduces a class of problems in which observations are drawn from a k-dimensional multivariate normal distribution, that is $Y \sim MVN_k(\eta, I)$ where I is a $k \times k$ identity matrix, and inference is desired for some real valued function $\theta = t(\eta)$. Efron's original motivation was to work out an analytical correction to the standard confidence intervals for θ, $\hat{\theta} \pm \hat{\sigma} z^{(\alpha)}$ ($\hat{\theta}$ is the *m.l.e.* of θ, $\hat{\sigma}$ is the estimate of its variance, and $z^{(\alpha)}$ is the α-percentile of the standard normal), and then to show the close agreement between the bootstrap confidence intervals and the corrected intervals. As the work evolved, the emphasis moved away from the bootstrap and toward an analysis of the correction term in the wider context of exponential family models.

As an example of the approach Efron takes in his 1985 paper, the Circular Normal problem (Example 4) will be considered. Assume that just one observation from the bivariate normal distribution is taken, so that the data comprises the pair (y_1, y_2). Denoting the means of the bivariate normal variate by η_1 and η_2, the parameter of interest ρ is $\sqrt{\eta_1^2 + \eta_2^2} = t(\eta_1, \eta_2)$. The unrestricted *m.l.e.*'s of η_1 and η_2, that is $\tilde{\eta}$, are y_1 and y_2 respectively, and the invariance of the *m.l.e.* implies that the *m.l.e.* of ρ is $t(y_1, y_2)$. If η_1 and η_2 are restricted to lie on a curve, say $\mathcal{C}_{\rho_0}$, where $\mathcal{C}_{\rho_0} = \{\eta_1, \eta_2 : t(\eta_1, \eta_2) = \rho_0\}$, then the restricted *m.l.e.*'s for η_1 and η_2, $\hat{\eta}$, are the values of η_1 and η_2 on the curve that are closest to the data. The situation is illustrated in Figure 1.1. Efron refers to the curves $\mathcal{C}_{\rho_0}$ as the "level curves", or in higher dimensional problems the "level surfaces".

If the signed distance between the unrestricted and restricted *m.l.e.*'s for η_1 and η_2 is denoted by X_{ρ_0}, then Efron suggests the use of the normalized signed

