Weighted Composite Likelihoods

Simon Harden
UCL

PhD Thesis

I, Simon Harden confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

For analysing complex multivariate data, the use of composite surrogates is a well established tool. Composite surrogates involve the creation of a surrogate likelihood that is the product of low dimensional margins of a complex model, and result in parameter estimators with acceptable properties (such as lack of bias and efficiency) that are relatively inexpensive to calculate. Some work has taken place in adjusting these composite surrogates to restore desirable features of the data generating mechanism, but the adjustments are not specific to the composite world: they could be applied to any surrogate. An issue that has received less attention is the determination of weights to be attached to each marginal component of a composite surrogate. This issue is the main focus of this thesis. We propose a weighting scheme derived analytically from minimising the Kullback-Leibler Divergence (KLD) between the data generating mechanism and the composite surrogate, treating the latter as a bona fide density which requires consideration of a normalising constant (a feature which is usually ignored). We demonstrate the effect of these weights for a simulation. We also derive an explicit formulation for the weights when the composite components are multivariate normal and, in certain cases, show how they can be used to restore the original data generating mechanism.

Contents

List of Figures

List of Tables

Heartfelt thanks to Dr Richard Chandler for his ideas and forbearance, and for not letting me stray too far from the statistical straight and narrow. Thanks also to Professor Trevor Sweeting for guidance at key moments and to the ESPRC for their financial support (EP/P504767/1).

Notation

All vectors are taken to be column vectors as in, for instance, Barndorff-Nielsen and Cox (1994). A function of p parameters (such as a loglikelihood) is differentiated (downwards) with respect to the parameters, resulting in a $p \times 1$ vector. It, in turn, is differentiated (sideways) with respect to the parameters to produce a $p \times p$ matrix. On occasion, we will differentiate that matrix with respect to the parameters. If $p > 1$, this will result in an array of p matrices. We describe this array as a p vector of $p \times p$ matrices. Matrix multiplication can then become a little unintuitive. For instance, if $p > 1$, $\boldsymbol{W}$ is a $p \times p$ matrix, and $\boldsymbol{\theta}$ and $\boldsymbol{V}$ are $p \times 1$ vectors, then:

$$\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\theta}} \boldsymbol{V} = \left(\frac{\partial \boldsymbol{W}}{\partial \theta_1} \boldsymbol{V}, \ldots, \frac{\partial \boldsymbol{W}}{\partial \theta_p} \boldsymbol{V} \right)$$

and

$$\boldsymbol{V}^T \frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\theta}} = \left(\boldsymbol{V}^T \frac{\partial \boldsymbol{W}}{\partial \theta_1}, \ldots, \boldsymbol{V}^T \frac{\partial \boldsymbol{W}}{\partial \theta_p} \right)^T.$$

are $p \times p$ matrices. We have omitted the tedious algebraic detail but more information can be found in, for instance Wei (1997).

Vectors and matrices are only emboldened if it is certain that they are not scalars.

Any subscript added to an expectation (or variance etc) refers to the distribution over which the expectation is to be taken. For instance,

$$\mathsf{E}_{\mathsf{G}}[Y]$$

refers to the expected value of random variable Y under the distribution **G**.

The $p \times p$ identity matrix is described as $I_{p\times p}$. $I_{\mathbf{G}}$ and $\hat{I}_n$ are the derivatives of the expected and observed estimating functions respectively.

We define a *dataset* to be a set of observations, $y_1, \ldots, y_n$, each *element* of which could be a cluster of *datapoints*, $y_i = (y_{i1}, \ldots, y_{im_i})$. The suffix i is used to represent clusters and j, points within clusters. The first covariate for y_{ij} is $x_{ij\,1}$ etc.

Certain letters are used to represent the same feature throughout this report:

- n - number of data elements
- m_i - length of cluster i
- m - length of clusters if all the same
- θ - parameters - vector or scalar
- p - number of parameters
- l - number of parameters for a test hypothesis (if not p)
- q - number of *components* in a composite surrogate
- **G** - distribution that generated the data, generally unknown
- **H** - preferred surrogate distribution - normally not used
- **F** - surrogate distribution. In the case of a composite surrogate, the term is used loosely so that
- $\mathbf{F}_K$ - surrogate distribution with constant of proportionality and thus well defined density
- K - constant of proportionality or normalising constant
- ψ - estimating function
- n as suffix - quantity calculated from n data elements

- cs as suffix - relating to composite surrogate
- $\boldsymbol{J}$ - second derivative of Kullback-Leibler Divergence with respect to the weights in new scheme.

Outline

For analysing complex multivariate data, the use of composite surrogates is a well established tool. Composite surrogates involve the creation of a surrogate likelihood that is the product of low dimensional margins of a complex model, and result in acceptable parameter estimators that are relatively inexpensive to calculate. Some work has taken place in adjusting these composite surrogates to restore desirable features of the data generating mechanism, but the adjustments are not specific to the composite world: they could be applied to any surrogate. An issue that has received less attention is the determination of weights to be attached to each marginal component of a composite surrogate. This issue is the main focus of this thesis. We propose a weighting scheme derived analytically from minimising the Kullback-Leibler Divergence (KLD) between the data generating mechanism and the composite surrogate, treating the latter as a bona fide density which requires consideration of a normalising constant (a feature which is usually ignored). We demonstrate the effect of these weights for a simulation. We also derive an explicit formulation for the weights when the composite components are multivariate normal and, in certain cases, show how they can be used to restore the original data generating mechanism.

In Chapter 1, we review results for surrogates, which are likely not to have generated the data. We use the KLD as a basis for parameter estimation, work with estimating functions and equations wherever possible, and are particular in establishing the assumptions made at every stage. We derive standard asymptotic results for consistency and distribution of parameter estimates, including the sandwich formulation for the variance, and provide estimators for the parameters and elements of the sandwich. We show how the usual test statistics are distributed in the case of surrogates. We review the use of a number of potential adjustments to surrogates that have been proposed in the literature. Finally, we examine how we might compare different surrogates for multivariate parameters. A simple univariate single parameter example illustrates the concepts and two continuing practical examples are described.

In Chapter 2, we examine the use of composite surrogates. Wherever possible we work with an unknown data generating mechanism and do not assume that the composite surrogate components are marginal for it. We look at bias, covariance estimation and, as a new contribution, the effect of introducing a normalising constant so that the surrogate can be regarded locally as a bona fide likelihood for a misspecified model. We illustrate this with reference to a composite normal surrogate. We explore the performance of the surrogate, and the adjustments, using a simulation based around clustered binary outcomes in a logistic regression with cluster-specific random effect. The effect of using higher order features for small samples is reviewed.

In Chapter 3, we examine the issue of weighting the components of a composite surrogate. We review the published work in this area, linking and extending it as required, resulting in two optimally efficient schemes for estimating functions, one a simplification of the other. We introduce a completely new weighting scheme, which takes into account the normalising constant and consists of a set of equations to be solved for the weights, proving its derivation from the KLD and explore how it might be used in practice. The performance of this new scheme is assessed in a simulation study based around probit regression and an autoregressive random effect, but the effect of the weights on the results of the simulation is not significant.

In Chapter 4, we apply the new weighting scheme analytically to composite surrogates whose components are multivariate normal and show that these surrogates represent distributions of transformations of the data. We derive elegant forms for the weights equations. We examine the circumstances under which the use of weights enables us to recover the distribution that generated the data. We apply these to the simulation from Chapter 3, again with no significant effect, and to autoregressive models.

In Chapter 5 we review the thesis. We suggest a reason for the results of the simulations and discuss the value of using weights at all. We suggest areas for further research.

Chapter 1

Surrogates

1.1 Introduction

Most data studied statistically arise from a mechanism that is at least partially unknown. They are often analysed using a parametric model that is a surrogate for that mechanism. In Section 1.2 we set up our terminology for the study of such surrogates, based around the use of estimating functions, and introduce some recurring examples. Basic results arise from minimising the Kullback-Leibler discrepancy between the data generating mechanism and the surrogate. In Section 1.3 we derive the standard asymptotic results and in Section 1.4 the observed equivalent to various theoretical expressions. The distributions of a number of test statistics are studied in Section 1.5. A range of adjustments to loglikelihoods and estimating functions which simplify the asymptotic distribution of test statistics is analysed in Section 1.6. They are shown to share common features. The choice of which surrogate to use for any particular dataset is important and methods for comparing these choices are outlined in Section 1.7. Finally, for completeness, two Bayesian approaches are described in Section 1.8.

1.2 Basics

We consider the analysis of data, $y_1, \ldots, y_n$, which are realisations of random variables, $Y_1, \ldots, Y_n$, observed from an unknown distribution $\mathbf{G}$, not necessarily belonging to a parametric family. Each data item could be a vector of, possibly dependent, measurements. We are interested in features of $\mathbf{G}$ - mean, variance etc - that we shall term *objects of interest*. As $\mathbf{G}$ is unknown, we investigate *objects of inference*, $\theta = \theta(\mathbf{G}) \in \Theta$, with dimension p, arising in $\mathbf{F}$, a *surrogate* for $\mathbf{G}$, using the data at hand. Although much of our analysis will be carried out using estimating functions, our primary area of interest is Maximum Likelihood Estimation, in which case we will use a likelihood, $\mathsf{L}_{\mathbf{F}}$, and loglikelihood, $\ell_{\mathbf{F}}$, based on $\mathbf{F}$. The related joint density, f, may not be fully described, particularly if the constant of proportionality (or normalising constant) is difficult to derive.

The two sets of objects are related in that the value of estimators for the objects of inference will depend upon the data and thus $\mathbf{G}$. One could indeed establish a functional from the set of possible $\mathbf{G}$s to the objects of inference so that $\theta_{\mathbf{G}}$ would be the value of θ that brings $\mathbf{F}$ 'closest' to $\mathbf{G}$.

The focus of this thesis is the study of *surrogate likelihoods* which may arise from *surrogate distributions* and *surrogate models*. One might consider a range of surrogates for a particular dataset using a variety of criteria to distinguish between them, such as mathematical tractability, computability, optimality (in a sense to be defined), information criteria or robustness.

Example I - Poisson Surrogate for Gaussian Data. Suppose that we have independent data, $y_1, \ldots, y_n$ from $\mathbf{G} \equiv N(\mu, \sigma^2)$. We have been explicit with the density, g derived from $\mathbf{G}$, for clarity's sake in this example, but it is generally unknown. Suppose that in the absence of knowledge of $\mathbf{G}$ our surrogate is $\mathbf{F} \equiv \mathrm{Poi}(\theta)$. Clearly, this will only work with non negative integer data and for this example, we will assume that is the case (for example, non negative data may have been rounded to integer values). Our object of inference is θ but our objects of interest will be whichever features of $\mathbf{G}$ we are interested in, say the mean and variance. We will need to ensure that θ has the desired interpretation in the context of $\mathbf{G}$. □

Define an *estimating function* (or *inference function*, McLeish and Small, 1980) as $\psi(\theta; y)$, a k-vector valued function ($k \geq p$) relating θ and y such that:

$$\mathsf{E}_{\mathbf{G}}[\psi(\theta_{\mathbf{G}}; Y)] = 0 \quad \text{some } \theta_{\mathbf{G}} \in \Theta. \tag{1.1}$$

the principal purpose of which is to define $\theta_{\mathbf{G}}$. Uniqueness is defined in Assumption 5. The subscript for θ denotes the fact that we have chosen an estimating function which we believe will have something of interest to say about **G** at parameter values $\theta_{\mathbf{G}}$, expectations being taken over the unknown **G**. Let $\mathcal{E}$ denote the class of all such estimating functions.

We are likely to have multiple instances of Y, $y_1, \ldots, y_n$. Each of these data elements will be generated from **G** and may represent a cluster of datapoints, $y_i = y_{i1}, \ldots, y_{im_i}$. Note that the data items may be of varying length. As shorthand we shall use m rather than the set of m_i and this may represent $\max(m_i), 1 \leq i \leq n$. We make no assumption about the independence of the Y_i. In practice the dependence that typically arises relates to time or space, and so one might assume stationarity at some level in those continuums (see Section 2.6).

We will make a number of assumptions in this thesis, some of which are specific to particular sections. They are mostly to keep us in the realm of well behaved functions that one would encounter in practice. For an analysis of many of the estimating function assumptions, especially as they relate to asymptotic results in Section 1.3, see Jesus and Chandler (2011). Where there are common exceptions to the assumptions, they will be noted. At various points we develop expressions based upon taking expectations over the distribution **G**. Where explicitly described, $g(y)$, the corresponding density, is assumed to be a continuous function in order to maintain simplicity with respect to integration.

Assumption 1. *$\psi(\theta; y)$ is continuous in θ for any y and measurable in y for any θ.*

An alternative definition for an estimating function (eg Song, 2007) just takes any function that satisfies Assumption 1 and describes it is as *unbiased* if $\mathsf{E}_\theta[\psi(\theta;Y)] = 0$ for all $\theta \in \Theta$. In that approach, as θ varies, so will **G** (which is taken to equal **F**), the distribution over which expectations are being taken (resulting in the different use of subscripts from this thesis). However, that assumes that a density g exists and is parameterised by θ, an assumption that, in general, we are not making.

Assumption 2. *Differentiation with respect to θ and integration with respect to y are interchangeable.*

This does not have to be the case, for instance where the range of integration depends upon θ, as in the example where we have a distribution over $[\theta - 0.5, \theta + 0.5]$.

In the familiar non surrogate situation (ie where **F** is **G**), $g(y;\theta)$ is a member of a parametric family with score $U(\theta) = \partial \ln g(y;\theta)/\partial\theta$ then:

$$\begin{aligned}
\mathsf{E}_{\mathbf{G}}[U(\theta)] &= \int \frac{\partial \ln(g(y;\theta))}{\partial\theta} g(y;\theta)\,\mathrm{d}y \\
&= \int \frac{\partial g(y;\theta)}{\partial\theta}\frac{1}{g(y;\theta)} g(y;\theta)\,\mathrm{d}y \\
&= \int \frac{\partial g(y;\theta)}{\partial\theta}\,\mathrm{d}y \\
&= \frac{\partial}{\partial\theta}\int g(y;\theta)\,\mathrm{d}y \quad \text{by Assumption 2} \\
&= \frac{\partial 1}{\partial\theta} \\
&= 0. \qquad (1.2)
\end{aligned}$$

so the score is an estimating function for all $\theta \in \Theta$ and, in particular, in an area around the value of of θ in which we are interested. (1.2) is sometimes known as Bartlett's first identity.

However, in general, where g, if it exists, is not necessarily a member of a known parametric family and we wish to use f (or its likelihood) as a parametric surrogate, we commence the same proof:

$$\begin{aligned} \mathsf{E}_{\mathbf{G}}\left[\frac{\partial \ln(f(\theta;Y))}{\partial\theta}\right] &= \int \frac{\partial \ln(f(\theta;y))}{\partial\theta} g(y)\,\mathrm{d}y \\ &= \int \frac{\partial f(\theta;y)}{\partial\theta}\frac{1}{f(\theta;y)} g(y)\,\mathrm{d}y \end{aligned} \tag{1.3}$$

and can go no further.

The approach that we shall then use (which we shall term *surrogate maximum likelihood estimation* or *SMLE*) is to find the value of θ, θ_*, that maximises the expected loglikelihood arising from **F** under **G**, ie knowing that we are working with a surrogate, we are looking for the parameter value that takes us as 'close' as possible to **G**. That involves trying to maximise:

$$\mathsf{E}_{\mathbf{G}}[\ell_{\mathbf{F}}(\theta)] = \int \ell_{\mathbf{F}}(\theta) g(y)\,\mathrm{d}y \tag{1.4}$$

sometimes known as the *Fraser Information* (Kent, 1982), over the parameters or, equivalently, to minimise:

$$\mathsf{E}_{\mathbf{G}}\left[\frac{\ln(g(Y))}{\ell_{\mathbf{F}}(\theta)}\right] = \mathsf{E}_{\mathbf{G}}[\ln(g(Y))] - \mathsf{E}_{\mathbf{G}}[\ell_{\mathbf{F}}(\theta)], \tag{1.5}$$

known as the *Kullback-Leibler Divergence* (Cox, 2006) or *KLD* (also used is Kullback-Leibler discrepancy in Davison 2003). The equivalence arises as the first term in (1.5) is constant with respect to the parameters and is sometimes omitted (White, 1982). There are other discrepancy functions that we could use but the Kullback-Leibler discrepancy seems the most natural, particularly as we are working with likelihood functions. Linhart and Zucchini (1986) describes many of these functions (such as those defined by Kolmogorov, and Cramér and von Mises) and provides an overview of this approach to inference.

To maximise (1.4), we differentiate with respect to the parameters and set equal to zero:

$$\begin{aligned}
0 &= \left.\frac{\partial \int \ell_{\mathsf{F}}(\theta) g(y)\,\mathrm{d}y}{\partial\theta}\right|_{\theta_*} \\
&= \int \left.\frac{\partial \ell_{\mathsf{F}}(\theta) g(y)}{\partial\theta}\right|_{\theta_*} \mathrm{d}y \quad \text{by Assumption 2} \\
&= \int \left.\frac{\partial \ell_{\mathsf{F}}(\theta)}{\partial\theta}\right|_{\theta_*} g(y)\,\mathrm{d}y \quad \text{as } g(y) \text{ is not a function of } \theta \\
&= \mathsf{E}_{\mathsf{G}}[\psi(\theta_*; Y)] \quad \text{say,}
\end{aligned}$$

and we have an estimating function, as defined in (1.1), with $\theta_{\mathsf{G}} = \theta_*$ so that under SMLE, the score of a surrogate is an estimating function, when expectations are taken over **G**.

Henceforth, we will assume that we are always working with surrogates and use definition (1.1). In maximum likelihood estimation, the estimating functions are the score functions. Nomenclature for the objects of interest and inference is from Royall and Tsou (2003), where the following example in which the two objects do not coincide is given. Let E$[Y]$ be our object of interest for **G**, which is not lognormal, and consider a surrogate, $\mathsf{F} \equiv \text{Lognormal}(\mu, \sigma^2)$, with density f. If we define μ as our object of inference in the surrogate, then $\mu_{\mathsf{G}} = \exp\left(\mathsf{E}_{\mathsf{G}}[\ln(Y)] + \sigma^2/2\right)$ which is not the same as the object of interest. As, in general, we do not know **G**, it should not therefore be assumed that objects of inference and interest always coincide.

We will formalise the uniqueness of θ_{G} in Assumption 5. This approach, whereby we work with a single target θ_{G}, is consistent with that used in Generalised Method of Moments (GMM) which arose originally through a least squares minimisation of orthogonal moment conditions (see, for instance, Hansen, 1982) and is discussed in Jesus and Chandler (2011).

Example I continued - Poisson Surrogate for Gaussian Data. The estimating function for the surrogate is:

$$\begin{aligned}\psi(\theta;y) &= \frac{\partial \ln f(y;\theta)}{\partial\theta} \\ &= \frac{\partial \ln(\theta^y \exp(-\theta)/y!)}{\partial\theta} \\ &= \frac{y}{\theta} - 1.\end{aligned}$$

The zero of the expected value, $\theta_{\mathbf{G}}$ satisfies:

$$\begin{aligned}0 &= \mathrm{E}_{\mathbf{G}}[\psi(\theta_{\mathbf{G}};Y)] \\ &= \mathrm{E}_{\mathbf{G}}\left[\frac{Y}{\theta_{\mathbf{G}}} - 1\right] \\ &= \frac{\mu}{\theta_{\mathbf{G}}} - 1.\end{aligned}$$

so that our object of inference $\theta_{\mathbf{G}} = \mu$ and we would use that value for estimating our first object of interest, the mean of $\mathbf{G}$. Our second object of inference, the variance, would then also have to be given value $\theta_{\mathbf{G}}$ due to the nature of the Poisson distribution. Clearly this will not, in general, equal σ^2, the object of interest. □

Estimating (or *inference*) *equations* for a dataset $Y_1, \ldots, Y_n$ are:

$$\psi_n(\theta;y) = \psi_n(\theta;y_1,\ldots,y_n) = 0. \tag{1.6}$$

They are described as *M-estimators* or *Z-estimators* ('Z' for zero) in van der Vaart (1998).

Under maximum likelihood estimation for instance, by setting the score functions equal to zero we have estimating equations which we can solve for the maximum likelihood estimator. In many cases, such as iid Y_i under maximum likelihood estimation, we have:

$$\psi_n(\theta;y) = \sum_{i=1}^{n} \psi(\theta;y_i) \tag{1.7}$$

but in general it will not be possible to express $\psi_n(\theta;y)$ in terms of the basic $\psi(\theta;y_i)$ function.

Example II - Weather Readings. Suppose we have weather readings (eg wind speed, temperature) at m sites. Over a fixed period every site records a reading hourly to give n clusters, each with m elements. In general, the clustered readings will be neither independent (storms typically last longer than an hour so there will be short term dependence for wind speed) nor identically distributed (seasonality results in higher temperatures in summer). Whatever form a surrogate might take, one could condition upon recent readings to eliminate the time dependence (see Section 2.6) and build in seasonality so that an iid approach for clusters might be justifiable (see, for instance Yan et al., 2002). There is no inherent order to the elements within a cluster. □

Example III - Longitudinal Study. Measurements (eg blood pressure, weight) are taken from n patients over a period of a year after an initial reading on day one. Readings are taken whenever a patient visits their GP, which will be irregularly and may be never: m_i will be greater than zero but not the same for all patients and we define the set of readings for a patient as a cluster. While it is reasonable to assume independence between clusters, they will certainly not be identically distributed and (1.7) will not apply. Within each cluster, there is an obvious ordering based upon time. □

As a generalisation of (1.6), we shall normalise all ψ_n to give $\bar{\psi}_n$, ie:

$$\bar{\psi}_n(\theta; y) \equiv A_n \psi_n(\theta; y) \tag{1.8}$$

where A_n are $k \times k$ symmetric invertible fixed matrices, possibly dependent upon n. In Section 1.3 we will choose A_n so that $\bar{\psi}_n$ converges to a deterministic function with a root at θ_{G} as $n \to \infty$. See Assumption 5 below for a description of asymptotic behaviour of A_n. The symmetry of A_n is required at various points in this chapter. While for estimating functions which are the derivatives of objective functions, such as loglikelihoods or moment conditions, A_n will be diagonal, we have allowed for the more general case but do require the assumption of symmetry. For the iid case in (1.7), we take $A_n = \frac{1}{n} I_{k \times k}$ and if θ is scalar:

$$\begin{aligned} \bar{\psi}_n(\theta; y) &= \frac{1}{n} \psi_n(\theta; y) \\ &= \frac{1}{n} \sum_{i=1}^{n} \psi(\theta; y_i). \end{aligned} \tag{1.9}$$

We will solve our estimating equations, (1.6), to give $\hat{\theta}_n$, an estimator of $\theta_{\mathbf{G}}$.