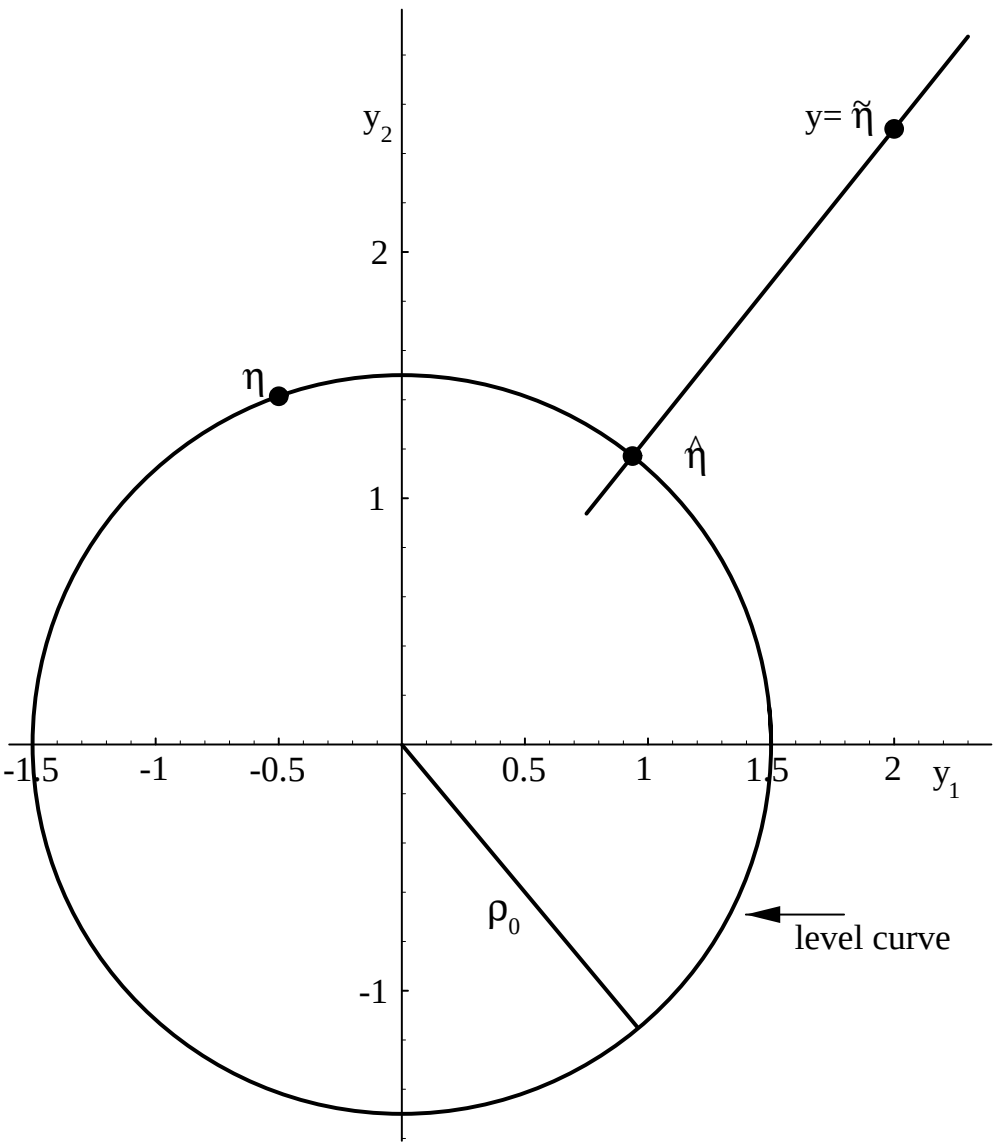

Figure 1.1: Circular normal model displaying the "level curve".

distance

$$W_{\rho_0} \equiv \frac{X_{\rho_0} - tr(d_{\hat{\eta}(\rho_0)})}{1 - tr(d^2_{\hat{\eta}(\rho_0)})}, \tag{1.15}$$

as a pivotal quantity for the construction of confidence intervals. The quantity $d_{\hat{\eta}(\rho)}$ is one half of the usual curvature matrix for the level curves at the model ρ_0. For normal theory models $X^2_{\rho_0}$ is in fact the likelihood ratio test statistic for the hypothesis that $\rho = \rho_0$, so that Barndorff-Nielsen's adjustment (1.10) can be thought of as a generalization of Efron's (Barndorff-Nielsen, 1986, p.321). In the circular normal problem the curvature is simply $\frac{1}{2\rho_0}$, so that the pivotal is

$$W_{\rho_0} = \frac{X_{\rho_0} - \frac{1}{2\rho_0}}{1 - \frac{1}{4\rho_0^2}}. \tag{1.16}$$

As formula (1.15) shows, the curvature of the level surfaces at the model plays a key role in the correction to the standard pivotal. If the curvature is globally 0 then the signed distance X_{ρ_0} has an *exact* standard normal distribution and no correction is needed. However, if there is any curvature whatsoever, then the signed distance has an approximate normal distribution, with mean greater than 0 and variance less than 1.

Using (1.16) a $100(1-\alpha)\%$ central confidence interval for ρ can be constructed by using the set $\{\rho : z^{(\frac{\alpha}{2})} \leq W_\rho \leq z^{(1-\frac{\alpha}{2})}\}$. In a similar fashion a hypothesis test for $\rho = \rho_0$ can be constructed by using the interval $\{z^{(\frac{\alpha}{2})}, z^{(1-\frac{\alpha}{2})}\}$ as an acceptance region for the test statistic W_{ρ_0}. The benefits of having used the standardized distance will be that the intervals and acceptance regions will be closer to the nominal coverage probability and α-level respectively.

DiCiccio's and Efron's (1992) paper concerns the construction of confidence intervals in multiparameter exponential families. These intervals are accurate to *second order*, meaning that they have asymptotic error probability of $\alpha/2+O(n^{-1})$ in

each tail, compared to the standard confidence intervals that have error probability of $\alpha/2+O(n^{-\frac{1}{2}})$ in each tail. Denoting the mean value parameter of the exponential family by μ and the parameter of interest as $\psi = t(\mu)$, the correction to the standard intervals (1.7) is composed of three parts:

1. A correction for bias in the *m.l.e.*, $\hat{\psi}$ of ψ.

2. A correction for the non-linearity of the function $\psi = t(\mu)$.

3. A correction for the fact that the standard error of $\hat{\psi}$ is not a constant, but varies with μ.

DiCiccio and Efron give two types of confidence interval for ψ, the ABC confidence intervals and the ABC_Q interval. Both are second order accurate, but the one that is easier to calculate, the ABC_Q interval, fails to have the correct transformation behavior. Correct transformation behavior is meant in the sense that if the parameter of interest is subject to a $1-1$ transformation, say $\psi \rightarrow m(\psi)$ then the confidence limit for the transformed parameter should be obtainable by transforming the original limits using $m(\cdot)$.

A technical report (DiCiccio & Efron, 1991) presents an algorithm for calculating the ABC confidence limits for exponential family problems. It also presents an argument that shows all the correction methods described so far are second order accurate, so that a strong criteria for choosing between the methods is that of computation ease.

The correction method proposed by McCullagh and Tibshirani (1990) is closer in spirit to the methods proposed in this thesis, in the sense that the focus of attention is on the score equations for the model. As mentioned earlier, the critical reason for bias in the *m.l.e.* for the parameter of interest is the substitution

bias displayed by the score equation (see 1.3). McCullagh and Tibshirani suggest a correction to the profile score equation which accomplishes the dual goals of making the corrected profile score both *unbiased* and *information unbiased*. If the profile log-likelihood is denoted by $l_p(\psi) = l(\psi, \hat{\lambda}_\psi)$ then the profile score is defined as $U(\psi) = \frac{\partial}{\partial \psi} l_p(\psi)$, and the corrected profile score $\tilde{U}$ satisfies:

$$E_{\psi,\hat{\lambda}_\psi} \tilde{U}(\psi) = 0$$

so that it is unbiased, and

$$\mathrm{Var}_{\psi,\hat{\lambda}_\psi}[\tilde{U}] = -E_{\psi,\hat{\lambda}_\psi}\left[\frac{\partial}{\partial \psi}\tilde{U}\right]$$

which is the information unbiasedness criterion.

This can be achieved by what amounts to a location and scale transform of the profile score

$$\tilde{U}(\psi) = \{U(\psi) - m(\psi)\}w(\psi), \tag{1.17}$$

where formulae for the functions $m(\cdot)$ and $w(\cdot)$ can be found in McCullagh and Tibshirani (1990, p.328).

These unbiasedness properties are judged desirable because they are a characteristic of regular maximum likelihood score functions, and further motivation for them can be found in the theory of optimal estimating equations; see Godambe (1960) and Lindsay (1982).

One important fact to notice is that the unbiasedness properties are required to hold at $(\psi, \hat{\lambda}_\psi)$ and not (ψ, λ), whereas, in this thesis, expectations and derivatives will always be taken at the model unless otherwise stated.

The adjusted profile score $\tilde{U}$ is integrated to obtain a likelihood type object

$$l_{ap}(\psi) = \int^{\psi} \tilde{U}(t)dt,$$

referred to as the "adjusted profile likelihood". The desire to integrate the adjusted score causes some difficulties, similar in nature to those experienced by the Cox and Reid methodology, which is unable at present to deal with a multivariate parameter of interest. The problem is that in general $\partial \tilde{U}_r/\partial \psi^s \neq \partial \tilde{U}_s/\partial \psi^r$ so that there is no likelihood function having $\tilde{U}_r$ as a gradient vector. (Technically the adjusted score functions form a non-conservative vector field.)

For some simple problems the corrected profile score can be calculated explicitly, but in general, Monte Carlo simulations are required to calculate the expectations and variances. A first-order approximation to the term $m(\psi)$ is derived which will prove to be closely connected to one of the correction terms considered in this thesis. A first order approximation is given for the variance adjustment $w(\psi)$ only in the case of full exponential family models.

McCullagh and Tibshirani present an interesting comparison for canonically parameterized exponential family models, between the first order approximation to McCullagh and Tibshirani's adjusted profile score and the approximate conditional profile score of Cox and Reid. The latter is obtained by differentiating (1.14). (See Liang (1987) and Ferguson, Reid, and Cox (1991) for a detailed discussion of the Cox and Reid's profile score.) In the cases of Normal and Poisson models the modifications are identical, but this is not true in general.