Assumption 3. *$\hat{\theta}_n$ exists and is unique for all n.*

This would normally require that we have the dimension of ψ (k) equal to the number of unknowns (p). This is generally the case with the GMM and maximum likelihood estimation approaches, where differentiation by each of the elements of θ results in that number of estimating equations. However, one could use the estimating equations directly to set up any number of moment conditions (eg Davidson and McKinnon, 2004, page 371) and then some manipulation, perhaps using linear combinations of conditions, is required to arrive at a set of p independent conditions. We will assume any such manipulation has taken place so that $k = p$.

Example I continued - Poisson Surrogate for Gaussian Data. The estimating equation for the n data items is:

$$\begin{aligned} 0 &= \psi_n(\theta; y_1, \ldots, y_n) \\ &= \frac{\sum_{i=1}^{n} y_i}{\theta} - n \end{aligned} \tag{1.10}$$

solving to give:

$$\hat{\theta}_n = \frac{\sum_{i=1}^{n} y_i}{n}.$$

□

It is worth noting the connection with *indirect inference* (see, for instance, Gourieroux et al., 1993). In that case, the density of **G**, $g(y; \zeta)$, is known and amenable to simulation but otherwise intractable, and parameterised by ζ, distinct from θ which parameterises the surrogate **F**. Observations are used to derive $\hat{\theta}$, the standard maximum likelihood estimator under **F**. For a range of values of ζ, one simulates data from **G** and estimates θ under **F**. The estimate of ζ for which the resulting estimate of θ is closest to $\hat{\theta}$ is then taken as optimal and is shown to have good asymptotic properties and enables hypothesis testing to be carried out. While the set up of **G** is different from that described in this thesis, this approach has similarities to those used for assessing robustness, described in Section 1.7.4.

Henceforth, we shall use E[.] rather than $\mathrm{E}_{\mathbf{G}}[.]$ for simplicity's sake but it is important to remember that we are taking expectations over the (unknown) distribution **G**, unless otherwise mentioned.

1.3 Asymptotic Results

In this section we deal with asymptotic results. There are times when we deal with small samples and the results discussed here do not apply. In particular, the asymptotic normal distribution, discussed at (1.18), is not appropriate and then the test statistics we derive in Section 1.5 are not available. These issues are exacerbated when we deal with surrogates, as poor model specification for small samples has a more deleterious effect than for large. However, standard techniques for dealing with such cases, such as t-tests, are available. Many of the applications that use composite likelihoods, such as Example II, have a wealth of data and we can comfortably make use of the asymptotic results in this section.

We want to understand the behaviour as $n \to \infty$ of $\hat{\theta}_n$, the solution to $\bar{\psi}_n(\theta) = 0$ (omitting the Ys and ys for brevity wherever possible). We will first show that the $\hat{\theta}_n$ are consistent for $\theta_{\mathbf{G}}$, the maximiser under SMLE of (1.4), and then examine their asymptotic distribution. The structure of the proofs in this section can be found in, for instance, van der Vaart (1998). We make use of a vector metric or distance function, $||.||_v$. In order to ensure that $\theta_{\mathbf{G}}$ is well defined, we assume

Assumption 4. Θ *is compact.*

As a result, the limit of any sequence of parameters in the space is itself in Θ. Jesus and Chandler (2011) note that there may still be problems if $\theta_{\mathbf{G}}$ lies on the boundary of Θ and a function of θ that we wish to differentiate (such as an estimating function) is undefined outside that space. There are alternatives to compactness, some of which are described in van der Vaart (1998, Chapter 5).

Assumption 5. *There exists a sequence, $\{A_n\}$, of symmetric invertible fixed matrices converging in probability to some constant symmetric invertible matrix such that $\bar{\psi}_n(\theta) \equiv A_n\psi_n(\theta)$ converges uniformly in probability to ψ_∞, a twice differentiable fixed function of θ with a unique zero at $\theta_{\boldsymbol{G}}$ under the metric, ie:*

$$\sup_{\theta}\left(\|\psi_\infty(\theta) - \bar{\psi}_n(\theta)\|_v\right) \xrightarrow{p} 0$$

where:

$$\inf_{\theta:\|\theta-\theta_{\boldsymbol{G}}\|_v \geq \epsilon}(\|\psi_\infty(\theta)\|_v) > \|\psi_\infty(\theta_{\boldsymbol{G}})\|_v = 0 \tag{1.11}$$

for any $\epsilon > 0$ and $\theta \in \Theta$.

There are alternatives to this (eg van der Vaart, 1998, Section 5.2) but, as in the approach described here, they all require more than simple point convergence (although with a scalar θ, Assumptions 1 and 3 are adequate). An early proof of the consistency result (Huber, 1967), albeit not defined in terms of estimating functions, offers two approaches each based around a different set of highly technical assumptions. Clearly, in cases such as (1.7), $\mathsf{E}[\psi(\theta)]$ satisfies the convergence conditions for ψ_∞, and $\psi_\infty = \mathsf{E}[\psi(\theta)]$ but we need to assume a unique zero. A consequence of the assumption is that $\lim_{n\to\infty} \mathsf{E}[\bar{\psi}_n(\theta)] = \psi_\infty(\theta)$.

We apply Assumption 5 to $\bar{\psi}_n$ at $\hat{\theta}_n$. As n varies, so will $\hat{\theta}_n$, but it will remain in Θ and as Assumption 5 applies to the supremum over Θ we have

$$\|\psi_\infty(\hat{\theta}_n) - \bar{\psi}_n(\hat{\theta}_n)\|_v \xrightarrow{p} 0$$

so by the definition of $\hat{\theta}_n$ as a zero of $\bar{\psi}_n$:

$$\|\psi_\infty(\hat{\theta}_n)\|_v \xrightarrow{p} 0. \tag{1.12}$$

However, if for some $\epsilon > 0$ and for all $n > n(\epsilon)$ for some fixed $n(\epsilon)$:

$$\|\hat{\theta}_n - \theta_{\mathbf{G}}\|_v \geq \epsilon$$

then by (1.11):

$$\|\psi_\infty(\hat{\theta}_n)\|_v > \|\psi_\infty(\theta_{\mathbf{G}})\|_v \equiv 0.$$

But, the probability of the left hand side being larger than zero tends to 0 by (1.12) and so for all ϵ:

$$\mathsf{P}(\|\hat{\theta}_n - \theta_{\mathbf{G}}\|_v \geq \epsilon) \to 0.$$

Thus, $\hat{\theta}_n$ are consistent for $\theta_{\mathbf{G}}$.

Example I continued - Poisson Surrogate for Gaussian Data. The results of solving the estimating equations, $\hat{\theta}_n = \frac{\sum_{i=1}^n y_i}{n}$, are consistent for μ by the Central Limit Theorem. □

We now turn to the asymptotic distribution of the estimators.

Assumption 6. *$\bar{\psi}_n'(\theta)$ and $\bar{\psi}_n''(\theta)$ exist and are continuous functions of θ in an area around $\theta_{\boldsymbol{G}}$.*

$\bar{\psi}_n''(\theta)$ is a p-vector of $p \times p$ matrices; see the Notation note at the beginning for this thesis for more details.

The existence of two derivatives is a strong assumption which fails, for instance, when $\psi(\theta; y) = \mathsf{sgn}(y - \theta)$, which has a zero at the median. There are a variety of alternative assumptions (eg van der Vaart, 1998, Section 5.3) to deal with such situations.

Assumption 7. *$\bar{\psi}'_n(\theta_{\boldsymbol{G}}) \xrightarrow{p} \psi'_\infty(\theta_{\boldsymbol{G}})$, which we shall define as $-I_{\boldsymbol{G}}(\theta_{\boldsymbol{G}})$, a generalisation of the expected Fisher information, not dependent upon the data. $I_{\boldsymbol{G}}(\theta_{\boldsymbol{G}})$ commutes with A_n and $A_n^{-\frac{1}{2}}$.*

We have implicitly used the same normalising matrix for $\psi'_n(\theta_{\mathsf{G}})$, A_n, as we did for $\psi_n(\theta)$ at (1.8) and in Assumption 5. This is not unreasonable as $\psi'_n(\theta)$ is well behaved around θ_{G} by Assumption 6. However, Jesus and Chandler (2011) do use a separate normaliser. In the case of iid Y_i the weak Law of Large Numbers (*LOLN*) means that the convergence in Assumption 7 will take place in distribution.

Assumption 8. *$I_{\boldsymbol{G}}(\theta_{\boldsymbol{G}})$ is nonsingular.*

An extension of this to its estimator is given as Assumption 14.

By the Lagrange form of Taylor's theorem:

$$\bar{\psi}_n(\hat{\theta}_n) = \bar{\psi}_n(\theta_{\mathsf{G}}) + \bar{\psi}'_n(\theta_{\mathsf{G}})(\hat{\theta}_n - \theta_{\mathsf{G}}) + \frac{1}{2}((\hat{\theta}_n - \theta_{\mathsf{G}})^T \bar{\psi}''_n(\breve{\theta}_n))(\hat{\theta}_n - \theta_{\mathsf{G}}) \tag{1.13}$$

for some $\breve{\theta}_n$ 'between' $\hat{\theta}_n$ and θ_{G} (ie $\|\breve{\theta}_n - \hat{\theta}_n\|_v \leq \|\theta_{\mathsf{G}} - \hat{\theta}_n\|_v$ and $\|\breve{\theta}_n - \theta_{\mathsf{G}}\|_v \leq \|\theta_{\mathsf{G}} - \hat{\theta}_n\|_v$). Then, because $\bar{\psi}_n(\hat{\theta}_n) = 0$ by definition:

$$\begin{aligned}(\hat{\theta}_n - \theta_{\mathsf{G}}) &= -\left(\bar{\psi}'_n(\theta_{\mathsf{G}}) + \frac{1}{2}(\hat{\theta}_n - \theta_{\mathsf{G}})^T \bar{\psi}''_n(\breve{\theta}_n)\right)^{-1} \bar{\psi}_n(\theta_{\mathsf{G}}) \\ &= \left(I_{\mathsf{G}}(\theta_{\mathsf{G}}) + o_p(1) - \frac{1}{2}(\hat{\theta}_n - \theta_{\mathsf{G}})^T \bar{\psi}''_n(\breve{\theta}_n)\right)^{-1} \bar{\psi}_n(\theta_{\mathsf{G}})\end{aligned}$$

where the second line follows by Assumption 7.

Assumption 9. *The second derivative of $\bar{\psi}_n$ is finite in an area around $\theta_{\boldsymbol{G}}$.*

Specifically, since $\hat{\theta}_n$ and thus $\breve{\theta}_n$ are consistent for θ_{G} (ie they are, in probability, in the area described in the assumption), $\|\bar{\psi}''_n(\breve{\theta}_n)\|$ is $O_p(1)$

Thus, we have:

$$\begin{aligned}(\hat{\theta}_n - \theta_{\mathsf{G}}) &= \left(I_{\mathsf{G}}(\theta_{\mathsf{G}}) + o_p(1) - \frac{1}{2}(\hat{\theta}_n - \theta_{\mathsf{G}})^T O_p(1)\right)^{-1} \bar{\psi}_n(\theta_{\mathsf{G}}) \\ &= \left(I_{\mathsf{G}}(\theta_{\mathsf{G}}) + o_p(1) + \frac{1}{2} o_p(1) O_p(1))\right)^{-1} \bar{\psi}_n(\theta_{\mathsf{G}}) \quad \text{by consistency} \\ &= (I_{\mathsf{G}}(\theta_{\mathsf{G}}) + o_p(1) + o_p(1))^{-1} \bar{\psi}_n(\theta_{\mathsf{G}}) \\ &= (I_{\mathsf{G}}(\theta_{\mathsf{G}}) + o_p(1))^{-1} \bar{\psi}_n(\theta_{\mathsf{G}}) \\ &= I_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}}) \bar{\psi}_n(\theta_{\mathsf{G}}) + O_p(1) o_p(1) \\ &= I_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}}) \bar{\psi}_n(\theta_{\mathsf{G}}) + o_p(1) \qquad (1.14)\end{aligned}$$

where the $o_p(1)$ term in the penultimate line comes from $\bar{\psi}_n(\theta_{\mathsf{G}})$ tending to zero in probability by Assumption 5. Therefore, $(\hat{\theta}_n - \theta_{\mathsf{G}})$ has asymptotic expectation:

$$\begin{aligned}I_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}}) \mathsf{E}[\bar{\psi}_n(\theta_{\mathsf{G}})] &\rightarrow I_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}}) \psi_\infty(\theta_{\mathsf{G}}) \quad \text{by Assumption 5} \\ &= 0. \qquad (1.15)\end{aligned}$$

Here and in later proofs, we have simplified the notation by not being explicit about the dimensions of the $o_p(1)$ and $O_P(1)$ expressions - this does not affect the results.

Assumption 10. *A further condition on ψ_n is:*

$$B_n Var[\psi_n(\theta)] B_n \xrightarrow{p} C_{\boldsymbol{G}}(\theta).$$

for some fixed $n \times n$ matrices B_n and $C_{\boldsymbol{G}}(\theta)$, the former symmetric and the latter not dependent upon the data. B_n commutes with $I_{\boldsymbol{G}}^{-1}(\theta_{\boldsymbol{G}})$

As with A_n, as discussed after (1.8), the rather awkward commutative requirements of our normalising matrix, B_n, will not be required in the usual circumstances when B_n is diagonal. With iid Y_i, $C_{\mathsf{G}}(\theta) = \mathsf{E}[\psi(\theta)\psi(\theta)^T]/n$ (ie $B_n = A_n^{\frac{1}{2}}$) and Assumption 10 states that the variance of the sample score tends to its estimating function equivalent in probability. A consequence of the assumption is that C_{G} is symmetric.

Assumption 11. *$\bar{\psi}_n'(\theta)$ is symmetric.*

We shall show in Section 1.4 that the limit in probability of $\bar{\psi}'_n(\theta)$ as $n \to \infty$ is $I_{\mathbf{G}}(\theta_{\mathbf{G}})$ which must therefore also be symmetric. If $I_{\mathbf{G}}(\theta_{\mathbf{G}})$ or $\bar{\psi}'_n(\theta)$ is a covariance matrix or a matrix of second derivatives of some function with a minimum at $\theta_{\mathbf{G}}$ (which is the case for our area of interest), then the assumption is not required.

Thus, the covariance matrix of the difference between our parameter estimator and its asymptotic limit, suitably normalised, is:

$$\begin{aligned}
\mathsf{Var}[B_n A_n^{-1}(\hat{\theta}_n - \theta_{\mathbf{G}})] &= B_n A_n^{-1} I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}})\mathsf{Var}[\bar{\psi}_n(\theta_{\mathbf{G}})] I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}})^T A_n^{-1} B_n \quad \text{from (1.14)} \\
&= I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) B_n A_n^{-1} \mathsf{Var}[\bar{\psi}_n(\theta_{\mathbf{G}})] A_n^{-1} B_n I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) \qquad (1.16) \\
&\quad \text{by Assumptions 7, 10 and 11} \\
&= I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) B_n \mathsf{Var}[\psi_n(\theta_{\mathbf{G}})] B_n I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) \\
&\to I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) C_{\mathbf{G}}(\theta_{\mathbf{G}}) I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) \quad \text{by Assumption 10.} \qquad (1.17)
\end{aligned}$$

So, our standardised $B_n A_n^{-1}(\hat{\theta}_n - \theta_{\mathbf{G}})$ has asymptotic expectation 0 and variance $I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) C_{\mathbf{G}}(\theta_{\mathbf{G}}) I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}})$. The form of the normalising matrix, $B_n A_n^{-1}$, looks cumbersome but reflects what is required. This ratio of normalising matrices is not apparent in many situations as, per the iid case analysed above and below, the net effect is $\sqrt{n} I_{p\times p}$.

In many situations Central Limit type arguments can be used to show that the $\bar{\psi}_n$ (or ψ_n) are asymptotically normally distributed at $\theta_{\mathbf{G}}$. In such cases, from (1.15) and (1.16):

$$\begin{aligned}
B_n A_n^{-1}(\hat{\theta}_n - \theta_{\mathbf{G}}) &\sim \mathsf{MVN}(0, I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) C_{\mathbf{G}}(\theta_{\mathbf{G}}) I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}})) \quad \text{or} \qquad (1.18) \\
&\sim \mathsf{MVN}(0, S_{\mathbf{G}}(\theta_{\mathbf{G}}))
\end{aligned}$$

where $S_{\mathbf{G}}(\theta_{\mathbf{G}}) = I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) C_{\mathbf{G}}(\theta_{\mathbf{G}}) I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}})$. Alternatively, and omitting **G** for simplicity:

$$Z_{\mathbf{G}} = S_{\mathbf{G}}^{-\frac{1}{2}} B_n A_n^{-1}(\hat{\theta}_n - \theta_{\mathbf{G}}) \sim \mathsf{MVN}(0, I_{p\times p}). \qquad (1.19)$$

Note that $S_{\mathbf{G}}$ is a covariance matrix and thus positive semidefinite, so there is a unique $S_{\mathbf{G}}^{\frac{1}{2}}$, a real positive semidefinite and symmetric square root (Horn and Johnson, 1987, theorem 7.2.6, page 405) which we will assume we have chosen. For independent Y_i

(the score, $U(\theta)$ is a sum of independent elements from a common distribution, and thus normal by the Central Limit Theorem) we have:

$$\sqrt{n}(\hat{\theta}_n - \theta_{\mathbf{G}}) \sim \mathsf{MVN}(0, \mathsf{E}[U'(\theta_{\mathbf{G}})]\mathsf{Var}[U(\theta_{\mathbf{G}})]\mathsf{E}[U'(\theta_{\mathbf{G}})]).$$

This result can be extended in a similar fashion to martingales, with their limited dependence (Crowder, 1986).

A number of papers, such as Sweeting (1980) and others cited therein, such as Bhat (1974), take a more general approach (ie do not assume the usual iid observations and the standard set of assumptions) to the asymptotic distribution of $\hat{\theta}_n$. Some of the assumptions are similar to those presented here but most of them also assume some variant on Bartlett's second identity (sometimes referred to as the *information identity*, eg in Varin and Vidoni, 2005):

$$\mathsf{E}[\psi(\theta_{\mathbf{G}})\psi(\theta_{\mathbf{G}})^T] = -\mathsf{E}[\psi'(\theta_{\mathbf{G}})]. \tag{1.20}$$

This does not hold generally in the environment we are working with here for the same reason that Bartlett's first identity fails in (1.3), namely that there is no reason to expect that the parametric family **F** used to form likelihoods will contain the unknown **G**. The Bartlett identities, arising from differentiating $\int f(y;\theta) = 1$ with respect to θ for a density f are also known as balance equations (Barndorff-Nielsen and Cox, 1994). Restoring the information identity plays a key role in adjusted surrogates as discussed in Section 1.6.

The results in this section assume that there are no nuisance parameters. In practice, that will not normally be the case and θ can be partitioned into (θ_1, θ_2) where θ_1 is a vector of parameters of interest and θ_2 of nuisance parameters. If it is not possible to integrate out the nuisance parameters (see the simulation in Section 2.7.1 where we do just that), then a common approach to maximum likelihood estimation is to profile them out (see, for instance, Garthwaite et al., 2006). This involves working with the profile loglikelihood:

$$\ell_p(\theta_1) = \ell(\theta_1, \hat{\theta}_2(\theta_1))$$

with maxima at $\hat{\theta}_1$ where $\hat{\theta}_2(\theta_1)$ is the vector of values of the nuisance parameter which maximise the loglikelihood for fixed θ_1. The resulting semiparametric likelihood is shown in Murphy and van der Vaart (2000), under some fairly general conditions, to have a quadratic expansion whose derivative terms are based around a score function shorn of its nuisance elements, so that the usual normal asymptotic and test statistic results (see Section 1.5) apply. This uses the standard result that $\hat{\theta}_1$ is equal to the estimate of the parameter of interest under maximum likelihood estimation for the full loglikelihood, $\ell(\theta_1, \theta_2)$.

1.4 Estimation of Theoretical Expressions

Most of the results from Section 1.3 include expectations of various expressions taken over the distribution **G**. However, in general, we will not know that distribution and thus need to estimate the expectations using the only information we have, namely the data. Specifically, the values we will need to estimate are:

- $\theta_{\mathbf{G}}$, the solution to $\mathsf{E}[\psi(\theta; Y)] = 0$.
- $I_{\mathbf{G}}(\theta_{\mathbf{G}}) = -\psi'_\infty(\theta_{\mathbf{G}})$.
- $C_{\mathbf{G}}(\theta_{\mathbf{G}})$, the limit in probability of $B_n \mathsf{Var}[\psi_n(\theta)] B_n$.

The approach is to consistently estimate each of these using the data. In the case of the first of the three, we shall use $\hat{\theta}_n$ from the consistency proof in Section 1.3.

Assumption 12. *The families of functions $\{\bar{\psi}_n(\theta)\}$ and $\{\bar{\psi}'_n(\theta)\}$ are, in probability, equicontinuous at $\theta_{\mathbf{G}}$.*

Equicontinuity means that in the neighbourhood of $\theta_{\mathbf{G}}$ nothing untoward occurs to members of the families as $n \to \infty$. Formally, a countable family of functions, $\mathcal{F}$, is equicontinuous at a point θ_0 if as $\theta_n \to \theta_0$, $\sup_{f\in\mathcal{F}}(f(\theta_n) - f(\theta_0)) \to 0$ (see, for instance Billingsley, 1999, Chapter 2). We extend this probabilistically at θ_{G} so that $\bar{\psi}_n(\hat{\theta}_n) - \bar{\psi}_n(\theta_{\mathsf{G}}) = o_p(1)$ and $\bar{\psi}'_n(\hat{\theta}_n) - \bar{\psi}'_n(\theta_{\mathsf{G}}) = o_p(1)$ as $\hat{\theta}_n \xrightarrow{p} \theta_{\mathsf{G}}$.

Now, for our second approximation:

$$\begin{aligned}
-\bar{\psi}'_n(\hat{\theta}_n) &= -\bar{\psi}'_n(\theta_{\mathsf{G}}) + o_p(1) \quad \text{by Assumption 12} \\
&\xrightarrow{p} -\psi'_\infty(\theta_{\mathsf{G}}) \quad \text{by Assumption 7} \\
&= I_{\mathsf{G}}(\theta_{\mathsf{G}})
\end{aligned}$$

to give our approximation.

Dealing with $C_{\mathsf{G}}(\theta_{\mathsf{G}})$ is more complex. If one were to follow the same route as for $I_{\mathsf{G}}(\theta_{\mathsf{G}})$, one might select:

$$B_n \psi_n(\hat{\theta}_n)\psi_n(\hat{\theta}_n)^T B_n.$$

Unfortunately, it suffers from the weakness that each of the middle terms is identically 0 by the definition of $\hat{\theta}_n$. In fact, the potential complexity of ψ_n means that in the general case one can only make an additional assumption.

Assumption 13. *There exists nonsingular symmetric matrix $\hat{V}_n(\theta)$, such that at $\theta = \hat{\theta}_n$ for all n it is not identically 0 and*

$$Var[\psi_n(\hat{\theta}_n)] \xrightarrow{p} \hat{V}_n(\hat{\theta}_n).$$

As a consequence:

$$\begin{aligned}
B_n \hat{V}_n(\hat{\theta}_n) B_n &= B_n \mathsf{Var}[\psi_n(\hat{\theta}_n)] B_n + o_p(1) \\
&\xrightarrow{p} C_{\mathsf{G}}(\theta_{\mathsf{G}}).
\end{aligned}$$

We are looking for a sequence of matrices, $\{\hat{V}_n(\theta)\}$ that is close to, or may even be, $\mathsf{Var}[\bar{\psi}_n(\theta)]$. In general, unlike the following case but similarly to the discussion after

(1.7), that variance will not be expressible in terms of the simpler $\psi(\theta; y_i)$ function.

The simplest example is for iid Y_i where $\hat{V}_n(\theta; Y)$ can be chosen as:

$$\hat{V}_n(\theta; Y) = \sum_{i=1}^{n} \psi(\theta; Y_i)\psi(\theta; Y_i)^T. \tag{1.21}$$

So, as $B_n = I_{p\times p}/\sqrt{n}$:

$$\begin{aligned} B_n\hat{V}_n(\hat{\theta}_n)B_n &= \frac{\hat{V}_n(\hat{\theta}_n)}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} \psi(\hat{\theta}_n; Y_i)\psi(\hat{\theta}_n; Y_i)^T \\ &= \frac{1}{n}\sum_{i=1}^{n} (\psi(\hat{\theta}_n; Y_i)\psi(\hat{\theta}_n; Y_i)^T - \mathsf{E}[\psi(\theta_{\mathbf{G}}; Y_i)]\mathsf{E}[\psi(\theta_{\mathbf{G}}; Y_i)]^T) \end{aligned} \tag{1.22}$$

since $\mathsf{E}[\psi(\theta_{\mathbf{G}}; Y_i)] = 0$. Now, the continuous mapping theorem states that for a continuous function $f(\theta)$, $\theta_n \xrightarrow{p} \theta_0$ implies that $f(\theta_n) \xrightarrow{p} f(\theta_0)$. Applying that with $f = \mathsf{E}[\psi]\mathsf{E}[\psi]^T$ at $\theta_{\mathbf{G}}$ we have:

$$\begin{aligned} B_n\hat{V}_n(\hat{\theta}_n)B_n &= \frac{1}{n}\sum_{i=1}^{n} \left(\psi(\hat{\theta}_n; Y_i)\psi(\hat{\theta}_n; Y_i)^T - \mathsf{E}[\psi(\hat{\theta}_n; Y_i)]\mathsf{E}[\psi(\hat{\theta}_n; Y_i)]^T\right)(1 + o_p(1)) \\ &= \mathsf{Var}[\psi(\hat{\theta}_n; Y_i)](1 + o_p(1)) \\ &= n\mathsf{Var}[\bar{\psi}_n(\hat{\theta}_n; Y)](1 + o_p(1)) \quad \text{by (1.9)} \\ &\xrightarrow{p} C_{\mathbf{G}}(\theta_{\mathbf{G}}) \quad \text{by its definition} \end{aligned}$$

and we have our estimator. There may be other cases where a simple estimator, as above, can be found. This example relies on $\psi_n(\theta; y)$ being a linear combination of independent $\psi(\theta; y_i)$s.

Our estimates for $I_{\mathbf{G}}(\theta_{\mathbf{G}})$ and $C_{\mathbf{G}}(\theta_{\mathbf{G}})$, $\hat{I}_n \equiv -\bar{\psi}'_n(\hat{\theta}_n)$ and $\hat{C}_n \equiv B_n\hat{V}_n(\hat{\theta}_n)B_n$ respectively, use a consistent estimator for the parameter as well as a usable form for the function ψ_∞. Our parameter variance estimator, $\hat{I}_n^{-1}\hat{C}_n\hat{I}_n^{-1}$ will require:

Assumption 14. *$\hat{I}_n$ is nonsingular.*

Assumption 8 is the equivalent for $\boldsymbol{I}_{\mathbf{G}}(\theta_{\mathbf{G}})$. Cox (2006) provides some counter examples for scalar θ but they can all be shown to have measure zero. Smith (1989) gives the example of a transformation of a bivariate extreme value distribution where the Fisher Information becomes infinite as the parameter value tends towards its limit in a closed space.

Example I continued - Poisson Surrogate for Gaussian Data. We can use the form for $\hat{V}_n(\theta; Y)$ given at (1.21) so that

$$\hat{V}_n(\theta; Y) = \sum_{i=1}^{n} \left(\frac{y_i}{\theta} - 1 \right)^2 .$$

Thus, the estimators are:

$$\begin{aligned}
\hat{I}_n &= -\bar{\psi}'_n(\hat{\theta}_n) \\
&= \frac{\sum_{i=1}^n y_i}{n\hat{\theta}_n^2} \\
&= \frac{1}{\hat{\theta}_n} \quad \text{and} \\
\hat{C}_n &= \frac{\hat{V}_n(\hat{\theta}_n)}{n} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\hat{\theta}_n} - 1 \right)^2 \\
&= \left(\frac{\sum_{i=1}^n y_i^2}{n\hat{\theta}_n^2} - 1 \right) .
\end{aligned}$$

Given n datapoints, we use for the variance of $\sqrt{n}(\hat{\theta}_n)$ in (1.18):

$$\frac{\sum_{i=1}^n y_i^2}{n} - \hat{\theta}_n^2 .$$

This is $(n-1)/n$ times the sample variance, and so is the same value that we would have arrived at if we had not used a surrogate but **G** with a known normal form. It is biased by a factor of $(n-1)/n$. If **G** had a known Poisson form then the variance would have been $\hat{\theta}_n$. □

1.5 Tests

In this subsection we shall assume that $B_n A_n^{-1}(\hat{\theta}_n - \theta_{\mathbf{G}})$ is asymptotically normally distributed resulting in (1.19). Thus, we have what one might term a *sandwich Wald statistic* for testing the null hypothesis $H_0 : \theta_{\mathbf{G}} = \theta_*$ (we discuss nuisance parameters briefly at the end of the section):

$$\begin{aligned} T_W &= (S_{\mathbf{G}}^{-\frac{1}{2}} B_n A_n^{-1}(\hat{\theta}_n - \theta_*))^T S_{\mathbf{G}}^{-\frac{1}{2}} B_n A_n^{-1}(\hat{\theta}_n - \theta_*) \qquad (1.23) \\ &= (\hat{\theta}_n - \theta_*)^T A_n^{-1} B_N S_{\mathbf{G}}^{-1} B_n A_n^{-1}(\hat{\theta}_n - \theta_*) \\ &\sim \chi_p^2 \quad \text{asymptotically under the null hypothesis.} \end{aligned}$$

where $S_{\mathbf{G}}(\theta_{\mathbf{G}}) = I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}}) C_{\mathbf{G}}(\theta_{\mathbf{G}}) I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}})$. The penultimate step is due to $S_{\mathbf{G}}$ being symmetric as shown after (1.19). By inverting $S_{\mathbf{G}}$, we have tacitly assumed that $C_{\mathbf{G}}(\theta_{\mathbf{G}})$ is invertible. We will formalise that and extend it to its estimate, which we require in Section 1.6:

Assumption 15. *$C_G(\theta_G)$ and $\hat{C}_n$ are nonsingular.*

Since $C_{\mathbf{G}}(\theta_{\mathbf{G}})$ is a transformation of covariance matrix, it is already positive semidefinite and Assumption 15 means it is positive definite.

Under SMLE, T_W is not generally asymptotically equivalent to the *naive likelihood ratio statistic* (see below for a Wald type statistic that is):

$$W_l = 2(\ell_{\mathbf{F}}(\hat{\theta}_n) - \ell_{\mathbf{F}}(\theta_*)). \qquad (1.24)$$

where $\ell_{\mathbf{F}}(\theta)$ is the loglikelihood of the data arising from **F** at θ, which takes no account of the sandwich adjustment to allow for any difference between **F** and **G**. In general for correctly specified models, the likelihood ratio test is preferred as it appears to give more accurate agreement between the true and asymptotic distributions ('substantial body of literature' - Young and Smith 2005), is invariant to reparameterisation (unlike the Wald test), is not as numerically unstable as the equivalent score test, and satisfies the Neyman-Pearson lemma for simple hypotheses extended to uniformly most powerful unbiased tests for composite ones (see Barndorff-Nielsen and Cox, 1994, chapter 4.2 for a more complete discussion).

It can be shown that, when $B_n^2 = A_n$, the distribution of what one might call a *naive Wald statistic*:

$$\begin{aligned} W_W &= (B_n A_n^{-1}(\hat{\theta}_n - \theta_*))^T I_{\mathbf{G}}(\theta_*) B_n A_n^{-1}(\hat{\theta}_n - \theta_*) \qquad (1.25) \\ &= (\hat{\theta}_n - \theta_*)^T A_n^{-1} B_N I_{\mathbf{G}}(\theta_*) B_n A_n^{-1}(\hat{\theta}_n - \theta_*) \\ &= (\hat{\theta}_n - \theta_*)^T A_n^{-\frac{1}{2}} I_{\mathbf{G}}(\theta_*) A_n^{-\frac{1}{2}}(\hat{\theta}_n - \theta_*) \end{aligned}$$

is asymptotically equivalent to that of the naive likelihood ratio statistic (under SMLE) and the *naive score statistic*:

$$W_s = \psi_n(\theta_*)^T A_n^{\frac{1}{2}} I_{\mathbf{G}}^{-1}(\theta_*) A_n^{\frac{1}{2}} \psi_n(\theta_*) \qquad (1.26)$$

(see Appendix A). The restriction on the normalising matrices, A_n and B_n is not is severe as it initially appears: A_n is used to normalise ψ_n and B_n^2 for Var$[\psi_n]$ so in many cases $B_n^2 = A_n$. For instance, with iid Y_i, $A_n = I_{p\times p}/n$ and $B_n = I_{p\times p}/\sqrt{n}$, and the condition holds for both simulations used in this thesis. The asymptotic distribution of W_W is not obvious but we shall now derive it.

Reorganising and simplifying notation by using ψ, $C_{\mathbf{G}}$ and $I_{\mathbf{G}}$ etc wherever possible:

$$\begin{aligned} W_W &= (S_{\mathbf{G}}^{\frac{1}{2}} S_{\mathbf{G}}^{-\frac{1}{2}} B_n A_n^{-1}(\hat{\theta}_n - \theta_*))^T I_{\mathbf{G}} (S_{\mathbf{G}}^{\frac{1}{2}} S_{\mathbf{G}}^{-\frac{1}{2}} B_n A_n^{-1}(\hat{\theta}_n - \theta_*)) \\ &= Z_{\mathbf{G}}^T (S_{\mathbf{G}}^{\frac{1}{2}})^T I_{\mathbf{G}} S_{\mathbf{G}}^{\frac{1}{2}} Z_{\mathbf{G}} \quad \text{per (1.19)} \\ &= Z_{\mathbf{G}}^T M Z_{\mathbf{G}} \quad \text{say.} \qquad (1.27) \end{aligned}$$

M is symmetric since $I_{\mathbf{G}}$ is. We can then use a spectral decomposition (Krzanowski, 2000, page 539):

$$M = L^T D L$$

where L is a matrix consisting of orthonormal eigenvectors of M as its columns and D

a diagonal matrix of the corresponding eigenvalues, so that expanding (1.27):

$$\begin{aligned} W_W &= Z_{\mathbf{G}}^T L^T D L Z_{\mathbf{G}} \\ &= Y_{\mathbf{G}}^T D Y_{\mathbf{G}} \end{aligned}$$

where $Y_{\mathbf{G}} = LZ_{\mathbf{G}}$. Since $Z_{\mathbf{G}} \sim \mathsf{MVN}(0, I_{p\times p})$ by (1.19):

$$\begin{aligned} Y_{\mathbf{G}} &\sim \mathsf{MVN}(0, LL^T) \\ &= \mathsf{MVN}(0, I_{p\times p}) \end{aligned}$$

as L is a matrix of orthonormal elements. As a consequence, asymptotically:

$$W_W \sim \sum_{i=1}^{p} d_i V_i$$

where the d_i are eigenvalues of $S_{\mathbf{G}}^{\frac{1}{2}} I_{\mathbf{G}} S_{\mathbf{G}}^{\frac{1}{2}} = M$ and the V_i are independent χ_1^2 variables. The result applies equally well to W_s and W_l by asymptotic equivalence as in Appendix A.

Clearly, if all the above eigenvalues of M are zero or one, then we have a standard χ^2 distribution with degrees of freedom the number of ones. This will be the case if M is *idempotent*, ie $M = M^2$ (indeed, idempotency is given as a necessary and sufficient condition for a standard χ^2 distribution in such a case in Mathai and Provost, 1992, Theorem 5.1.1).

It is worth noting that the only point in this section at which we use asymptotic normality is in describing the distribution of $Z_{\mathbf{G}}$. If we drop that condition, then we still have $W_W = Y_{\mathbf{G}}^T D Y_{\mathbf{G}}$, but $Y_{\mathbf{G}}$ is not necessarily distributed as $\mathsf{MVN}(0, I_{p\times p})$.

Kent (1982) extends this result to the situation where there are nuisance parameters and outlines a proof using the same approach as above for profile likelihoods where the Hessian and variance matrices are partitioned into blocks corresponding to the parameters of interest and the nuisance parameters.

In the case of composite surrogates (see Chapter 2), Aerts and Claeskens (1999) suggest that using the parametric bootstrap for clustered binary data will give a consistent estimator for the likelihood ratio statistic which is simpler to calculate and more robust than the standard approach using eigenvectors outlined in this section. However, that requires one to be able to simulate from the full multivariate distribution.

1.6 Adjusted Surrogates

1.6.1 General Approach

We saw in Section 1.5 that Wald tests (or, asymptotically, likelihood ratio tests) for surrogates are either, (1.23), based upon the sandwich matrix

$$S_{\mathbf{G}}(\theta_{\mathbf{G}})^{-1} = I_{\mathbf{G}}(\theta_{\mathbf{G}})C_{\mathbf{G}}(\theta_{\mathbf{G}})^{-1}I_{\mathbf{G}}(\theta_{\mathbf{G}}),$$

in which case, for null hypothesis of dimension l, the resulting statistic is asymptotically χ^2_l, or, (1.25), upon $I_{\mathbf{G}}(\theta_{\mathbf{G}})$ in which case the resulting statistic is asymptotically a weighted sum of χ^2_1 distributions. Ideally, we would like to use the statistic from the latter case but the distribution from the former as they are the simplest. That would be possible if

$$I_{\mathbf{G}}(\theta_{\mathbf{G}}) = C_{\mathbf{G}}(\theta_{\mathbf{G}}),$$

ie the information identity, (1.20), held, as then

$$S_{\mathbf{G}}(\theta_{\mathbf{G}})^{-1} = I_{\mathbf{G}}(\theta_{\mathbf{G}}). \tag{1.28}$$

A number of papers, discussed in this section, begin with a surrogate likelihood, likelihood ratio or estimating function and then adjust it so that (1.28) holds. Clearly, the adjusted estimating function will still need to have an expected zero at $\theta_{\mathbf{G}}$. If ψ_a is our adjusted estimating function we will then require that

$$-\mathsf{E}[\psi_a'(\theta_{\mathbf{G}})] = \mathsf{Var}[\psi_a(\theta_{\mathbf{G}})]$$

or for n data elements, estimated parameters, θ_{an}, and estimating function, ψ_{an},

$$-\left.\frac{\partial \psi_{an}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{an}} = \mathsf{Var}[\psi_{an}(\hat{\theta}_{an})] \tag{1.29}$$

ie asymptotically we restore equality between the variance of the estimating function and the inverse of Fisher's information matrix. A consequence is that the variance of the parameter estimator will be, from a version of (1.14) without the normalisation, for n data elements

$$\begin{aligned}
\mathsf{Var}[\hat{\theta}_{an}] &= \left(\left.\frac{\partial \psi_{an}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{an}}\right)^{-1} \mathsf{Var}[\psi_{an}(\hat{\theta}_{an})] \left(\left.\frac{\partial \psi_{an}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{an}}\right)^{-1} (1+o_p(1)) \\
&= -\left(\left.\frac{\partial \psi_{an}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{an}}\right)^{-1} (1+o_p(1)) \\
&= -\psi_{an}'(\hat{\theta}_{an})(1+o_p(1)) \quad \mathsf{say} \qquad (1.30)
\end{aligned}$$

whereas for the unadjusted estimator, $\hat{\theta}_n$

$$\begin{aligned}
\mathsf{Var}[\hat{\theta}_{n}] &= \left(\left.\frac{\partial \psi_{n}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{n}}\right)^{-1} \mathsf{Var}[\psi_{n}(\hat{\theta}_{n})] \left(\left.\frac{\partial \psi_{n}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{n}}\right)^{-1} (1+o_p(1)) \\
&= \psi_n'(\hat{\theta}_n)^{-1}\mathsf{Var}[\psi_n(\hat{\theta}_n)]\psi_n'(\hat{\theta}_n)^{-1}(1+o_p(1)) \qquad (1.31)
\end{aligned}$$

Further justification for using the information identity, which may be necessary if the parameter estimates are not normally distributed as assumed in Section 1.5, has been given as:

Optimality McCullagh and Tibshirani (1990) point out that the sandwich information (see Section 1.7.2) is maximised when the identity holds (see, for instance Song, 2007), for a "limited" class of estimating functions. It assumes a known **G** and reflects an extension of maximum likelihood estimation.

Robustness Stafford (1996) and Royall and Tsou (2003) claim that the identity im-proves robustness. The former suggests invariance to certain forms of reparameterisation as a definition while the latter introduces a more formal approach: a 'bump' function assesses the probability of misleading evidence (likelihood ratios greater than, say, k) for a range of potential parameter values, which is bounded over all θ as a function of k when the information identity holds, but bounded by a function that depends upon the data and $\theta_{\mathbf{G}}$ otherwise. See Section 1.7.4 for a longer discussion on the use of the term 'robust'.

Approximation An informal argument is used in McCullagh and Tibshirani (1990) to suggest that an adjusted likelihood that conforms to this criterion is likely to give a more accurate approximation of both the variance and the χ^2 distribution of the loglikelihood ratio statistic than one that does not. Some exponential family examples are given. This argument is described in the case of an adjusted profile loglikelihood, but, as we shall see, has wider application.

Adjustments are made either to the surrogate loglikelihood or to the estimating function directly, which we shall term *vertical*, or to the parameter within the likelihood, which we shall call *horizontal*, the names arising in Chandler and Bate (2007) and describing the effect of a change on a plot of parameter values versus loglikelihood in low dimensions.

1.6.2 Horizontal Parameter Adjustments

We consider adjustments of the form:

$$\ell_{ah}(\theta; y) = \ell(\theta_a; y)$$

where

$$\theta_a = c + L\theta \tag{1.32}$$

for some $p \times 1$ vector c and $p \times p$ non zero matrix B, both functions of the data but not θ, as the adjustment is linear. We will seek values of c and L such that (1.29) holds.

The adjusted estimating function is, with the ys omitted for simplicity

$$\psi_{ah}(\theta) = \frac{\partial \ell_{ah}(\theta)}{\partial \theta} = \left(\frac{\partial \theta_a}{\partial \theta}\right)^T \frac{\partial \ell_{ah}(\theta)}{\partial \theta_a} = \left(\frac{\partial \theta_a}{\partial \theta}\right)^T \frac{\partial \ell(\theta_a)}{\partial \theta_a} = L^T \frac{\partial \ell(\theta_a)}{\partial \theta_a},$$

where the transposed first term arises from our convention of treating the estimating function as a column vector (see the Notation note at the beginning of this thesis), with its observed equivalent for n data points:

$$\psi_{ahn}(\theta; y_1, \ldots, y_n) = L_n^T \frac{\partial \ell_n(\theta_a; y_1, \ldots, y_n)}{\partial \theta_a} \tag{1.33}$$

where $L_n \to L$ and, by Assumption 3, the left hand side has a unique root, $\hat{\theta}_{ahn}$ say, so that

$$\psi_{ahn}(\hat{\theta}_{ahn}; y_1, \ldots, y_n) = 0.$$

Note that for iid data, $\psi_{ahn}(\theta; y_1, \ldots, y_n) = \sum_{i=1}^n \psi(\theta_a; y_i)$. Repeating the differentiation

$$\frac{\partial \psi_{ah}(\theta)}{\partial \theta} = L^T \frac{\partial \psi(\theta_a)}{\partial \theta_a} L$$

with its observed equivalent for n data points:

$$\frac{\partial \psi_{ahn}(\theta)}{\partial \theta} = L_n^T \frac{\partial \psi_n(\theta_a)}{\partial \theta_a} L_n. \tag{1.34}$$

We will require that the estimating equations, (1.33), have zeroes at $\theta = \hat{\theta}_n$, ie

$$\hat{\theta}_{ahn} = \hat{\theta}_n. \tag{1.35}$$

An alternative would be to allow (1.33) to have new zeroes (ie not $\hat{\theta}_n$) which are also consistent for $\theta_{\mathbf{G}}$, but our current approach gives rise to a simple form of adjustment.