The discussion here follows McLeish and Small (1988, Chapters 2 and 4), but see also Small and McLeish (1988, 1989) for additional developments.

The overall objective of their work is to provide a method of choosing an inference function that satisfies certain preconceived desirable properties. These properties may be information maximality, computational ease, linearity in the observations, or a host of other things. The authors extend the traditional statistical

constructs of ancillarity, sufficiency and completeness to what they describe as an "inference function" setting. Even though an important aspect of their work is its generality, it will be described here in a standard parametric framework, in which there is a family of probability measures $\mathcal{P}$ indexed by a parameter $\theta \in \Theta$ which is taken to be of finite dimension. This will facilitate comparison with some of the results presented in this thesis.

Attention is restricted to an "inference function space", Ψ, which is a Hilbert space of unbiased functions, all of which satisfy the condition of having "constant covariant structure". (See McLeish and Small (1988, p.13) for the definition of constant covariant structure.) A function $f \in \Psi$ is said to be unbiased if it is uncorrelated (orthogonal) to the 1 function, that is

$$< f(\theta), 1 >_\theta = 0 \quad \text{for all } \theta \in \Theta.$$

Here, $< \cdot, \cdot >_\theta$ denotes the inner product in the Hilbert space, and is defined for two functions f_1 and f_2 as

$$< f_1, f_2 >_\theta = E_\theta[f_1 \, f_2].$$

An ancillary statistic is by definition a function of the data, say $A(x)$ such that the distribution of A is free of θ. The expectation analog of this is *E-ancillarity* where an inference function is defined to be E-ancillary if

$$E_\theta[f(\theta_0)] = 0 \text{ for all } \theta \text{ and } \theta_0. \tag{1.18}$$

This definition is an extension of what Lehmann (1981) refers to as *first order ancillarity* and is a much weaker requirement than the usual definition of ancillarity. In fact Small and McLeish show that any function of an ancillary statistic must be an E-ancillary function.

Basu's classic theorem (1955, 1958) states that if a sufficient statistic is boundedly complete, then all ancillary statistics are independent of it. Writing the complete sufficient statistic as T and an ancillary as V, then the normalized statistics $T^* = T - E[T]$ and $V^* = V - E[V]$, which are mean zero, satisfy

$$E_\theta[V^* T^*] = 0 \text{ for all } V^* \text{ and } \theta. \tag{1.19}$$

This demonstrates a relationship between ancillarity and complete sufficiency in terms of expectations and orthogonality, in particular (1.19) can be rewritten as $< T^*, V^* >= 0$ for all V^* and θ. This motivates the definition of an *E-sufficient* inference function as one that is uncorrelated only with E-ancillary functions. That is, a function f is E-sufficient if $< f, g >_\theta = 0$ for all θ implies that g is E-ancillary. An inference function is said to be *complete E-sufficient* if in addition to being E-sufficient, it possesses the property that *all* E-ancillary statistics are orthogonal to it.

Let $\mathcal{A}$ denote the subspace of E-ancillary functions and $\mathcal{S}$ the subspace of complete E-sufficient functions, which will be closed by construction. Assuming that all the probability distributions are absolutely continuous with respect to each other, and letting μ denote a reference measure, then the orthogonality of $\mathcal{A}$ and $\mathcal{S}$ ensures a μ-almost sure decomposition of any inference function f into two parts:

$$f = f_A + f_S$$

where $f_A \in \mathcal{A}$ and $f_S \in \mathcal{S}$. The function f_S is the analog of having reduced the problem by sufficiency, and since it is orthogonal to all E-ancillary functions will hopefully provide an efficient distillation of the data. It is in theory obtainable by projecting the function f onto the subspace $\mathcal{S}$, but of course from a practical

perspective it is the feasibility of this projection that will determine the usefulness of the approach.