Since the left hand side of (1.33) has a unique value of θ for which it equals zero and θ_a is linear in θ, the right hand side must have a unique value of θ_a for which it, in turn, also equals zero (L_n is invertible - see after Assumption 16). Since $\theta_a = \hat{\theta}_n$ will certainly make the right hand side zero, it is a unique root so that, from (1.32)

$$\begin{aligned} \hat{\theta}_n &= c + L_n\hat{\theta}_n \qquad \text{and} \\ c &= \hat{\theta}_n - L_n\hat{\theta}_n. \end{aligned}$$

In turn, our adjusted parameter will be

$$\begin{aligned} \theta_a &= \hat{\theta}_n - L_n\hat{\theta}_n + L_n\theta \\ &= \hat{\theta}_n + L_n(\theta - \hat{\theta}_n) \end{aligned}$$

which is the form of adjustment proposed in Chandler and Bate (2007).

For our adjusted parameter estimator we would like the information identity and thus (1.30) to hold so that

$$\begin{aligned} \text{Var}[\hat{\theta}_{ahn}] &= -\psi'_{an}(\hat{\theta}_{ahn})(1 + o_p(1)) \\ &= -(L_n^T\psi'_n(\hat{\theta}_n)L_n)^{-1}(1 + o_p(1)) \quad \text{from (1.34), as } \hat{\theta}_{ahn} = \hat{\theta}_n \\ &= -L_n^{-1}\psi'_n(\hat{\theta}_n)^{-1}(L_n^T)^{-1}(1 + o_p(1)) \\ &= -L_n^{-1}(\psi'_n(\hat{\theta}_n))^{-1}(L_n^T)^{-1}(1 + o_p(1)) \end{aligned}$$

Assuming, for the moment, that M_n and N_n, $p \times p$ matrices, exist we define

$$\begin{aligned} M_n^T M_n &= -\psi'_n(\hat{\theta}_n) \\ N_n^T N_n &= \psi'_n(\hat{\theta}_n)\text{Var}[\psi_n(\hat{\theta}_n)]^{-1}\psi'_n(\hat{\theta}_n) \\ L_n &= M_n^{-1}N_n \end{aligned}$$

so that

$$\begin{aligned}
-L_n^{-1}\psi_n'(\hat{\theta}_n)^{-1}(L_n^T)^{-1} &= -N_n^{-1}M_n\psi_n'(\hat{\theta}_n)^{-1}M_n^T(N_n^T)^{-1} \\
&= N_n^{-1}(N_n^T)^{-1} \\
&= (N_n^TN_n)^{-1} \\
&= (\psi_n'(\hat{\theta}_n)\mathsf{Var}[\psi_n(\hat{\theta}_n)]^{-1}\psi_n'(\hat{\theta}_n))^{-1} \\
&= \psi_n'(\hat{\theta}_n)^{-1}\mathsf{Var}[\psi_n(\hat{\theta}_n)]\psi_n'(\hat{\theta}_n)^{-1} \\
&= \mathsf{Var}[\hat{\theta}_n](1+o_p(1)) \quad \text{per (1.31)}
\end{aligned}$$

and our unadjusted and adjusted parameter estimators have the same first two moments for large enough sets of data.

Now, we address the existence of L_n. We make

Assumption 16. *$\psi_n'(\hat{\theta}_n)$ is negative definite and Var$[\psi_n(\hat{\theta}_n)]$ is positive definite.*

As a result, and using Assumption 11, M_n^TM and $N_n^TN_n$ must be symmetric and positive definite. Therefore, M_n and N_n must exist as they could be genuine square root matrices (Horn and Johnson (1987), Theorem 7.2.6), or formed by a Cholesky decomposition. These roots must be positive definite and so $L_n = M_n^{-1}N_n$ exists, is positive definite and invertible.

Our horizontal adjustment then becomes:

$$\ell_{ah}(\theta; y) = \ell(\theta_a; y)$$

with

$$\theta_a = \hat{\theta}_n + M_n^{-1}N_n(\theta - \hat{\theta}_n) \tag{1.36}$$

for n data points. This adjustment was proposed in Chandler and Bate (2007) as an adjustment to a composite likelihood (see Chapter 2) consisting of univariate margins of some distribution: they show that the adjusted loglikelihood is a marked improvement (shown via power curves) on the unadjusted one and, where a comparison is possible, close to (but, clearly, not an improvement on) the true loglikelihood in calculating parameter estimates, standard errors and test statistics. We have derived it as the only linear adjustment of the parameters which matches the $\hat{\theta}_n$ from the unadjusted estimating equations, subject to Assumption 16, for any loglikelihood. We shall examine its use as an adjustment to a bivariate surrogate in Simulation I in Section 2.7.

Example I continued - Poisson Surrogate for Gaussian Data. We can easily derive

$$L_n = \sqrt{\frac{n\hat{\theta}_n}{\sum_{i=1}^n y_i^2 - n\hat{\theta}_n^2}} \tag{1.37}$$

leading to an adjusted loglikelihood

$$\ell_{ahn}(\theta) \propto \sum_{i=1}^n y_i \ln\left(\hat{\theta}_n + (\theta - \hat{\theta}_n)L_n\right) - n\left(\hat{\theta}_n + (\theta - \hat{\theta}_n)L_n\right)$$

and estimating equation

$$\psi_{ahn}(\hat{\theta}_n) = \frac{\sum_{i=1}^n y_i L_n}{\hat{\theta}_n + (\theta - \hat{\theta}_n)L_n} - nL_n = 0.$$

We solve for $\hat{\theta}_h$, the horizontally adjusted parameter estimate so that

$$\frac{L_n\left(\sum_{i=1}^n y_i - n\left(\hat{\theta}_n + (\hat{\theta}_h - \hat{\theta}_n)L_n\right)\right)}{\hat{\theta}_n + (\hat{\theta}_h - \hat{\theta}_n)L_n} = 0 \tag{1.38}$$

which has unique solution

$$\hat{\theta}_h = \hat{\theta}_n.$$

□

One could investigate nonlinear adjustments to the parameters by considering

$$\theta_a = c + L(\theta)\theta$$

in contrast to (1.32) where L was not a function of the parameters. However, there appear to be no simple forms akin to (1.36).

1.6.3 Vertical Estimating Function Adjustments

We now examine adjustments of the form

$$\psi_{av}(\theta) = D(\theta)\psi(\theta) + e(\theta) \tag{1.39}$$

for some $p \times 1$ vector $e(\theta)$ and $p \times p$ nonsingular non random matrix $D(\theta)$, ie we reshape and then centre the estimating functions. This is based upon, although more general (ie for multidimensional θ) than, adjustments proposed by Stafford (1996) and Royall and Tsou (2003), which, in turn, were based on an adjustment to the profile loglikelihood given in McCullagh and Tibshirani (1990). Note that adjustments of the form (1.39) are made to the estimating function rather than the likelihood function, as in the preceding section.

Firstly, we require that the adjusted estimating function has expected value 0 at $\theta_{\mathbf{G}}$:

$$\begin{aligned} 0 &= \mathsf{E}[\psi_{av}(\theta_{\mathbf{G}})] \\ &= \mathsf{E}[D(\theta_{\mathbf{G}})\psi(\theta_{\mathbf{G}}) + e(\theta_{\mathbf{G}})] \\ &= D(\theta_{\mathbf{G}})\mathsf{E}[\psi(\theta_{\mathbf{G}})] + e(\theta_{\mathbf{G}}) \\ &= e(\theta_{\mathbf{G}}) \end{aligned}$$

and so, no centring is required, ie our parameter estimates are unchanged by the adjustment.

Secondly, we would like the information identity to hold:

$$\begin{aligned}
\mathsf{E}\left[\frac{\mathrm{d}\psi_{av}(\theta)}{\mathrm{d}\theta}\right]\Bigg|_{\theta=\theta_{\mathbf{G}}} &= \\
\mathsf{E}\left[\frac{\mathrm{d}(D(\theta)\psi(\theta))}{\mathrm{d}\theta}\right]\Bigg|_{\theta=\theta_{\mathbf{G}}} &= \\
\mathsf{E}\left[\left(\frac{\mathrm{d}D(\theta)}{\mathrm{d}\theta}\right)^T \psi(\theta_{\mathbf{G}}) + D(\theta_{\mathbf{G}})\frac{\mathrm{d}\psi(\theta)}{\mathrm{d}\theta}\right]\Bigg|_{\theta=\theta_{\mathbf{G}}} &= \\
\left(\frac{\mathrm{d}D(\theta)}{\mathrm{d}\theta}\right)^T\Bigg|_{\theta=\theta_{\mathbf{G}}} \mathsf{E}[\psi(\theta_{\mathbf{G}})] + D(\theta_{\mathbf{G}})\mathsf{E}\left[\frac{\mathrm{d}\psi(\theta)}{\mathrm{d}\theta}\right]\Bigg|_{\theta=\theta_{\mathbf{G}}} &= \\
D(\theta_{\mathbf{G}})\mathsf{E}\left[\frac{\mathrm{d}\psi(\theta)}{\mathrm{d}\theta}\right]\Bigg|_{\theta=\theta_{\mathbf{G}}} &= -\mathsf{Var}[\psi_{av}(\theta_{\mathbf{G}})] \\
&= -\mathsf{Var}[D(\theta_{\mathbf{G}})\psi(\theta_{\mathbf{G}})] \\
&= -D(\theta_{\mathbf{G}})\mathsf{Var}[\psi(\theta)]D(\theta_{\mathbf{G}})^T
\end{aligned}$$

where $\frac{\mathrm{d}D(\theta)}{\mathrm{d}\theta}$ is a p-vector of $p \times p$ matrices as defined in the Notation section at the start of this thesis. Thus:

$$\begin{aligned}
-\mathsf{Var}[\psi(\theta_{\mathbf{G}})]^{-1}\mathsf{E}\left[\frac{\mathrm{d}\psi(\theta)}{\mathrm{d}\theta}\right]\Bigg|_{\theta=\theta_{\mathbf{G}}} &= D(\theta_{\mathbf{G}})^T \\
&= D(\theta_{\mathbf{G}}) \qquad (1.40)
\end{aligned}$$

the final equality resulting from the symmetry of both elements in the left hand side. Clearly, if the information identity, (1.20), holds for $\psi(\theta_{\mathbf{G}})$, the adjustment is just the identity matrix.

We estimate both expected terms in the expression (1.40) per Section 1.4 to give:

$$\psi_{av}(\theta) = \hat{C}_n^{-1}\hat{I}_n\psi(\theta) \qquad (1.41)$$

so that we have used a consistent estimator, $\hat{\theta}_n$, for the parameter as well as a surrogate for the functions of ψ. $\hat{C}_n$ is non singular by Assumption 15 as it is the product of nonsingular matrices. The adjustment will have no effect on parameter estimates as we are just multiplying the estimating function by a constant. The variance of the estimator remains unchanged as the adjustments are cancelled out by the sandwich form. Vertical adjustments recover the information identity and therefore benefit from all the advantages set out at the beginning of Section 1.6.1. Note the similarity with the proposed optimal efficiency improving weights given at the end of Section 3.2.4.

Stafford (1996) then shows that the resulting likelihood (for scalar θ so that we can integrate the adjusted estimating function) has the following features:

- It is invariant under transformation. Pace et al. (2011) extend this to the multiple parameter case. Clearly, by definition, the horizontal adjustment is not parameter invariant, although that, in itself, does not preclude subtlety. It is direction dependent but is a matrix rather than scalar adjustment of the parameters.
- Test statistics can be calculated for the adjusted likelihood as described in Section 1.5.
- To study the effectiveness of the adjustment, estimates of test statistics for both the adjusted and the usual unadjusted likelihood are compared when the correct model is used. Their first and third cumulants are similar but the adjusted statistic has a larger variance and a formula for the relative efficiency is derived.

The situation for vector $\boldsymbol{\theta}$ is discussed in Section 1.6.4.

Example I continued - Poisson Surrogate for Gaussian Data. The vertical adjustment is

$$\frac{n\hat{\theta}_n}{\sum_{i=1}^n y_i^2 - n\hat{\theta}_n^2}.$$

This is the square of the horizontal adjustment at (1.37). The adjusted estimating equation becomes

$$\psi_{av}(\hat{\theta}_n) = \left(\frac{\sum_{i=1}^n y_i}{\theta} - n\right)\frac{n\hat{\theta}_n}{\sum_{i=1}^n y_i^2 - n\hat{\theta}_n^2} = 0$$

which we solve for $\hat{\theta}_v$, the vertically adjusted parameter estimate to give the unique solution

$$\hat{\theta}_v = \hat{\theta}_n$$

which is the same as that resulting from the horizontally adjusted and, obviously, unadjusted estimating equations. □

1.6.4 In Practice

In practice, one would prefer to use the likelihood ratio test, as opposed to the Score or Wald tests, for the reasons given in Section 1.5. For the horizontal adjustment, the likelihood ratio statistic for nested models is

$$\begin{aligned}\Delta_{ah} &= 2(\ell_{ah}(\hat{\theta}_{ahn}) - \ell_{ah}(\tilde{\theta}_{ahn})) \\ &= 2(\ell_{ah}(\hat{\theta}_n) - \ell_{ah}(\tilde{\theta}_{ahn}))\end{aligned}$$

as parameter estimates are unchanged by the adjustment, as discussed in Section 1.6.2, and where $\tilde{\theta}_{ahn}$ maximises $\ell(\theta)_{ahn}$ subject to the restriction that we are testing, typically $\Delta\theta = \delta_*$. This requires an additional maximisation process to calculate $\tilde{\theta}_{ahn}$. Instead, Chandler and Bate (2007) suggest using

$$\Delta^*_{ah} = 2c(\ell_{\mathsf{F}}(\hat{\theta}_n) - \ell_{\mathsf{F}}(\tilde{\theta}_n))$$

where $\tilde{\theta}_n$ maximises the unadjusted surrogate loglikelihood $\ell_{\mathsf{F}}(\theta)$ subject to $\Delta\theta = \delta_*$, which would be calculated anyway for the unadjusted bivariate test, and c is a ratio of quadratic approximations to

$$\frac{\ell_{ah}(\hat{\theta}_n) - \ell_{ah}(\tilde{\theta}_{ahn})}{\ell_{\mathsf{F}}(\hat{\theta}_n) - \ell_{\mathsf{F}}(\tilde{\theta}_n)}$$

which is easily calculated. This can produce substantial improvements to using the unadjusted test statistic (Chandler and Bate, 2007) where comparisons are made by power curves, but, as we shall see in Section 2.7, that is not always the case.

For the vertical adjustment, things are more complex. For scalar θ, one can integrate the estimating function to give a loglikelihood function, $\ell_{av}(\theta)$ and one can use the likelihood ratio test as usual. This will also be true for some vector $\boldsymbol{\theta}$s but only where an integrated function can be consistently derived. It is worth noting that for quadratic loglikelihoods with vector θ, a necessary and sufficient condition for the adjusted estimating function to integrate consistently is that the adjusted Hessian is symmetric.

Chandler and Bate (2007) get round the issue of integrating an adjusted estimating function for vector $\boldsymbol{\theta}$ by suggesting a similar looking vertical adjustment, but applied to the loglikelihood:

$$\ell^*_{av2}(\boldsymbol{\theta}) = \ell_{\mathsf{F}}(\hat{\boldsymbol{\theta}}_n) + \frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^T \hat{\boldsymbol{I}}_n \hat{\boldsymbol{C}}_n^{-1} \hat{\boldsymbol{I}}_n (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)}{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^T \hat{\boldsymbol{I}}_n (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)} (\ell_{\mathsf{F}}(\boldsymbol{\theta}) - \ell_{\mathsf{F}}(\hat{\boldsymbol{\theta}}_n)) \tag{1.42}$$

where the terms in used in the adjustment at (1.41) are converted to quadratic forms. The result is an adjusted likelihood ratio statistic that can be calculated in all cases. If θ is a scalar, then ℓ^*_{av2} equals the likelihood ratio statistic that results from using the adjustment at (1.41).

However, Pace et al. (2011) point out that the approximation (1.42) is not invariant to reparameterisation and, again, we shall see in Section 2.7 that it does not always produce improvements over the unadjusted test statistic. Pace et al. (2011) suggest a further form of vertical adjustment which is parameterisation invariant and whose distribution is asymptotically χ^2_l. However, it requires an additional maximisation process and is thus not as computationally efficient as the horizontal and vertical adjustments described earlier in this section.

As a way of understanding the use of adjusted surrogates in practice, if we are testing $\theta = \theta^*$ (a similar approach is taken for testing a subset of θ), we can use the sandwich Wald statistic

$$\begin{aligned} T_W &= (\hat{\theta}_n - \theta_*)^T A_n^{-\frac{1}{2}} S_{\mathbf{G}}^{-1} A_n^{-\frac{1}{2}} (\hat{\theta}_n - \theta_*) \qquad (1.43) \\ &\sim \chi^2_l \quad \text{asymptotically under the null hypothesis} \end{aligned}$$

or we can apply the adjustment

$$(\hat{C}_n^{-1}\hat{I}_n)^{-1} = (\hat{I}_n^{-1}(\hat{I}_n\hat{C}_n^{-1}\hat{I}_n))^{-1} \tag{1.44}$$

resulting in the simple Wald statistic, (1.25), with the weighted χ_1^2 distribution, and then we can apply the inverse of (1.44), or $\hat{C}_n^{-1}\hat{I}_n$, resulting in, say, the vertically adjusted surrogate with the original distribution.

1.7 Comparing Surrogates

1.7.1 Introduction

For any particular dataset, one might wish to use and then compare a number of surrogates. In certain well behaved theoretical models, analytical results based around comparisons of estimates and efficiency can be extracted (eg Cox and Reid, 2004). Otherwise, criteria used for comparison include, in simulations where one often has the maximum likelihood estimator from **G** for comparison and in practice:

Power Power curves are created by plotting the power of tests as a function of the parameters of interest (eg Chandler and Bate, 2007). We use this in Section 2.7. The values at $\theta = 0$, testing the null hypothesis $H_0 : \theta = 0$,should equal α, the Type I error rate, and this is sometimes used as a criterion for comparison (eg Aerts and Claeskens, 1999). Power is a useful tool for comparing procedures (Cox, 2006, page 25). However, there are situations where such comparisons are misleading (Young and Smith, 2005), dealt with by use of the conditionality principle (if the minimal sufficient statistic can be partitioned as (S, C), where C does not depend upon θ, then inference about θ should be based upon $S|C$), but they do not apply to the examples we study in this thesis.

Confidence Intervals Coverage of confidence intervals (CIs) for parameter estimates can be compared numerically or graphically (eg Heagerty and Lumley, 2000).

Efficiency We analyse options for the multivariate parameter case in Section 1.7.2 and make use of some of them in Sections 3.4.3 and 4.4. The scalar parameter case is often used (eg Bevilacqua et al., 2012).

Model Selection One can extend model selection criteria such as the AIC to a comparison of surrogates. These are reviewed in Section 1.7.3.

Robustness There are a number of concepts of robustness used in the literature and these are reviewed in Section 1.7.4.

In comparing estimating functions, we first understand whether classes of them are equivalent so that their outcomes under various of the criteria listed above are the same. For estimating functions, $\psi(\theta)$, and any nonsingular non random $p \times p$ matrix $D(\theta)$, at θ_{G}:

$$\mathsf{E}[D(\theta_{\mathsf{G}})\psi(\theta_{\mathsf{G}})] = D(\theta_{\mathsf{G}})\mathsf{E}[\boldsymbol{\psi}(\theta_{\mathsf{G}})] = 0$$

as $\mathsf{E}[\psi(\theta_{\mathsf{G}})] = 0$, so that $D(\theta)\psi(\theta)$ is also an estimating function with the same zero. We establish an equivalence relation, $\sim$, on $\mathcal{E}$, the class of estimating functions defined in Section 1.2, whereby for $\psi, \phi \in \mathcal{E}$:

$$\psi \sim \phi \qquad \text{iff} \qquad \psi(\theta) = D(\theta)\phi(\theta)$$

for some $D(\theta)$, resulting in a set of equivalence classes for use in comparisons.

It is simplest to represent each class by an estimating function with a particular feature and here we shall use gradient as in, for instance, Davison (2003, page 318). We select a standardised estimating function (approximations for observed data are given at the end of Section 1.7.2), namely:

$$(\mathsf{E}[\psi'(\theta)])^{-1}\psi(\theta). \tag{1.45}$$

Note that $\mathsf{E}[\psi'(\theta; Y)]$ is nonsingular in an area around θ_{G}: from White (1982, Theorem 3.1(i)) it is negative definite (subject to a number of the assumptions set out earlier in this thesis) and thus it has an inverse. Intuitively, in the situations where $\psi(\theta; Y)$ is a gradient vector, $\mathsf{E}[\psi'(\theta; Y)]$ is the change in gradient around a maximum in all dimensions and it makes sense for it to be negative definite.

Then, for any other equivalence class member, $D(\theta)\psi(\theta)$ at $\theta_{\mathbf{G}}$ (abbreviated to $D\psi$ etc):

$$\begin{aligned}
\left(\mathsf{E}\left[\frac{\partial(D\psi)}{\partial\theta^T}\right]\right)^{-1} D\psi &= (\mathsf{E}[(D\psi)'])^{-1}D\psi \quad \text{say} \\
&= (\mathsf{E}[D'\psi + D\psi'])^{-1}D\psi \\
&= (D'\mathsf{E}[\psi] + D\mathsf{E}[\psi'])^{-1}D\psi \\
&= (\mathsf{E}[\psi'])^{-1}D^{-1}D\psi \\
&= (\mathsf{E}[\psi'])^{-1}\psi
\end{aligned}$$

so that normalising any member of a class by its expected gradient will give rise to the same class representative at $\theta_{\mathbf{G}}$.

The reason for selecting (1.45) as a representative is that the expected gradient at $\theta_{\mathbf{G}}$ (omitting the $\theta_{\mathbf{G}}$s to simplify notation) is then:

$$\begin{aligned}
\mathsf{E}\left[\frac{\partial((\mathsf{E}[\psi'])^{-1}\psi)}{\partial\theta^T}\right] &= \mathsf{E}\left[\frac{\partial(\mathsf{E}[\psi'])^{-1}}{\partial\theta^T}\psi\right] + \mathsf{E}[(\mathsf{E}[\psi'])^{-1}\psi'] \\
&= \frac{\partial(\mathsf{E}[\psi'])^{-1}}{\partial\theta^T}\mathsf{E}[\psi] + (\mathsf{E}[\psi'])^{-1}\mathsf{E}[\psi'] \\
&= I_{p\times p}
\end{aligned}$$

as $\mathsf{E}[\psi(\theta_{\mathbf{G}})] = 0$.

For n observations, our theoretical class representative will be $I_{\mathbf{G}}^{-1}(\theta)\bar{\psi}_n(\theta)$, which at $\theta_{\mathbf{G}}$ is the asymptotic estimation error of $\hat{\theta}_n$ for $\theta_{\mathbf{G}}$ per (1.14) and in practice we will estimate the representative at $\boldsymbol{\theta}_{\mathbf{G}}$ by

$$\hat{I}_n^{-1}\bar{\psi}_n(\hat{\theta}_n).$$

1.7.2 Efficiency

One method for choosing between estimating functions might be by comparing their variances. However, within an equivalence class, variances could be arbitrarily large or small depending upon $D(\theta)$ so, using our class representative from (1.45) we have that at θ_{G} (omitted for brevity):

$$\begin{aligned} \mathsf{Var}\left[\mathsf{E}[\psi']^{-1}\psi\right] &= (\mathsf{E}[\psi'])^{-1}\mathsf{Var}[\psi](\mathsf{E}[\psi']^{T})^{-1} && (1.46)\\ &= \mathsf{E}[\psi']^{-1}\mathsf{Var}[\psi]\mathsf{E}[\psi']^{-1} \quad \text{by Assumption 11} && (1.47)\end{aligned}$$

which is the variance of our parameter estimates as in (1.16). We can estimate (1.46) by $\hat{I}_n^{-1}\hat{C}_n\hat{I}_n^{-1}$ which is our estimate of the variance of the parameter estimates as described in Section 1.4. Thus, by selecting an estimating function class representative with minimum variance we are selecting a representative whose parameter estimates also have minimum variance. Comparing these quantities for different choices of ψ enables us to compare estimating function variances between equivalence classes. If θ is scalar, we can use efficiency for a direct comparison. However, we need to extend that to multiple parameters.

We do that using a positive semidefinite matrix condition: if we have two estimating functions from different equivalence classes, $\boldsymbol{\psi}_1(\boldsymbol{\theta})$ and $\boldsymbol{\psi}_2(\boldsymbol{\theta})$, both of whose expected values have zeroes at $\boldsymbol{\theta}_{\mathsf{G}}$, such that:

$$\mathsf{Var}\left[(\mathsf{E}[\boldsymbol{\psi}_1'(\boldsymbol{\theta}_{\mathsf{G}})])^{-1}\boldsymbol{\psi}_1(\boldsymbol{\theta}_{\mathsf{G}})\right] \prec \mathsf{Var}\left[(\mathsf{E}[\boldsymbol{\psi}_2'(\boldsymbol{\theta}_{\mathsf{G}})])^{-1}\boldsymbol{\psi}_2(\boldsymbol{\theta}_{\mathsf{G}})\right]$$

where $\boldsymbol{C} \prec \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{C}$ is positive definite (similarly $\preceq$ denotes positive semidefiniteness) for matrices $\boldsymbol{B}$ and $\boldsymbol{C}$, then the smaller variance for $\boldsymbol{\psi}_1$ might lead us to prefer it over $\boldsymbol{\psi}_2$. In particular, if the difference between any two matrices is positive semidefinite, then any linear combination of parameters is estimated at least as precisely using $\boldsymbol{\psi}_1$ as it is under $\boldsymbol{\psi}_2$. We compare variances for individual parameters in Section 3.4.3. This positive semidefinite ordering is also known as Loewner ordering (see, for instance, Lindsay et al., 2011).

It has been shown (Chandrasekar and Kale, 1984; Joseph and Durairajan, 1991), subject to certain assumptions already made in this thesis, that in comparing estimating functions whose covariance matrices are positive definite the use of the Loewner ordering for comparing matrices (described as M-optimality) produces the same results as the use of either the trace (T-optimality), the determinant (D-optimality) or a quadratic loss function (Q$_\text{C}$-optimality where $\boldsymbol{\psi}^T\boldsymbol{C}\boldsymbol{\psi}$ is compared for some positive definite $\boldsymbol{C}$). As all but the Loewner ordering criterion are scalars then they can be used to calculate some measure of relative 'efficiency'. We explore M, T and D-optimality for a simulation in Section 4.4.

The variance comparison has a parallel in the notion of *Godambe Information* which was introduced in Godambe (1960) for scalar θ: further details are given in, for instance, Song (2007). For this we assume that **G** is parameterised by the objects of inference θ so we have $g(Y;\theta)$. We have seen in (1.2) that the score function for g, $U(\theta;Y)$, is an estimating function for all $\theta \in \Theta$. Define the variance of that score, the Fisher Information Matrix, as $i(\theta)$ (the Y being dropped for simplicity of notation). For any surrogate estimating function, $\psi(\theta;Y)$, the *Godambe Information Matrix* or *GIM* is defined as:

$$j_\psi(\theta) \equiv \mathsf{E}_{\mathbf{G}}[\psi'(\theta)]\mathsf{Var}_{\mathbf{G}}^{-1}[\psi(\theta)]\mathsf{E}_{\mathbf{G}}[\psi'(\theta)] \tag{1.48}$$

where, for obvious reasons, $\psi'(\theta)$ is defined as the sensitivity of $\boldsymbol{\psi}$. Assumptions are made consistent with those previously described. Then the Godambe inequality states that:

$$j_\psi(\theta) \preceq i(\theta) \tag{1.49}$$

with equality holding iff $\psi \sim U$. The proof is based around the Cauchy Schwartz inequality.

There is a danger with using information criteria, such as Godambe Information, for comparing objects in that there is no guarantee that any criterion is a good one (ie will give the ordering that one would 'like' in every circumstance), as acknowledged in Godambe (1960). In the case of Godambe Information, justification for (1.48) as an optimality criterion can be most easily seen with a scalar θ. In that case, for the estimating function, small variance ($\mathsf{Var}_{\mathbf{G}}[\psi(\theta)]$) and large sensitivity ($\mathsf{E}_{\mathbf{G}}[\psi'(\theta)]$), both of which could be seen as desirable at $\theta_{\mathbf{G}}$ as they define the function (and thus the parameters) more sharply, will increase the value of the information.

Now, (1.48) at $\theta_{\mathbf{G}}$ for n observations is proportional to the inverse of the asymptotic limit of the variance of $\hat{\theta}_n$ per (1.16). So, minimising our variance comparator is equivalent to maximising the Godambe information at $\theta_{\mathbf{G}}$ and surrogates can be compared to each other by relative efficiency or relative Godambe information as described earlier in this section, for instance using the determinant of one over the other. Note that the pth root of the determinant is often used (Davison, 2003, page 113) in order to keep the order of the efficiency correct with respect to n. The major difference between the variance and Godambe Information approaches is that in the former we have not assumed explicit knowledge of **G**, while in the latter that is not the case and there is consequently a bound for the information as described at (1.49).

1.7.3 Predictive ability

A common method for selecting a model is to use one of a number of information criteria, the most well known being the Akaike Information Criterion or AIC. Varin and Vidoni (2005) have proposed extending that to a particular form of surrogate, namely the composite surrogates (see Chapter 2) by adapting Takeuchi's Information Criteria or *TIC* (sometimes known as the *Network Information Criterion* or *NIC*, see Davison, 2003). However, their proof, based on that of Takeuchi, works equally well for any surrogate - the composite element is not critical. The overall idea, of both AIC and TIC, is to select models that best forecast a future random variable (Y_{n+1} say), where the judgment is made by minimising the Kullback-Leibler Divergence between **F** and **G** for the prediction. However, as in (1.5), the **G** term is a constant and so one can compare surrogates by maximising:

$$\ell_{\mathbf{F}}(\hat{\theta}_n) - \mathrm{tr}(\hat{C}_n \hat{I}_n^{-1})$$

where **F** ranges over the functions under consideration. The trace term is an approximation to the expected value of the likelihood ratio statistic under **G**. We are thus centring the values for comparison purposes (AIC penalises with a cruder but simpler factor, p).

Example I continued - Poisson Surrogate for Gaussian Data. TIC is:

$$\sum_{i=1}^{n} y_i \ln(\hat{\theta}_n) - n\hat{\theta}_n - \left(\frac{\sum_{i=1}^{n} y_i^2}{n\hat{\theta}_n} - \hat{\theta}_n \right).$$

□

One should take care before using TIC:

1. Burnham and Anderson (2002, page 65) point out that in fact AIC is a good approximation to TIC and, as a consequence, we might prefer to use that as calculation of the penalty term (namely p) is far more computationally efficient than the inversion and multiplication of potentially large matrices.

2. TIC involves the approximations $\hat{C}_n$ and $\hat{I}_n$ which can be slow to approach their asymptotic limit and again AIC may be more appropriate.

3. All the information criteria mentioned require the forecast future random variable to be independent of those for which we have observations. This may not always be the case (for instance in Example II, we have described short term temporal dependence and seasonality) in this thesis and so these criteria only apply when we have dealt with any dependence and are working with the resulting residuals.

A precursor to the currently used model selection criteria is given in Cox (1962) where a test statistic is derived for testing the null hypothesis that the density of the mechanism giving rise to the data belongs to a particular family of densities ($f_1(\theta)$) against the alternative hypothesis that it belongs to another separate family ($f_2(\omega)$). This is expanded upon slightly in Cox (2006, page 142) where a likelihood ratio statistic, $\ell_{f_1}(\hat{\theta}) - \ell_{f_2}(\hat{\omega})$, is suggested. Obviously, this can take positive or negative values and is asymptotically normal by a central limit theorem argument (given in Cox, 1962, for iid and dependent Y_i). While not bearing directly on our surrogate questions, this does suggest a way of distinguishing between potential **F**s. Along similar lines, but varying **G** rather than **F**, Foutz and Srivastava (1977) consider comparing the efficiency of the likelihood ratio test for data arising from various possible $\mathbf{G}_i$ by considering the ratio of the likelihood test statistics either exactly or, more likely, approximately in each case as $n \to \infty$. It assumes a known likelihood and also works with $\mathbf{G}_i$ parameterised, at least in part, by θ.

1.7.4 Robustness

The term robustness is widely used but, equally, has a wide range of definitions. It is used in at least the following senses for surrogate likelihoods or models:

Data robustness The effect of a small change in the data gives rise to only a small variation in $\hat{\theta}_n$. This is the traditional definition as discussed in, for instance, Maronna et al. (2006) and often involves the study of the effect of outliers. As, frequently, the data is the only knowledge we have of **G**, this definition is closely related to:

Model robustness We would like to work with a surrogate that fits well with a range of **G**s that might have generated the data under consideration. For instance, Copas and Eguchi (2010) propose a loglikelihood envelope for parameters of interest which is based upon the possibility that the data arises from a **G** contained in a tubular neighbourhood of radius ϵ around **F**. Models are then treated as equivalent to **F** if they satisfy the hypothesis that $\epsilon = 0$ at a particular acceptance level and this results in a loglikelihood with a plateau rather than a peak.

Likelihood robustness If we are testing a hypothesis for θ_* of dimension l, the likelihood ratio follows a χ^2_l distribution asymptotically (Kent, 1982). See the discussion at the end of Section 1.5 for a summary of when this might be the case for surrogates.

Information robustness $\hat{\theta}_n$ is consistent for $\theta_{\mathbf{G}}$, asymptotically normal and the information identity holds empirically (Royall, 1986). We investigate the effect of the last condition in Section 1.6.

Parameter robustness The parameters are invariant to certain forms of reparameterisation (Stafford, 1996). We discuss this with respect to adjustments in Sections 1.6.3 and 1.6.4.

Dimension robustness This can be applied to other forms of robustness and means that, for instance, a composite surrogate (as defined in Chapter 2) is robust to misspecification of the bivariate distributions as long as the univariate distributions are correctly specified (Kuk, 2007).

Distribution robustness Use of the sandwich estimator for the estimated parameter variance is described as robust (Chandler and Bate, 2007) compared to use of the *naive* inverse Fisher information. This is studied in more detail in Section 1.6.2.

1.8 Bayesian Approaches

Two Bayesian techniques have been developed which make use of surrogate distributions or likelihoods: variational Bayes and approximate Bayesian computation. Both are outlined here although neither will be explored further in this thesis.

Variational Bayes is analysed in more detail in, for instance, Beal (2003). It is used where one is taking a Bayesian approach to choose between models. Then one might be interested in the posterior probability of a model, m, given the data:

$$p(m|y) = \frac{p(m)p(y|m)}{p(y)}.$$

Maximising that expression over all models would lead one to a preferred model. To do that, amongst other things, one would need to evaluate:

$$p(y|m) = \int p(y|\theta, m)p(\theta|m)\mathrm{d}\theta$$

known as the *marginal likelihood*, where θ parameterises m. This is potentially complex. Equally intractable integrals may also arise in calculating predictive distributions or densities for latent variables. Rather than approximating using Monte Carlo methods, one approximates the integral using a density, f, that is integrable and which forms a bound for the target expression, and could thus be regarded as a form of surrogate. The idea is to minimise the 'distance' between the target density, g, and approximating expressions, where both are parameterised by θ. So, we seek to maximise:

$$\mathsf{E}_f \left[\ln \left(\frac{g(\theta; y)}{f(\theta)} \right) \right] = \int \ln \left(\frac{g(\theta; y)}{f(\theta)} \right) f(\theta)\, \mathrm{d}\theta$$

over f. This quantity arises in information theory where it is defined as *information* or in statistical physics where its negation is known as *free energy*, both referring to a departure from randomness or entropy, and that appears to be the justification for its more general variational Bayes usage. The usual approach (known as *mean field*) is to assume in the simplified f that all the parameters are independent, which generally permits the required integration. It is sometimes noted, for instance in Chappell et al. (2007), that the above quantity plus:

$$\mathsf{E}_f \left[\ln \left(\frac{f(\theta)}{g(\theta|y)} \right) \right] \tag{1.50}$$

is a constant, known as the expectation of the *log evidence*, $\mathsf{E}[\mathsf{P}(\theta)]$, and thus maximising the free energy is equivalent to minimising (1.50). However, that expression is not the KLD defined in (1.5) as expectations here are taken using the surrogate density, f, and not the target density, g, so variational Bayes is not a direct equivalent of our surrogate approach. Justification for using f is that it produces the best approximation to the true posterior (Chappell et al., 2007), and the approach is used as a computational aid.

Approximate Bayesian computation, or *ABC*, however, is a closer parallel to surrogate methodology, a good introduction being Kennedy and O'Hagan (2001), where it is described under the rubric of calibrating likelihood parameters in a Bayesian model, although more recently the technique has been applied to model selection (see, for instance Toni et al., 2009). It is used in the situation where data, y_g are derived from an unknown distribution with density, $g(y)$. One then works with a known density, $f(y;\theta)$, simulates data, y_f, from that and under a variant of rejection sampling (Lee, 2004, Section 9.5) accepts θ (drawn from a prior) if the distance between y_g and y_f is small enough, ie

$$d(y_g, y_f) \leq \delta$$

for some distance function, d, and scalar δ. A variation involves comparing summary statistics about the observed and simulated data rather than the data themselves. A further refinement known as *generalised ABC* involves assuming that the data is observed with a measurement error, ϵ, specified by some prior, π_ϵ and θ is accepted with probability proportional to $\pi_\epsilon(|y_g - y_f|)$. Having chosen our f, we then can influence the outcome by choosing the summary statistics (in a similar fashion to generalised method of moments) and then either d and δ, or π_ϵ.

In a similar fashion to composite surrogates (see Chapter 2), ABC is used because it gives usable results for complex problems but appears to be carried out heuristically, for instance the choice of distance function (eg Toni et al., 2009). The nature of the simulation means that it is only used for relatively simple models with a low number of parameters.

1.9 Summary

We have examined the use of surrogates using estimating functions, in situations where the distribution that generated the data, **G**, is not either known, tractable, computable, or available. Working with a detailed set of assumptions, we have derived the usual asymptotic results together with approximations to expected values, that can be used in practice. We have described the distribution of the standard test statistics for surrogates. We have then studied the options for adjusting surrogates to recover the information identity, a feature of **G**. As data may support a range of surrogates, we have reviewed how they may be compared, particularly for vector parameters. For completeness, we have looked at some Bayesian equivalents to surrogates.

We now move to studying a particular form of surrogate, the composite surrogate that has been widely used of late. We take forward all of the assumptions and many of the results that we have derived in this chapter, particularly those for asymptotics, adjusted surrogates and comparison of surrogates and, having applied them to composite surrogates, analyse their effectiveness in a simulation.

Chapter 2

Composite Surrogates

2.1 Introduction

A particular form of surrogate is the composite surrogate where the loglikelihood is the weighted sum of a number of marginal loglikelihoods, not necessarily marginal for the data generating mechanism (*DGM*). If the latter is the case, then the distribution for which the loglikelihoods are marginal will presumably have been chosen as a plausible parametric approximation for the DGM. In Section 2.2 we introduce the basic concepts together with a simple example. In Sections 2.3 and 2.4 we examine the bias and variance of estimators arising from composite surrogates. In the case of bias we introduce a new assumption around the compatibility of estimates arising from the marginal components, that continues to allow us to work with an unknown **G**. We explore the consequences of adding a constant of proportionality to create a true density arising from a composite surrogate in Section 2.5. In Section 2.6 we analyse how to deal with data elements with short term dependence so that they can be treated as independent. In Section 2.7 we carry out a simulation, examining the effect of adjusting the bivariate surrogate as described in Chapter 1 and examine the use of higher order asymptotics for testing small samples.

2.2 Basics

The use of a surrogate consisting of the product of low dimensional marginals of a complex distribution is of particular value in the study of clustered data, ie where each data element or observation is a vector of probably dependent datapoints, and is examined in this section. Good summaries of the area are Varin (2008); Varin et al. (2011), where the term *composite marginal likelihoods* is used. They also refer to a large number of applied papers in a variety of different fields, particularly genetics, that exploit the techniques outlined here.

In many situations the surrogate of choice arises from a complex parametric joint distribution, say **H**, which may not be the same as the potentially unknown **G**, and is not easily handled mathematically or computationally and which may not be robust to misspecification. In that case, one could try simplifying the procedure in order to make it more analytically tractable computable and / or robust. We define a *surrogate composite likelihood*, associated with a distribution **F**, to be a weighted product of lower dimensional likelihoods with which we can work more comfortably:

$$\mathsf{L}_{sc}(\theta; y) = \left(\prod_{C \in \mathcal{C}} \mathsf{L}_C(\theta; y)^{w_C} \right) \tag{2.1}$$

where $\mathcal{C}$ is a set, of dimension q, of subsets of the dataset indices $\{1, \dots, m\}$ and each L_C acts on the appropriate subset of dependent elements within an observation and, possibly, a subset of the parameters under consideration. Note the abuse of notation whereby we have previously used a subscript on the (log)likelihood to refer to the surrogate **F**: as we are certainly working with misspecified distributions here, the **F** is assumed and shall be dropped henceforth. Also, L_C is often parameterised by a subset of θ, θ_C say, but for ease of notation we shall continue to use θ except where use of the subset is specifically required.

The likelihood at (2.1) is easily set out but the related density may not be easily recovered so as to be well defined. We have therefore restricted ourselves to the likelihood, L_{sc}, as the constant of proportionality may not be known. This issue is explored further in

Section 2.5. Clearly, **F** is misspecified but we are allowing **H** to be so as well.

The resulting loglikelihood, for an observation (or vector of observations), y_i, is a weighted sum of *component* loglikelihoods:

$$\ell_{sc}(\theta; y_i) = \sum_{C \in \mathcal{C}} w_C \ell_C(\theta; y_i).$$

The loglikelihoods, ℓ_C are usually derived from low dimensional commonly used distributions, $\mathbf{F}_C$ which are frequently identical in form but act on different subsets of the cluster, and are marginal for **H**. We shall occasionally consider components that are derived from conditional marginal distributions but will be clear when that is the case. The composite loglikelihood is consequently analytically or computationally tractable and shares many of the features of **H**, which are explored in this chapter. These are the reasons for its use. The weights, w_C, and the vector of them, w, allow for the possibility that different subsets of the cluster might vary in importance, but are often all 1. These form the basis of many of the new contributions in this thesis and are studied in Chapters 3 and 4.

So, for instance, one might prefer to use a high (m) dimensional multivariate normal surrogate but deem the loglikelihood too complex to deal with. We could then define ℓ_{sc} to be the product of $q = m(m-1)/2$ bivariate normal likelihoods (one for each pair of elements in y) with weights, w_C all equal to 1. That would retain all the parameters from the high dimensional surrogate but be more manageable and is examined in more detail in Chapter 4. A different and even simpler model might assume an unknown correlation parameter that is constant across all pairs of variables.

Example IV - Bivariate Normal Composite Surrogate. For random variables $(Y_1, \ldots, Y_m)$, where we are interested in studying a common correlation between them, consider taking as a surrogate a weighted composite loglikelihood consisting of the sum of the possible standard bivariate normal loglikelihoods with a common ρ, with, for instance, weight w_{12} being that for the bivariate distribution for y_1 and

y_2. Then for $m = 3$

$$\begin{aligned}
\mathrm{L}_{sc}(\rho) &= \frac{1}{(1-\rho^2)^{w_{12}/2}} \exp\left(-\frac{w_{12}}{2(1-\rho^2)}(y_1^2 - 2\rho y_1 y_2 + y_2^2)\right) \\
&\cdot \frac{1}{(1-\rho^2)^{w_{13}/2}} \exp\left(-\frac{w_{13}}{2(1-\rho^2)}(y_1^2 - 2\rho y_1 y_3 + y_3^2)\right) \\
&\cdot \frac{1}{(1-\rho^2)^{w_{23}/2}} \exp\left(-\frac{w_{23}}{2(1-\rho^2)}(y_2^2 - 2\rho y_2 y_3 + y_3^2)\right) \\
&= \frac{\exp\left(\frac{-(y_1^2(w_{12}+w_{13})+y_2^2(w_{12}+w_{23})+y_3^2(w_{13}+w_{23})-2\rho(y_1y_2w_{12}+y_1y_3w_{13}+y_2y_3w_{23}))}{2(1-\rho^2)}\right)}{(1-\rho^2)^{3/2}} \\
&= \frac{1}{(1-\rho^2)^{3/2}} \exp\left(-\frac{-y^T\boldsymbol{Q}y}{2}\right),
\end{aligned}$$

incorporating a quadratic form where:

$$\boldsymbol{Q} = \frac{1}{1-\rho^2}\begin{pmatrix} w_{12}+w_{13} & -\rho w_{12} & -\rho w_{13} \\ -\rho w_{12} & w_{12}+w_{23} & -\rho w_{23} \\ -\rho w_{13} & -\rho w_{23} & w_{13}+w_{23} \end{pmatrix}.$$

We can see that the likelihood $\mathrm{L}_{sc}(\rho)$ can be considered as corresponding to the likelihood from a trivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{Q}^{-1}$. This result is generalised and analysed further in Chapter 4. Theorem 4.3.1, which generalises this example, sets out conditions on the parameters for the distributions described here to be genuine.