The ancillarity and sufficiency definitions can be redefined in a "local" way; specifically, an inference function is said to be *locally E-ancillary* if $E_\theta[f(\theta_0)] = o(\theta - \theta_0)$ as $\theta \to \theta_0$. That is, the coefficient of $\theta - \theta_0$ in the Taylor series expansion of $E_\theta[f(\theta_0)]$ about θ_0 is zero:

$$\frac{\partial}{\partial\theta} E_\theta[f(\theta_0)]\bigg|_{\theta=\theta_0} = 0.$$

Assuming that the derivative can be taken through the integral associated with the expectation, this can be rewritten as $E_{\theta_0}[f(\theta_0)\,U(\theta_0)] = 0$ where $U(\theta_0)$ is the score function. The interpretation is that locally E-ancillary functions are *orthogonal* to the score function. The concepts of *local E-sufficiency* and *complete local E-sufficiency* are defined analogously to E-sufficiency and complete E-sufficiency. The authors extend local ancillarity and sufficiency definitions to higher orders, where for example *second order E-ancillarity* is defined as $E_\theta[f(\theta_0)] = o((\theta - \theta_0)^2)$ as $\theta \to \theta_0$.

Chapter 4 of McLeish and Small (1988) gives details on the use of their methodology in nuisance parameter problems. In some nuisance parameter problems, by conditioning on an appropriate complete sufficient statistic, the nuisance parameter can be eliminated. (This will be referred to as a Type I problem, see Section 2 of Chapter 2.) The authors show that in such cases, the conditional score function is locally E-sufficient.

As noted earlier, in many nuisance parameter problems, such a reduction is not possible. Small and McLeish refer to this situation as one of "Obstructing Nuisance Parameters", and suggest the use of inference functions that depend on

both the parameter of interest and nuisance parameter, and satisfy the condition

$$E_{\psi,\lambda}[f(\psi,\lambda_0)] = 0 \quad \text{for all } \psi, \lambda \text{ and } \lambda_0.$$

Treating ψ as fixed, this condition demands that the inference function should be E-ancillary for the nuisance parameters. This is a strong condition and it is at times useful to relax it to that of *local unbiasedness* where for fixed ψ the inference function is required to be only locally E-ancillary, or equivalently orthogonal to the score functions for the nuisance parameters. *Second order unbiasedness* is defined by forcing the inference function to second order E-ancillary for the nuisance parameters treating ψ again as fixed. Obvious extensions to higher orders of unbiasedness follow. These degrees of unbiasedness will be considered in much greater detail in the thesis. In their notation they show that the function $S_1(\theta,\zeta) - c(\theta,\zeta)S_2(\theta,\zeta)$ where S_1 and S_2 are the score functions for the parameter of interest and nuisance parameter respectively, and c is an appropriate constant is locally unbiased. They also show that the function $S_1(\theta,\zeta) - c_1 S_2 - c_2(S_2^2 - I_2)$, where $I_2 = -\frac{\partial}{\partial\zeta}S_1$ and c_1, c_2 are appropriate constants is second order unbiased. These two functions will be exhaustively investigated in this thesis. The first function forms the basis of Neyman's $C(\alpha)$ test (Neyman, 1959), and what may be the earliest example of the second type can be found in Godambe and Thompson (1974), where it is presented as an optimal estimating equation in a very restricted class of problems.

The benefits of using an inference function that is locally unbiased or second order unbiased are that it should demonstrate a reduced sensitivity to the nuisance parameters. As an indication of this (Small & McLeish, 1989) includes a result (see section 4) that essentially states that for a second order unbiased inference function g, $E[g(\psi,\hat{\lambda}_\psi) - g(\psi,\lambda)] = o(1)$. If the inference function were only first

order unbiased the preceding expectation would in general be $O(1)$. The idea is that when λ is replaced by an estimate ($\sqrt{n}$ consistent) $\hat{\lambda}_\psi$ the change as measured by the difference in expectations is negligible, and it is in this sense that the sensitivity to the nuisance parameters has been reduced.

In order to make inference, the authors suggest drawing an inference function from the *N-sufficient subspace* where inference functions in the N-sufficient subspace have the property that each component is E-ancillary to the required order for the other parameters. For example, with two parameters in the model ψ and λ, the first order N-sufficient subspace is generated by the inference function $(U_\psi - c_{12}U_\lambda, U_\lambda - c_{21}U_\psi)$, here c_{ij} is a constant determined so that the functions satisfy the appropriate orthogonality constraints. The first component $U_\psi - c_{12}U_\lambda$ is E-ancillary with respect to λ since it is orthogonal to the λ-score U_λ, and the second component is E-ancillary with respect to ψ since it is orthogonal to the ψ-score U_ψ.

The dual themes of projection and orthogonality are a constant motif of the work of McLeish and Small. These themes remain at the core of this thesis.