This can be extended relatively simply to composite surrogates where m takes any integer value of at least 2 so that:

$$\mathrm{L}_{sc}(\rho) = \frac{1}{(1-\rho^2)^{q/2}} \exp\left(-\frac{-y^T\boldsymbol{Q}y}{2}\right)$$

where

$$\boldsymbol{Q} = \frac{1}{1-\rho^2}\begin{pmatrix} \sum_{i=1}^m w_{i1} - w_{11} & -\rho w_{12} & \dots & -\rho w_{1m} \\ -\rho w_{12} & \sum_{i=1}^m w_{i2} - w_{22} & \dots & -\rho w_{2m} \\ \dots & \dots & \dots & \dots \\ -\rho w_{1m} & -\rho w_{2m} & \dots & \sum_{i=1}^m w_{im} - w_{mm} \end{pmatrix}$$

and $w_{ij} = w_{ji}$. □

The $\mathbf{F}_C$s could differ in form. For instance, they might represent all the univariate and bivariate margins of **H**. This more general situation is considered in Cox and Reid (2004) where the conditional composite surrogate introduced in Besag (1974) for dealing with spatial data, probably the earliest example of the composite approach, is treated as a subcase. We study this case for a particular example in Section 4.7.

In practice the use of bivariate composite components is becoming more widespread. Recent examples that belong to categories we have already considered include

Example II continued - Weather Readings. Padoan et al. (2010) apply bivariate composite techniques to the modelling of spatial extremes using max-stable processes. The theory is applied to rainfall readings in part of the United States. The results allow flexible models, show good estimate behaviour compared with traditional approaches and are computationally inexpensive. □

Example III continued - Longitudinal Study. Vasdekis et al. (2012) use bivariate composite likelihood estimation to study ordinal longitudinal responses. Time dependent latent variables and random effects are considered. The promising techniques are applied to extracts from the British Household Panel Survey. □

2.3 Bias

Clearly, it would be ideal if any estimates arising from using a composite loglikelihood are unbiased compared to those which would arise from **H**, or **G** if known. That is the subject of this section. We can differentiate each component loglikelihood, $\ell_C(\theta; y)$, and from the results for general surrogates in Section 1.2, treat the result as an estimating function, $\psi_C(\theta; y)$, with a zero under **G** at $\theta_{\mathbf{G}_C}$ say, so that

$$\mathsf{E}_{\mathbf{G}}[\psi_C(\theta_{\mathbf{G}_C}; y)] \quad = \quad 0. \tag{2.2}$$

In practice, we will generally be using the data as the source of our knowledge of **G**. We make the following assumption to ensure unbiasedness as described above. Most papers, for instance as summarised in Varin (2008), treat **G** = **H** as known and so, as we shall see later in this section, the assumption is not required. It is stronger than it need be for identifiability but is convenient for Section 2.5.

Assumption 17. *The* $\boldsymbol{\theta}_{\boldsymbol{G}_C}$*, which are the zeroes of* $E_{\boldsymbol{G}}[\boldsymbol{\psi}_C(\boldsymbol{\theta})]$, $C \in \mathcal{C}$ *are mutually compatible over all C Ie, where any element, say* θ_t*, of* θ *appears in more than one component loglikelihood, with indices* l_1 *and* l_2 *say, the values of* θ_t *for which* $E_{\boldsymbol{G}}[\boldsymbol{\psi}_{l_1}(\theta)] = E_{\boldsymbol{G}}[\boldsymbol{\psi}_{l_2}(\theta)] = 0$ *are the same in each case.* [1].

If we do not make this assumption, then different components will give rise to parameter estimates that may not be consistent for the same $\theta_{\mathbf{G}}$. In that case, the introduction of weights may mean that our overall parameter estimates have a different limit from the unweighted ones.

Example II continued - Weather Readings. For instance, if one decides that standard bivariate normal components with common correlation, ρ, representing all the possible pairs of locations should be used in the composite loglikelihood, then as there is in practice likely to be distance based correlation, the estimates for ρ from each component may vary considerably and the overall composite surrogate estimate may not be susceptible to appropriate interpretation. □

Thus, Assumption 17 is partially about model choice - one needs to think about the situation being analysed before deciding upon an appropriate model and if that has happened, then the assumption may prove to be unnecessary.

We then define our composite estimating function:

$$\psi_{sc}(\theta : y) \equiv \sum_{\mathcal{C} \in \mathcal{C}} w_{\mathcal{C}} \psi_{\mathcal{C}}(\theta ; y). \tag{2.3}$$

[1] An equivalent condition that will lead to a compatible set of zeroes is that up to pq equations in p variables, the component estimating equations, would have to be solved consistently. For instance, if our composite surrogate arises from q bivariate normal likelihoods with a common correlation parameter, ρ, over m-dimensional Y (so that $q = m(m-1)/2$), then our $p = 2m+1$ dimensional $\theta_{\mathbf{G}}$ (m means and variances, μ_i, σ^2 and ρ) will arise from solving $(m-1)(2+m/2)$ equations consistently ($m-1$ for each μ_i and σ_i^2 and $m(m-1)/2$ for ρ).

We define $\theta_{\mathbf{G}}$ to be the compatible values from Assumption 17, so that by (2.2)

$$\mathsf{E}_{\mathbf{G}}[\psi_{sc}(\theta_{\mathbf{G}}; Y)] = 0 \tag{2.4}$$

and ψ_{sc} accords with our definition of an estimating function given at (1.1). Clearly, the choice of weights will have no effect on 2.4.

We might describe this as a *bottom up* approach to building our ψ_{sc} - we construct it from lower dimensional elements with compatible zeroes. An alternative is a *top down* line of attack. In that case, where we begin with ψ_{sc}, elements of the overall zero, $\theta_{\mathbf{G}}$, would then need to be a compatible set of zeroes of the component estimating functions. In that case we would need to make assumptions about the nature of the families the component likelihoods arise from in comparison to the family containing **G** (for instance, the f_C which give rise to the ℓ_C are marginal densities arising from a single joint density g arising from **G**). Since, we are assuming an unknown **G**, that would be awkward. With our approach, to assess compatibility of estimates we either approximate E[] from the data or make assumptions about the goodness of fit of **H** for **G**.

The known **G** approach is also standard in *inference from the margins* (see, for instance Joe, 1997) wherein parameters are estimated dimensionwise: univariate from a univariate composite surrogate, then bivariate from a bivariate composite surrogate using the univariate parameter values just calculated (via a profile loglikelihood approach), etc. Our approach still allows the composite and choice surrogates to be compared, but in the light of **G**, represented by the data.

In a similar fashion we define our estimating equation, $\psi_{scn}(\theta; y_1, \ldots, y_n)$, for n data points:

$$\psi_{scn}(\theta; y_1, \ldots, y_n) \equiv \sum_{C \in \mathcal{C}} w_C \psi_C(\theta; y_1, \ldots, y_n) \tag{2.5}$$

with a zero at $\hat{\theta}_n$. Where the Y_i are iid:

$$\begin{aligned} \psi_{scn}(\hat{\theta}_n) &\equiv \sum_{C \in \mathcal{C}} w_C \sum_{i=1}^{n} \psi_C(\hat{\theta}_n; y_i) \\ &= 0. \end{aligned} \tag{2.6}$$

If we do not make Assumption 17 then one would prefer the expected values of parameter estimators arising from the use of a surrogate composite ($\hat{\theta}_n$) to match those arising from the corresponding preferred complex multivariate surrogate, **H**, ($\hat{\theta}_{\mathbf{H}}$). Unfortunately, this is not always the case. Mardia et al. (2009) show that this issue is complex and describe two cases in terms of the estimators

1. **H** belongs to a canonical exponential family for y with sufficient statistic $t(y)$ that is closed. Closure is defined so that if y is not scalar, then for any subvector y_B of $\boldsymbol{y}$, its distribution is also a member of a canonical exponential family with sufficient statistic $t_B(y_B)$ a subvector of $t(y)$. The requirement demands that the individual elements of y are not too closely intertwined. Then, $\hat{\theta}_n$ is unique and $\hat{\theta}_n = \hat{\theta}_{\mathbf{H}}$ so that the estimator is unbiased only if:

   (a) Each of the $\mathbf{F}_C$s includes all of the elements of $\boldsymbol{y}$ (for instance, each is the distribution of a single element of y, conditional upon all the other elements, such as $\mathbf{F}_1(y_1|y_2, \ldots, y_n)$) and each element of $t(y)$ is excluded from at least one of the sets of sufficient statistics of the components of the composite surrogate.

   (b) The composite is either a product of all the marginal pairs or of all the conditional pairs and for which all the elements of the sufficient statistic contain at most 2 elements of y. For instance, the multivariate normal distribution $MVN_p(\mathbf{0}, \boldsymbol{\Sigma})$ has sufficient statistics based around the sample covariance matrix, any marginal or conditional density is also normal and so fulfils the criteria.

2. For all models, including those that are not closed, with the usual regularity conditions (specified here in Section 1.2) and parameter identifiability, $\hat{\theta}_n \rightarrow \hat{\theta}_{\mathbf{H}}$ as

$n \to \infty$. Identifiability requires that the composite surrogate contains all the information about θ. For instance, if θ includes parameters representing interactions between more than two elements of Y, then for the bivariate composite surrogate, $\mathsf{E}[\hat{\theta}_n]$ may not tend to $\mathsf{E}[\hat{\theta}_{\mathbf{H}}]$.

We would expect this Case 2 to apply to the models under consideration in this thesis.

Note that expectations in (2.4) are being taken under the unknown **G** and that the surrogate of choice, some high dimensional multivariate likelihood, **H**, is not mentioned. If **G** is known, then each of the composite elements arises as a marginal distribution of **G** and so when we solve the estimating equations for each marginal, the solutions for the appropriate subsets of θ will be the same as if we had solved estimating equations for **G** directly, subject to the conditions of the preceding paragraph, and are thus the same across all sets of equations, satisfying Assumption 17. Combining these in the composite estimating equations will have the same effect. This is often given (Varin, 2008) as a justification for a composite approach but our more general approach - an unknown data generating mechanism, **G**, with a known preferred distribution, **H**, but using the Kullback-Leibler Divergence between **G** and **F** as a justification - gives similar outcomes. This approach is also taken in Kent (1982) and Xu and Reid (2011).

2.4 Covariance Matrix Estimation

Having examined the bias of a composite likelihood estimator, we now investigate its covariance matrix. This will be needed for hypothesis testing as described in Section 1.5 and will also help us assess whether the composite surrogate approach is useful for any particular set of data. If elements of the covariance matrix are extremely large then another approach might be appropriate.

The normalised asymptotic distribution of $\hat{\theta}_n$ has covariance matrix, from (1.16):

$$I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}})C_{\mathbf{G}}(\theta_{\mathbf{G}})I_{\mathbf{G}}^{-1}(\theta_{\mathbf{G}})$$

Clearly, as **G** is unknown, this will need to be estimated and a number of techniques have been developed.

The first is to extend the estimation of individual terms as in Section 1.4. Composite surrogates add a further level of complexity to that estimation:

$I_{\mathbf{G}}(\theta_{\mathbf{G}})$. We estimate this generally by $-A_n\psi_n'(\hat{\theta}_n)$ per Section 1.4. For composite surrogates, we can expand the expression using (2.5) to give

$$-A_n\left(\sum_{C\in\mathcal{C}} w_C\psi_C'(\hat{\theta}_n;y_1,\ldots,y_n)\right),$$

which is easily calculated;

$C_{\mathbf{G}}(\theta_{\mathbf{G}})$. Our usual approximating function, per Section 1.4 is $B_n\hat{V}_n(\hat{\theta}_n)B_n$. In the iid case we have from (1.22):

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{sc}(\hat{\theta}_n;y_i)\psi_{sc}(\hat{\theta}_n;y_i)^T.$$

Substituting the definition of $\psi_{sc}(\hat{\theta}_n;y_i)$ from (2.3):

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{C\in\mathcal{C}} w_C\psi_C(\hat{\theta}_n;y_i)\right)\left(\sum_{C\in\mathcal{C}} w_C\psi_C(\hat{\theta}_n;y_i)\right)^T \tag{2.7}$$

where the component estimating functions, $\psi_C(\hat{\theta}_n;y_i)$, may well be correlated with each other (unlike in the partial likelihood (Cox, 2006), an early version of the composite approach) and so (2.7) cannot be reduced to a sum of $\psi_C^2(\hat{\theta}_n;y_i)$ terms. An example of how one might deal with uncorrelated components in certain circumstances is given in Section 2.6.

As set out in Section 1.4, we estimate $I_{\mathbf{G}}(\theta_{\mathbf{G}})$ and $C_{\mathbf{G}}(\theta_{\mathbf{G}})$ by $\hat{I}_n$ and $\hat{C}_n$ respectively, combining them to give the variance of the parameter estimator, $\hat{I}_n^{-1}\hat{C}_n\hat{I}_n^{-1}$.

A second technique for estimating the variance of the parameter estimator, outlined in Joe (1997), is to use the jackknife. In this the y_i are omitted in turn (for larger samples one could omit larger groups of data), leading to parameter estimates, $\hat{\theta}_n^{(i)}$, $i=1,\ldots,n$, and the covariance estimator is:

$$\sum_{i=1}^{n}(\hat{\theta}_n^{(i)} - \hat{\theta}_n)^T(\hat{\theta}_n^{(i)} - \hat{\theta}_n)$$

using arguments similar in form to those for the asymptotic distribution of $B_n A_n^{-1}(\hat{\theta}_n - \theta_{\mathbf{G}})$.

A third technique, with similarities to both the first two, applies when the data can be considered as ordered in some sense, perhaps by time and / or space. This approach to estimation of $C_{\mathbf{G}}$ is called window subsampling (Heagerty and Lumley, 2000) and involves splitting the range of the order into r overlapping subranges, R_i and our estimator is:

$$\frac{1}{r}\sum_{i=1}^{r} r_i \psi_{scr_i}(\hat{\theta}_n; y \in R_i)\psi_{scr_i}(\hat{\theta}_n; y \in R_i)^T$$

where r_i denotes the size of the subrange.

Both the latter two methods are working with subsets of the data that are treated as independent: as long as the data are well mixed, estimating an expected value by these methods appears to be effective.

In the context of Generalized Estimating Equations (*GEE*), Lu et al. (2007) examine two alternatives to the sandwich estimator (1.16), for small samples involving bias correction. Crowder (2001) suggests that as an alternative to GEE estimation for longitudinal studies, the covariance matrix for the parameters that are coefficients of covariates could be estimated by adjusting that arising from Gaussian estimation. This has potential in examples such as that studied in Section 2.7.

2.5 Constant of Proportionality

Composite and adjusted composite surrogate loglikelihoods do not necessarily arise from a well defined density. If we wish to make use of such a density, as we do in Section 3.4 to derive a set of optimal weights, then we need to calculate a constant of proportionality

(*CoP*), K^{-1}, so that in the composite case:

$$\frac{1}{K}\int_y \exp(\ell_{sc}(\theta; y))\mathrm{d}y = 1,$$

ie:

$$\begin{aligned} K &= \int_y \exp(\ell_c(\theta; y))\mathrm{d}y \\ &= \int_y \prod_{C \in \mathcal{C}} f_C(y_C; \theta)^{w_j}\mathrm{d}y \end{aligned} \tag{2.8}$$

where y_C represents the subset of the vector y that appears in component *C*.

Assumption 18. *The constant of proportionality is finite.*

While this need not be the case, one would hope that the choice of **H** and **F** would make it so. For the multivariate normal case, a more specific assumption is given as part of the main result, Theorem 4.3.1.

K does not depend upon the data, they have been integrated out, but may be a function of the parameters and the weights. We describe the resulting distribution as $\mathbf{F}_K$ with density f_K, loglikelihood ℓ_K and estimating function ψ_K. We retain, by an abuse of notation, **F** to refer to the composite surrogate and ℓ_{sc} to its loglikelihood. Clearly, we can work with ℓ_K to derive parameter estimators etc as we have done with ℓ_{sc}. However, there is no guarantee that the resulting parameter estimators, $\hat{\theta}_{n_K}$ and $\hat{\theta}_n$ respectively, would tend to the same limits, $\theta_{\mathbf{G}_K}$ and $\theta_{\mathbf{G}}$ (remember that we have built composite surrogates from the ground up rather than treating them as marginal for the distribution that generated the data). In fact, for $\theta_{\mathbf{G}_K}$ and $\theta_{\mathbf{G}}$ to be equal we would need for non zero K:

$$\begin{aligned}
0 &= \mathsf{E}[\psi_K(\theta_{\mathbf{G}_K})] \\
&= \mathsf{E}_{\mathbf{G}}\left[-\frac{\mathrm{d}\ln(K)}{\mathrm{d}\theta}\right]\bigg|_{\theta=\theta_{\mathbf{G}}} + \mathsf{E}[\psi(\theta_{\mathbf{G}_K})] \\
&= \mathsf{E}_{\mathbf{G}}\left[-\frac{\mathrm{d}\ln(K)}{\mathrm{d}\theta}\right]\bigg|_{\theta=\theta_{\mathbf{G}}} + \mathsf{E}[\psi(\theta_{\mathbf{G}})] \\
&= \mathsf{E}_{\mathbf{G}}\left[-\frac{\mathrm{d}\ln(K)}{\mathrm{d}\theta}\right]\bigg|_{\theta=\theta_{\mathbf{G}}} \\
&= \frac{\mathrm{d}\ln(K)}{\mathrm{d}\theta}\bigg|_{\theta=\theta_{\mathbf{G}}} \qquad \text{as } K \text{ is non random and by Assumption 2}
\end{aligned}$$

or

$$0 = \frac{\mathrm{d}K}{\mathrm{d}\theta}\bigg|_{\theta=\theta_{\mathbf{G}}} \tag{2.9}$$

so that either K is not dependent upon θ or its derivative has a factor of $\theta - \theta_{\mathbf{G}}$. Thus, although one might use $\mathbf{F}_K$ for calculating weights, as in Section 3.4, care should be taken about using it for estimating parameters.

The surrogate composite loglikelihood with CoP could also be viewed as a standard composite loglikelihood where one of the components (ie $-\ln(K)$), rather than being a marginal density, is just a function of the parameters and weights. For that to fit within our definition of a composite surrogate and thus be used for parameter estimation, the extra term would have to satisfy Assumption 17, which, as we have seen, would require (2.9) to hold, which is often not the case.

The assumptions we have made in this thesis to derive the asymptotic results in Section 1.3 have been made with reference to estimating functions in order to set out the theory as generally as possible. However, in common with most papers (eg Varin (2008)) we have described composite surrogates in terms of loglikelihoods and then derived estimating functions. It is worth noting here that if the asymptotic theory based around surrogate densities (as opposed to surrogate estimating functions) is applied to the composite case, Xu and Reid (2011) show that only minor adjustments to the equivalents of our assumptions are required to cope with the fact that composite surrogates without a CoP are not genuine distributions.

Cox and Reid (2004) explore analytically the relative efficiency of a generalisation of Example IV, as set out in Section 2.2, to any dimension of Y. They examine the situation where **G** is standard multivariate normal with exchangeable correlation and the composite surrogate loglikelihood is the sum of all the bivariate marginals. It shows that as the length of the multivariate random variable, m, increases, the efficiency of the bivariate composite surrogate, compared with that of the equivalent full multivariate normal distribution, decreases, albeit fairly slowly.

Example IV continued - Bivariate Normal Composite Surrogate. Adding in the CoP to the composite surrogate affects both the mean (as shown at (2.9)) and variance (as shown by Cox and Reid (2004)) of the estimator. We ran 1000 simulations each of 1000 datapoints generated from a standard multivariate normal distribution with common correlation $\rho = 0.5$ for a range of lengths of random variable from 3 to 10 and examined the bias and efficiency of estimators arising from bivariate normal composite surrogates with and without CoP. The results are shown in Table 2.1. The bias of the bivariate normal composite surrogate with CoP illustrates the result at (2.9), ie it is significant since neither of the conditions for the estimate to be unbiased are met. There is no significant bias for the estimators arising from the multivariate normal (as expected) or the bivariate surrogate without CoP as Assumption 17 is satisfied - parameter estimates arising from marginal distributions of the multivariate normal distribution are compatible. The efficiency of the bivariate surrogate without CoP parameter estimator is consistent with the results in Cox and Reid (2004) while that for the bivariate surrogate with CoP is extremely poor. □

The bias and efficiency results from the example show bear out the main result of this section - one should not use a composite surrogate with CoP for parameter estimation. However, the CoP can be useful in deriving weights and this is explored in Section 3.4.

	Bias			Efficiency compared to MVN	
m	MVN	BVN no CoP	BVN CoP	BVN no CoP	BVN CoP
3	-0.00017	0.00002	-0.08522	99.6%	69.7%
5	0.00054	0.00111	-0.15459	96.9%	63.0%
8	-0.00007	-0.00007	-0.18939	89.2%	61.8%
10	0.00027	0.00037	-0.19864	84.5%	60.1%

Table 2.1: Comparison of mean and variance of parameter estimator for bivariate composite normal surrogates (BVN) with and without constant of proportionality (CoP), with multivariate normal distribution (MVN).

2.6 Ordered Dependence

In Example II in Section 1.2 we saw how one might need to condition upon other data elements in order to ensure that each of the Y_i are independent from each other. This situation was discussed in Chandler and Bate (2007). *Ordered dependence* sits between iid and unrestrictedly dependent Y_i and falls under the umbrella of partial likelihood outlined for instance, in Cox (2006, Section 7.6.5). In this situation, one ascribes some sort of order to the Y_i so that they are independent of each other, conditional upon a set, $\mathcal{D}_i$, consisting of any or all of the $Y_{i'}$ for $i' < i$. One then works with a univariate composite likelihood with the dependence conditioned out resulting in the standard surrogate asymptotic distribution and χ^2 test statistics (see Chapter 1). For instance, taking the conditioning into account,

$$\begin{aligned}
\ell_{scn}(\theta) &= \sum_{i=1}^{n} \sum_{C \in \mathcal{C}} w_C \ell_C(\theta; y_i | \mathcal{D}_i) \\
\psi_{scn}(\theta) &= \sum_{i=1}^{n} \sum_{C \in \mathcal{C}} w_C \psi_C(\theta; y_i | \mathcal{D}_i)
\end{aligned}$$

and one can apply the results outlined for composite surrogates in earlier sections of this chapter.

Chandler et al. (2007, pages 200-201) show that, if the univariate composite components are marginal for $\mathbf{G}$ (ie $\mathbf{H} = \mathbf{G}$), then $\mathsf{E}[\psi_C(\theta_{\mathbf{G}})]$, ie the expected value of the estimating function contribution from each cluster at $\theta_{\mathbf{G}}$, is zero. Adapting that proof, as suggested in the reference, we show that the estimating function contributions from different components are uncorrelated at $\theta_{\mathbf{G}}$ (see Appendix B) and as a result we can estimate $C_{\mathbf{G}}(\theta_{\mathbf{G}})$ with, adapting (2.7):

$$B_n^{\frac{1}{2}} \sum_{i=1}^{n} \sum_{C \in \mathcal{C}} \psi_C(\hat{\theta}_n; Y_i|\mathcal{D}_i) \psi_C(\hat{\theta}_n; Y_i|\mathcal{D}_i)^T B_n^{\frac{1}{2}}.$$

Example II continued - Weather Readings. Hourly weather readings will certainly have local time dependence. One could assess the extent of this (ie how many hours) in practice by treating recent readings as covariates and examining their significance. One could then condition upon that number of hours' readings and treat the resulting $Y_i|\mathcal{D}_i$ as independent, subject to any further dependencies such as seasonality. □

2.7 Simulation I

In order to compare the effectiveness of various composite surrogates, particularly those that have been adjusted horizontally and vertically to restore the information identity, (1.28), as described in Section 1.6, we use a simulation set out in Chandler and Bate (2007) (where the versions of the adjustments that we use here were described for composite surrogates, although, as we have seen, their applicability is more general). There, univariate and horizontally adjusted univariate models are compared. Here we add bivariate, horizontally and vertically adjusted bivariate, and the maximum likelihood estimator from the data generating mechanism from **G** (*MLE*) into the mix. The simulation and other calculations have been carried out in R (R Development Core Team, 2012) unless otherwise mentioned.

2.7.1 Composite Surrogates and Adjustments

The model used to generate the data has binary responses, together with covariates and a random effect. We define

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij\,1} + \beta_2 x_{ij\,2}.$$

where i represents a cluster or vector of observations and j a datapoint within that cluster and

- β_0 is taken to be 0.25.
- The length of each cluster is $1+Z$ where $Z \sim$ Poi(λ). One representation would be a longitudinal study where measurements for the ith patient (cluster) are taken at time 0, t_{i0}, and then at each subsequent visit, $t_{i1}, \ldots, t_{im_i}$, which occur with intervals determined by a Poisson process with arrival rate λ, up to time 1. Here, we have used $\lambda = 4$.
- The first covariate, $x_{ij\,1}$, consists of independent realisations of a Bernoulli random variable with mean $0.2 + 0.6t_{ij}$.
- The second covariate, $x_{ij\,2}$, is a linear trend, t_{ij}. There is thus dependence between the two covariates.
- The response variables, Y_{ij}, are taken from a Bernoulli distribution with mean p_{ij} where:

  $$p_{ij} \sim \text{Beta}\left(a, a\frac{1-\mu_{ij}}{\mu_{ij}}\right). \tag{2.10}$$

  so that we introduce a random effect. For each cluster we randomly select u_i from $U[0,1]$ and then $p_{ij} = F_{ij}^{-1}(u_i)$ where F_{ij} is the cumulative density function for the beta distribution just described. The random effect is thus common within each cluster, introducing intra cluster dependence. Here, we use $a = 0.1$ which means that the p_{ij} tend to take on values closer to 0 or 1 than the corresponding μ_{ij}.
- The probabilities for cluster i with l elements are given by:

  $$\pi(y_i = (y_{i1}, \ldots, y_{il})) = \prod_{j=1}^{l} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}.$$

Inference for the parameters $(\beta_0, \beta_1, \beta_2, a)$ was carried out, for data generated from 25 values of $\beta_1 = \beta_2$ regularly in $[-0.6, 0.6]$, for 1000 simulations each of 30 clusters, by maximising loglikelihoods. The results were compared for likelihood ratio tests, testing $H_0 : \beta_1 = \beta_2 = 0$, using power curves at the 5% level for the following:

Univariate For reference purposes, as previously studied in Chandler (2004), entries in a cluster were treated as if they were independent and the loglikelihood was the sum of the loglikelihood for each element of a component. The likelihood ratio test statistic has an asymptotic distribution which is a weighted sum of χ^2_1 distributions as described in Section 1.5. Varin (2008) proposes using the Satterthwaite approximation for this weighted sum while Chandler and Bate (2007) use an approximation described in Bowman and Azzalini (1997), a practice that we have followed in this thesis.

Vertically Adjusted Univariate Chandler and Bate (2007)'s version of the vertical adjustment described in Section 1.6.4 is applied to the univariate likelihood. This results in a χ^2_l distribution for the test statistic where l represents the number of covariate coefficients hypothesised as 0. The horizontally adjusted univariate model is not shown here but the results are very similar to those from the vertical adjustment.

Bivariate Unadjusted A bivariate composite loglikelihood for cluster i, containing m_i elements, would be a sum of bivariate Bernoulli loglikelihoods

$$\begin{aligned} \ell_i &= \sum_{j \neq k} (y_{ij} \ln p_{ij} + (1 - y_{ij}) \ln(1 - p_{ij}) + y_{ik} \ln p_{ik} + (1 - y_{ik}) \ln(1 - p_{ik})) \\ & \quad 1 \leq j, k \leq m_i \end{aligned} \tag{2.11}$$

where each pair has a common random effect, $u_{i,jk}$, to generate the p_{ij} and p_{ik}. Ideally, one would like to integrate out the random effect. This would require knowing the explicit form of F^{-1}_{ij}, which is not possible. However, with $a = 0.1$ and μ_{ij} varying between 0.26 and 0.82 (which is the case here), $p_{ij} = F^{-1}_{ij}(u_{i,jk})$ is sigmoid and can be approximated by the function

$$1/(1 + \exp(-a_n(u_{i,jk} - 1 + \mu_{ij}))), \tag{2.12}$$

where a_n is a nuisance parameter in place of a, a shifted and scaled inverse logistic function, which retains the dependence upon μ_{ij}. The random effect could be integrated out from the resulting loglikelihood, (2.11), and details of the integration are given in Appendix C. If a cluster had only one member, an integrated out univariate likelihood was taken, again summarised in Appendix C. The resulting test statistic was compared to a weighted sum of χ_1^2 distributions.

Bivariate Adjusted The loglikelihood in the previous item is adjusted horizontally (Section 1.6.2) and vertically (Section 1.6.3) resulting in χ_l^2 distributions for the test statistics.

MLE We would like to carry out maximum likelihood estimation for the model that generated the data. However, the problem of integrating out the random effect, as described in the Bivariate Unadjusted point, was present for the distribution that generated the data but each random effect was common across the whole of a cluster. The technique of analytically integrating out that random effect from a sigmoid approximation, (2.12), was not practical for clusters of size greater than three due to algebraic complexity. Therefore, a numerical approach was taken and the loglikelihood of the data in each simulation was maximised with the random effects being integrated out numerically. It did prove possible to do the same with the original data generating mechanism (DGM), but each simulation then took over two hours to run. For 10 simulations, the numerical approach was compared for the DGM and the sigmoid approximation. The resulting differences in p-values and parameter estimates were transformed to be approximately normal and t-tests in both cases gave p-values over 0.4 for the hypotheses that there were no differences between the two approaches, ie the use of the sigmoid function, (2.12), is justified as an approximation to the inverse beta distribution.

In order to calculate the adjustments or the eigenvalues for the weighted χ_1^2 distributions, $\hat{I}_n$ and $\hat{C}_n$ were required, as described in Section 1.5. The former was taken as the Hessian matrix in the nonlinear minimisation routines used in R (R Development Core Team, 2012). The latter was calculated as the variance of the estimating function for the data under consideration per (2.7). Differentiation was carried out numerically using Ridders' method (Lourmas and Chandler, 2006).

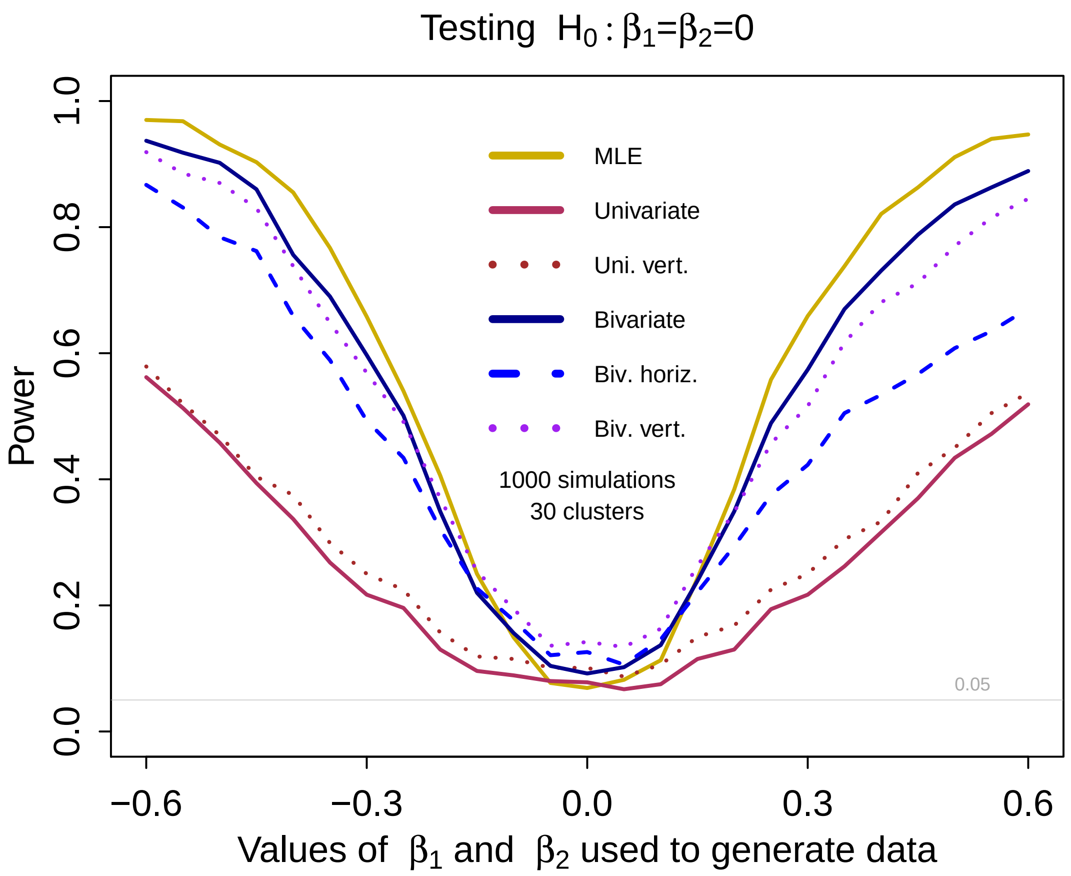

Figure 2.1: Simulation I. Power plots at level 0.05 for MLE and range of surrogates.

The resulting power plot is shown in Figure 2.1 and the following are of note:

1. Rejection rates for data generated at $\beta_1 = \beta_2 = 0$ are given in Table 2.2. As the null hypothesis uses the same values, one would expect the rejection rates to be 0.05. Chandler and Bate (2007) suggest that for small numbers of clusters, the covariance estimator can be inefficient leading to slightly liberal tests. This is borne out for a similar example in Section 3.4.5. However, the rates for the three adjusted surrogates are more than double what they should be. We review higher order features in Section 2.7.2.

2. As one might expect, the differences between the levels of complexity in the surrogates are reflected in the power curves: MLE is more powerful than the bivariate surrogates which are more powerful than the independence ones.

Univariate	0.078
Univariate Vertically Adjusted	0.101
Bivariate	0.092
Bivariate Horizontally Adjusted	0.126
Bivariate Vertically Adjusted	0.142
MLE	0.069

Table 2.2: Simulation I. Rejection rates for each of five tests at level 0.05 with data generated from $\beta_1 = \beta_2 = 0$.

3. What is surprising is that the effect of adjusting the bivariate surrogate reduces the power of the test. This is consistent with results in Padoan et al. (2010) and Pace et al. (2011). It leads us to ask whether there are other methods that could improve the power. One possibility is to vary the weights attached to each component; this is investigated in Chapters 3 and 4. The most likely explanation for the reduction in power is based around the Bartlett identities. We have only restored the first two of them but they exist for all moments and those of order three or more may have significant impact in this example. Note that the Bartlett Correction discussed in the following Section is concerned with higher order approximations for test statistics, not for Bartlett identities.

2.7.2 Bartlett Correction

We saw in Simulation I in Section 2.7.1 that the rejection levels for data generated from $\beta_1 = \beta_2 = 0$ are higher than the expected 0.05, particularly for adjusted surrogates. One approach for smaller sized samples (we have been using 30 clusters per simulation) is to adjust the likelihood ratio statistic, W, to take into account higher order features of the data than the two so far considered, per Appendix A. The standard approach to this is to use the *Bartlett correction* or adjustment (details from McCullagh and Nelder (1989) and Barndorff-Nielsen and Cox (1994)) which reduces the relative error from $O(n^{-1})$ to $O(n^{-2})$. This result only holds where the distribution of the likelihood ratio statistic, W, is asymptotically χ^2_l, some l, and so we can only use it for our adjusted surrogates, not the unadjusted ones.

The idea is to multiply W by $l/\mathsf{E}_0[W]$, to give W_B, where $\mathsf{E}_0[W]$ is the expected value of W under the null hypothesis. The new statistic has the same mean as W and faster convergence to χ^2_l throughout the distribution.

The usual approach is to approximate the correction by a complex expression involving multiple terms and products of high order cumulants. Deriving these analytically for our simulation is analytically daunting and possibly impossible. However, there are alternatives:

1. McCullagh and Nelder (1989, 15.3.2) derive a version of the correction, for Generalized Linear Models (*GLM*s), that mostly involves matrices arising in the standard theory of GLMs. This simplifies the calculations considerably. While our simulation do not involve distributions of response variables from the exponential family, they are similar and one could adapt the theory to those cases.

2. One could estimate the values of the required cumulants from a large data set.

3. Finally, we could estimate $\mathsf{E}_0[W]$ directly by taking a large set of simulated datapoints and calculate the mean of the likelihood ratios generated for the adjusted surrogates under consideration.

The last two alternatives can only be used where one has the ability to simulate from **G**, which is not generally possible in real applications. However, in order to understand whether there is any value in using the Bartlett correction for adjusted surrogates we will use the final approach in this case.

We apply the adjustment to the three adjusted models: univariate vertical, bivariate vertical and bivariate horizontal. The resulting power curves are shown in Figure 2.2 and the rejection levels at $\beta_1 = \beta_2 = 0$, compared to those without the Bartlett correction are given in Table 2.3.

The overall effect is to shift the adjusted surrogate power curves slightly downwards. This means that the rejection rate levels are brought much closer to the anticipated 0.05 although for the bivariate adjusted models they have overshot somewhat and are

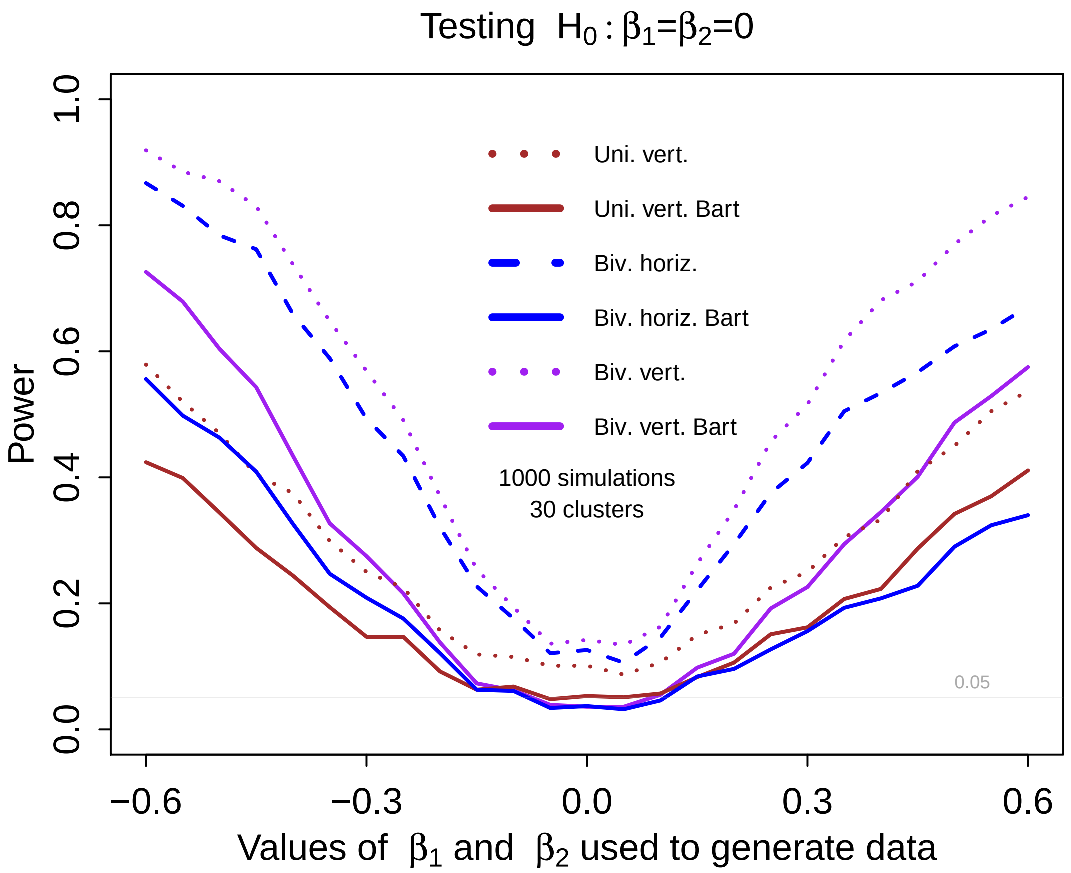

Figure 2.2: Simulation I. Power plots at level 0.05 for adjusted surrogates with and without Bartlett correction.

between 0.036 and 0.037. However, the effect on the poor power performance of the bivariate adjusted surrogates is to make it marginally worse.

A Bartlett type correction applied to the unadjusted surrogates would presumably have a similar (although probably less marked as the rejection rates are not so poor) effect. While not possible under the standard theory (Barndorff-Nielsen and Cox, 1994), such a correction has been suggested in Viraswami and Reid (1998b) but only for scalar θ. Extension to vector θ is discussed but not taken further as it is unlikely to be

Surrogate	Original	Bartlett Corrected
Univariate Vertically Adjusted	0.101	0.053
Bivariate Horizontally Adjusted	0.126	0.037
Bivariate Vertically Adjusted	0.142	0.036

Table 2.3: Simulation I. Rejection rates before and after applying a Bartlett correction for $H_0 : \beta_1 = \beta_2 = 0$, target 0.05

an improvement on the similar corrected score statistic derived in Viraswami and Reid (1998a). In both papers, it is required that $\theta_{\mathsf{G}} = \theta_0$, the latter being the the value of similarly defined parameters from the distribution that generated the data.

2.8 Summary

We have reviewed the theory behind composite surrogates, placing it in the context of the more general theory of surrogates described in Chapter 1. We have examined the bias and covariance matrix of the composite parameter estimators, features which will help us determine whether the use of the composite surrogate has value. We have seen in a simulation that composite surrogates, with and without the adjustments described in Section 1.6 are not always as powerful as maximum likelihood estimation based on the mechanism that generated the data and, indeed, the adjustments may reduce the power. We have analysed the effect of creating a well defined density related to the composite loglikelihood, ie one with a constant of proportionality. The use of this will be studied in Chapter 3 where we analyse the effect of weighting each component of a composite surrogate.

Chapter 3

Weights

3.1 Introduction

Having reviewed how one might adjust the basic composite surrogate to recover features of a preferred complex surrogate (such as the Information Identity), we now examine how, by weighting components of the composite surrogate, we might improve one of a number of desirable measures, such as efficiency. We begin, in Section 3.2, by surveying different approaches to weighting used in the literature. In Section 3.3 we demonstrate a more generally applicable version of one of those approaches, based on applying weights to estimating functions. We also show that a computationally cheaper scheme is optimal if dependence between estimating function components is not taken into account. In Section 3.4 we suggest a new scalar weighting scheme based on taking into account the constant of proportionality for the composite surrogate and minimising a Kullback-Leibler Divergence (KLD). We study situations in which this scheme does not give rise to unique weights. The effectiveness of the new scheme in practice is assessed through simulation.

3.2 General Approaches

3.2.1 Introduction

We recall that the standard definition of a composite surrogate loglikelihood (Section 2.2) is

$$\ell_{sc}(\theta; y) = \sum_{C \in \mathcal{C}} w_C \ln f_C(\theta : y_C) \tag{3.1}$$

where each of the q components, f_C, from **F**, our composite surrogate, acts on a subset of y, y_C. There is a weight, w_C, corresponding to each of the q components. While composite surrogates are widely used (see Varin (2008) for a summary), formal methods for choosing the weights have received comparatively little attention. Those approaches that do exploit weights cover a greater range of schemes than that given in (3.1) and they are described in the following subsections under the headings of scalar, component type, estimating function and cluster. In practice, papers that do not focus on weights (for instance Padoan et al. (2010)) will have tended to set most of the w_C to one, with the remainder, for components whose contribution to overall information is deemed negligible, zero.

In much of this chapter, we will use multivariate efficiency, described in Section 1.7.2, as the criterion for assessing weighting schemes. This is equivalent to maximising the sandwich information under the positive semidefinite or Loewner ordering.

It is worth noting that the weights we consider depend upon the data through the particular parameter values estimated for the dataset under consideration (Sections 3.2.2 and 3.2.4) as well as the length of the data clusters (Sections 3.2.3 and 3.2.5). In practice we will estimate parameters from the data and then use those estimates to calculate weights. Section 3.4.4, for example, explores this in more detail.

Also, we place no constraint on the support for the weights. In Section 4.8 we shall see an example with a negative weight. However, note that Lindsay et al. (2011) states that "if we were to include sub-likelihoods with negative weights, the guarantee of Fisher consistency would be lost". Indeed, arbitrary negative weights could give rise to such issues. This is most easily seen geometrically (see Section 3.2.4 for a longer geometric discussion). The weighted loglikelihood for a component with a positive weight will, in the area around $\theta_{\mathbf{G}}$ have decreasing gradient and a maximum. However, with a negative weight, there will be an increasing gradient and a minimum. Informally, as we wish the weighted composite loglikelihood to have a maximum, the components with negative weights must not overwhelm those with negative weights. Formally, Assumption 3 states that $\hat{\theta}_n$ exists and is unique and so the sum of the component loglikelihoods must have the decreasing gradient around $\theta_{\mathbf{G}}$. We thus permit negative weights subject to that assumption as, if the negatively weighted components predominated, there would be no unique maximum. See Section 3.2.2 for an example of what that might mean for scalar parameters. Our structured approach to assumptions and derivation of weights (see Section 3.4) ensures that any negative weights do not cause the problems described above.

3.2.2 Scalar Weights

Where θ is a scalar, the relative efficiency of $\hat{\theta}$ is just a number, rather than the matrix resulting from vector $\boldsymbol{\theta}$, and is a straightforward way to compare weighting schemes. Denote by $\boldsymbol{\psi}_S(\theta)$ the $q \times 1$ vector of stacked $\psi_C(\theta)$s from each of the composite components. Lindsay (1988) assumes that our data generating distribution $\mathbf{G}$ is parameterised by θ, with density g and score $U(\theta)$, and then shows that the vector of optimal efficiency improving weights, for a composite surrogate with estimating function $\psi(\theta)$, is:

$$\boldsymbol{w}_S^* = \mathsf{Var}^{-1}[\boldsymbol{\psi}_S(\theta_{\mathbf{G}})]\mathsf{E}[U(\theta_{\mathbf{G}})\boldsymbol{\psi}_S(\theta_{\mathbf{G}})] \tag{3.2}$$

where $\mathsf{E}[U(\theta_{\mathbf{G}})] = 0$ (we have shown that $\hat{\theta}_n$ is consistent for $\theta_{\mathbf{G}}$ in Section 1.3). The vector $\boldsymbol{w}_S^*$ represents $(w_1^*, \dots, w_q^*)^T$ where the w_j^* are the optimal component weights. The result is derived by minimising $\mathsf{E}[U(\theta_{\mathbf{G}}) - \boldsymbol{w}_S^T \boldsymbol{\psi}_S(\theta_{\mathbf{G}})]^2$, ie by maximising the Godambe information of the estimating function over the weights and treating the score as optimal per Section 1.7.2.

Clearly this result requires knowledge of **G** and only applies to scalar θ, but is a useful starting place for an understanding of the effect of varying weights.

It is worth exploring further the notion of negative weights, introduced in Section 3.2.1, for this simple situation. It is quite possible that (3.2) would have negative elements. For instance if we take a randomly generated covariance matrix with inverse

$$\mathsf{Var}^{-1}[\boldsymbol{\psi}_S(\theta_{\mathbf{G}})] = \begin{pmatrix} 0.154 & -0.057 & -0.006 & -0.008 \\ -0.057 & 0.322 & -0.023 & -0.015 \\ -0.006 & -0.023 & 0.148 & -0.005 \\ -0.008 & -0.015 & -0.005 & 0.158 \end{pmatrix}$$

and set $\mathsf{E}[U(\theta_{\mathbf{G}})\boldsymbol{\psi}_S(\theta_{\mathbf{G}})] = (0.1, 0.9, 0.1, 0.1)^T$ then

$$\boldsymbol{w}_S^* = (-0.0373, 0.2803, -0.007, 0.001)^T.$$

The constraint that prevents any negatively weighted components overwhelming those with positive weights is that $\mathsf{Var}^{-1}[\boldsymbol{\psi}_S(\theta_{\mathbf{G}})]$ must be positive semidefinite, in fact positive definite as the inverse exists, and so:

$$\begin{aligned} (\mathsf{E}[U(\theta_{\mathbf{G}})\boldsymbol{\psi}_S(\theta_{\mathbf{G}})])^T\mathsf{Var}^{-1}[\boldsymbol{\psi}_S(\theta_{\mathbf{G}})]\mathsf{E}[U(\theta_{\mathbf{G}})\boldsymbol{\psi}_S(\theta_{\mathbf{G}})] &\geq 0 \qquad \text{or} \\ (\mathsf{E}[U(\theta_{\mathbf{G}})\boldsymbol{\psi}_S(\theta_{\mathbf{G}})])^T\boldsymbol{w}_S^* &\geq 0. \end{aligned} \tag{3.3}$$

In the example above we have negative weights but $(\mathsf{E}[U(\theta_{\mathbf{G}})\boldsymbol{\psi}_S(\theta_{\mathbf{G}})])^T\boldsymbol{w}_S^* = 0.24794$.

Alternative approaches that relax the need for θ to be scalar have been suggested in specific contexts. For instance, the use of weighted bivariate composite likelihoods for large space time datasets has been studied by Bevilacqua et al. (2012). A simple weighting scheme, allocating weights per (3.1), is reviewed, whereby each weight is either 0 or 1 depending upon whether the distance and time between the pair of data points in that particular component are less than (d_t, d_s), say, respectively. These tuning parameters can be chosen by minimising numerically the trace of the variance matrix of the surrogate. In certain cases, this simple scheme is shown to be more efficient than that involving weights from the whole of $[0, 1]$. A more sophisticated version of this is applied to the surrogate's estimating function and is reviewed in Section 3.2.4.

We propose a new scalar weighting scheme, that is more generally applicable than those reviewed above, in Section 3.4.

3.2.3 Component Type Weights

In the case of composite likelihoods whose components are of more than one data dimension (eg a mixture of univariate and bivariate), several authors have worked with a subset of scalar weights where the weights differ only with the dimension of the data in the marginal components to which they are attached. The number of weights is the number of different marginal dimensions present. Varying the weights, may result in a variety of related distributions and in some cases will give rise to plausible interpretations (see, for instance, Section 4.8).

For instance, Cox and Reid (2004) consider the case where it is possible to specify the univariate and bivariate distributions but none of a higher dimension for a particular dataset. In that case, for m elements in a cluster, we have a weighted surrogate loglikelihood

$$\ell_{sc}(\theta; y) = \sum_{s>t} \ln f_{st}(y_s, y_t : \theta) - wm \sum_{s=1}^{m} \ln f_s(y_s; \theta) \tag{3.4}$$

where the suffix for f consists of the elements of y for which that distribution is marginal, and w is chosen by solving an optimality problem. The weight, w, can be considered as a relative weight attached to the univariate margins, and different choices lead to different interpretations. A particular instance, the pseudo-likelihood consisting of the product of all combinations of conditioning one data element upon another, was examined in Besag (1974) in the context of spatially interacting random variables. In that case (omitting the ys and θs for simplicity) the surrogate loglikelihood is

$$\begin{aligned}\ell_{sc} &= \sum_{s\neq t} \ln f_{s|t} \qquad 1 \leq s,t \leq m \\ &= \sum_{s\neq t}(\ln f_{st} - \ln f_t) \\ &= 2\sum_{s>t} \ln f_{st} - (m-1)\sum_t \ln f_t\end{aligned}$$

and, as loglikelihoods are equivalent, for parameter estimation, up to a multiplicative constant, we can recover (3.4) by taking $w = (m-1)/2m$ (note that Cox and Reid (2004) suggest using $w = 1/2$ for the same example). Technically, as we use the loglikelihood by differentiating it and setting the result equal to zero, we could reduce the number of weights required in all component type weight situations by one, by setting the weight of one of the composite terms to one and adjusting the rest of the weights accordingly.

This approach was extended in Lindsay et al. (2011), using the more general additive estimating function framework rather than loglikelihoods. At $\theta \in \mathbf{\Theta}_{ind}$, the subset of parameters at which all the $Y_{\mathcal{C}}$s are independent from each other, it is shown that certain values of weights (*Hoeffding scores*) maximise the sandwich information for the related parameter estimates. This elegant result is based upon starting with univariate components and then adding components of higher dimension (eg pairs, triplets) that are orthogonal to all previous components. For instance, if we just consider univariate and bivariate margins, the estimating function would be

$$\begin{aligned}\psi_2^* &= \sum_t \psi_t + \sum_{s>t}(\psi_{st} - \psi_s - \psi_t) \qquad 1 \leq s,t \leq m \qquad (3.5) \\ &= \sum_{s>t} \psi_{st} - (m-2)\sum_t \psi_t\end{aligned}$$

where the suffixes are as in (3.4), and we have recovered an estimating function version of (3.4) with $w = (m-2)/m$. The result arises as the residuals, $U - \psi_2^*$, where U is the score from **G**, are shown to be orthogonal to the basis of the set of additive estimating functions under consideration and thus to create an estimating function that is closer to U, one would have to add margins of higher dimension than two.

However, the result does not apply away from $\boldsymbol{\Theta}_{ind}$. If we only considered the values of $\boldsymbol{\Theta}_{ind}$ for which the Y_Cs are independent, then one would not bother with multivariate analysis at all. So, the value in this approach lies in how it can be extended to other parts of $\boldsymbol{\Theta}$, which is discussed in Section 3.2.4.

Note the difference in w for the Besag (1974) ($(m-1)/2m$) and Lindsay et al. (2011) ($(m-2)/m$) schemes. Besag (1974, Section 7.2.2), who was not aiming for optimality, does discuss overdependence on certain components in his model and this accounts for the difference in weights.

3.2.4 Estimating Function Weights

In Section 3.2.2, we have seen how scalar weights can optimise efficiency for a scalar parameter. More generally, θ will be vector valued and, in order to optimise efficiency, subtler weighting schemes may be required, so that weights can affect individual elements of the parameter vector within each composite component. The estimating function, as well as arising from a wider range of situations than just the loglikelihood, allows us to use these more complex weighting schemes.

One way of thinking about a loglikelihood surface for a composite surrogate is as the sum of a series of surfaces for each of the composite components. Each component surface will have expected maxima at the same place as the summary surface. Scalar weighting schemes multiply each component surface by a constant whereas estimating function schemes enable one to manipulate the shape of each surface, albeit, generally, through the surface representing the derivative of the loglikelihood. So, if we consider one of the elements that contributes to efficiency, namely the sharpness of the loglikelihood at $\boldsymbol{\theta}_{\mathbf{G}}$, expressed through the matrix of second derivatives there, a scalar weighting scheme just permits us to place greater emphasis on those components with greater sharpness (or sensitivity), while an estimating function weighting scheme actually allows us to improve the sharpness of each component. The latter scheme will thus improve sharpness and, consequently, efficiency of the composite surrogate more effectively than the former.

Lindsay et al. (2011) extends the incremental orthogonal scheme described in Section 3.2.3 away from $\mathbf{\Theta}_{ind}$ by applying more complex weights (*modified Hoeffding scores*) to estimating functions. So, each element in the right hand in (3.5), $\psi_{st} - \psi_s - \psi_t$, is replaced by

$$\psi_{st} - B_s\psi_s - B_t\psi_t$$

where B_s and B_t are $p \times p$ matrices derived by minimising

$$\mathsf{E}[(\psi_{st} - B_s\psi_s - B_t\psi_t)(\psi_{st} - B_s\psi_s - B_t\psi_t)^T].$$

Application to specific multivariate normal examples is examined in more detail in Section 4.2.

A more general weighting scheme would be to apply matrix valued weights to the vector components of estimating functions so that they sum to a weighted composite surrogate (*wcs*) estimating function

$$\boldsymbol{\psi}_{wcs}(\boldsymbol{\theta}; y) = \sum_{C \in \mathcal{C}} \boldsymbol{W}_C \boldsymbol{\psi}_C(\boldsymbol{\theta} : y_C)$$

so that each weight, $\boldsymbol{W}_C$, is a $p \times p$ matrix and we have q of these matrices, each with p^2 weights.

Lindsay et al. (2011) claims that the set of $\boldsymbol{W}_j$ which maximises efficiency, which we shall describe as the *Best Weighted Estimating Function* (*BWEF*), can be found as follows. As in Section 3.2.2, define $\boldsymbol{\psi}_S(\boldsymbol{\theta}) = (\psi_1(\boldsymbol{\theta}), \ldots, \psi_q(\boldsymbol{\theta}))^T$ as the $pq \times 1$ vector formed by stacking the component estimating functions in order, and then set

$$\begin{aligned} \boldsymbol{C}_S(\boldsymbol{\theta}) &= \mathsf{Var}[\boldsymbol{\psi}_S(\boldsymbol{\theta})] \quad \text{and} \\ \boldsymbol{I}_S(\boldsymbol{\theta}) &= -\mathsf{E}\left[\frac{\partial \boldsymbol{\psi}_S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \quad \text{or} \quad -\mathsf{E}[\psi'_S(\theta)], \end{aligned}$$

being $pq \times pq$ and $pq \times p$ matrices respectively. Then, set:

$$\boldsymbol{W}_B = \boldsymbol{I}_S(\boldsymbol{\theta}_{\mathbf{G}})^T \boldsymbol{C}_S(\boldsymbol{\theta}_{\mathbf{G}})^{-1}, \tag{3.6}$$

where $\boldsymbol{W}_B = (\boldsymbol{W}_{B_1}, \ldots, \boldsymbol{W}_{B_q})$ is a $p \times pq$ matrix of the weighting matrices stacked horizontally. This results in the BWEF, $\boldsymbol{\psi}_B(\boldsymbol{\theta})$, which is the weighting scheme resulting in the most efficient parameter estimates under a positive definite matrix partial order:

$$\begin{aligned} \boldsymbol{\psi}_B(\boldsymbol{\theta}) &= \boldsymbol{W}_B \boldsymbol{\psi}_S(\boldsymbol{\theta}) \qquad (3.7) \\ &= \sum_{C \in \mathcal{C}} \boldsymbol{W}_{B_C} \boldsymbol{\psi}_C(\boldsymbol{\theta}). \qquad (3.8) \end{aligned}$$

A similar approach is taken in the Generalized Method of Moments (Hansen, 1982) for the derivation of optimal weighting matrices. In that case, which is not restricted to composite estimating functions, $\theta_{\mathbf{G}}$ minimises $\psi(\theta)^T \Omega \psi(\theta)$ where Ω is a $p \times p$ matrix of weights. It can then be shown that the Ω which maximises efficiency is $\text{Var}^{-1}[\psi(\theta_{\mathbf{G}})]$.

The result at (3.7) is merely stated and not proved in Lindsay et al. (2011). In addition, $\boldsymbol{W}_B$ is only defined if $\boldsymbol{C}_S(\boldsymbol{\theta}_{\mathbf{G}})$ is non singular. This assumption is never stated but appears to be made a number of times in the paper. A more general version of this result is suggested and proved in Section 3.3.1.

The main disadvantage of (3.6), as noted in Lindsay et al. (2011), is its computational inefficiency - it requires the inversion of a $pq \times pq$ matrix. One way to avoid that inversion is to ignore any dependence between elements of $\boldsymbol{\psi}_S(\boldsymbol{\theta})$, irrespective of whether those elements belong to the same component. Bevilacqua et al. (2012) do that, amongst other things, to extend their simple weighting scheme, described in Section 3.2.2, to estimating functions, resulting in a weighted composite score, $e_W(\boldsymbol{\theta})$:

$$e_W(\boldsymbol{\theta}) = \text{diag}(\boldsymbol{I}_S(\boldsymbol{\theta}))\boldsymbol{\psi}_S(\boldsymbol{\theta}).$$

Justification is through minimising an upper bound for the asymptotic variance of the surrogate parameter estimates. The scheme appears to improve efficiency in certain cases compared with unweighted and scalar weighted composite surrogates.

A half way house that takes into account dependence between estimating function elements of the same component is considered in Section 3.3.2.

3.2.5 Cluster Weights

Most references for composite likelihoods (eg Lindsay, 1988; Varin, 2008; Varin et al., 2011) just specify multivariate data, y, each instance or cluster being of some fixed length, m. In some situations, such as repeated weather readings from a fixed number of stations, this is appropriate, although missing or unreliable data could cause problems. However, in, for instance, longitudinal data studies, the clusters are very likely to be of unequal length. In that case, the use of equal weights in composite likelihoods will implicitly derive more information from clusters with greater length. This may not always be appropriate.

For instance, le Cessie and van Houwelingen (1994) studied the effects of various (mostly peri-natal) effects on binary outcomes of mortality and morbidity of children over the first few years of their lives. As some of the children were members of twins, triplets etc, there was correlation between the observations of the effects for members of the same multiple birth groups. The impact of each of these individuals compared with the single birth event individuals thus needed to be reduced. This particular situation was then generalised to consider blocks or clusters of length m_i in bivariate composite loglikelihoods and it was proposed, heuristically, that weights of $1/(m_i - 1)$ were applied to cluster i to reduce the effect of dependent individuals from larger clusters. Thus, the weights are attached to the clusters rather than the components and there are q of them.

Joe and Lee (2009) introduced more sophisticated functions of the cluster size to cope with situations where there is one large and many small clusters. These are derived heuristically from multivariate normal models where the data generating mechanism and the bivariate composite surrogates have common means and variances, and exchangeable correlations. The various weighting schemes are compared for relative efficiency. This comparison is analytical for one (correlation) unknown parameter. For all three parameters simultaneously, the results are through simulation. Results vary for possible length of clusters and parameter values but, in general, weights of $1/(m_i - 1)$ and $1/((m_i - 1)(1 + (m_i - 1)/2)$ perform best.

Example III continued - Longitudinal Study. If we apply the second of the preferred weighting schemes from Joe and Lee (2009) to patient measurements where one patient has had one measurement and a second many, then the weighted contribution of the second to the loglikelihood used for parameter estimation will be of the same order as that from the first (the weighting of the second is of the same order as the number of composite components). This does not seem appropriate as the second patient will be contributing far more information to the study. □

There is potential for exploring combined cluster length and, for instance, scalar weights.

3.2.6 The Way Ahead

We have seen a variety of proposed weighting schemes for composite surrogates. Many of them are developed heuristically or apply only to specific distributions. The only one that meets some general optimality criterion, in this case efficiency, analytically is the BWEF described in Section 3.2.4. However, that scheme, as proposed but not proved, makes the implicit assumption that each component varies with every parameter . That is not always the case: for instance, where we have bivariate components derived from a full multivariate distribution, each correlation parameter will only appear in one component. We suggest and prove a more general result that does not make that assumption in Section 3.3.1. The scheme is potentially computationally expensive and so we propose a restricted version that assumes no dependence between component estimating functions in Section 3.3.2 and show that it is still best in its class.

The theory of surrogates as set out in Chapter 1 is based around minimising the KLD between the distribution that generated that data and the surrogate. We take that principle and apply it to a composite surrogate with constant of proportionality in Section 3.4. This results in a completely new set of equations that can be solved to derive weights under the KLD optimality criterion.

It is worth noting that the number of weights in the different schemes reviewed so far varies from the number of component dimensions (usually two) for component type weights to p^2q^2 for the BWEF proposal. The new KLD proposal has q weights and the partially dependent scheme qp^2. The number of weights does not necessarily reflect the complexity of calculation, and the weights may well have to be calculated for each cluster if the cluster distributions are not iid.

3.3 Estimating Function Weighting Schemes

3.3.1 A Fully Dependent Weighting Scheme

We saw in Section 3.2.4 that the proposed optimal estimating function weighting scheme in Lindsay et al. (2011) required $\boldsymbol{C}_S(\boldsymbol{\theta}_\mathbf{G})$ to be non singular. This requires, inter alia, that we are working with a full composite surrogate which we define in the following:

Full composite surrogate Each component of the composite varies with the full set of parameters under consideration (ie the full set used in the composite surrogate, not the set used in some idealised joint distribution). This will often be the case where we are working with covariates and a link function, such as in the Generalized Linear Model

Projected composite surrogate Each component of the composite varies with only a proper subset of the parameters under consideration. An example of this is where we have multivariate data (dimension greater than two) with a bivariate composite surrogate where each component is bivariate normal and will thus depend upon parameters relating to only two of the elements of the data. The term "projected" is used to denote the fact that we have collapsed the parameter space for any component onto the dimensions for which parameters exist.

If we are working with a projected composite surrogate, then there will be at least one zero element in $\boldsymbol{\psi}_S(\boldsymbol{\theta}_{\mathsf{G}})$ and equivalent zero row and column in $\boldsymbol{C}_S(\boldsymbol{\theta}_{\mathsf{G}})$, so the latter will be singular. To address this problem we here provide a more precise statement and proof of Lindsay et al. (2011)'s proposal.

Define $\boldsymbol{\psi}_S^0(\boldsymbol{\theta})$ to be $\boldsymbol{\psi}_S(\boldsymbol{\theta})$ with the zero rows removed, ie $\boldsymbol{\psi}_S^0(\boldsymbol{\theta}) = \boldsymbol{B}\boldsymbol{\psi}_S(\boldsymbol{\theta})$ where $\boldsymbol{B}$ is a $pq \times pq$ identity matrix with the rows corresponding to zeroes in $\boldsymbol{\psi}_S(\boldsymbol{\theta})$ deleted. Similarly, define $\boldsymbol{I}_S^0(\boldsymbol{\theta}_{\mathsf{G}}) = \boldsymbol{B}\boldsymbol{I}_S(\boldsymbol{\theta}_{\mathsf{G}})$ and $C_S^0(\boldsymbol{\theta}_{\mathsf{G}}) = \boldsymbol{B}C_S(\boldsymbol{\theta}_{\mathsf{G}})\boldsymbol{B}^T$. Finally, set

$$\boldsymbol{W}_B^0 = \boldsymbol{I}_S^0(\boldsymbol{\theta}_{\mathsf{G}})^T \boldsymbol{C}_S^0(\boldsymbol{\theta}_{\mathsf{G}})^{-1} \tag{3.9}$$

so that

$$\boldsymbol{\psi}_B^0(\boldsymbol{\theta}) = \boldsymbol{W}_B^0 \boldsymbol{\psi}_S^0(\boldsymbol{\theta}).$$

Assumption 19. $\boldsymbol{C}_S^0(\boldsymbol{\theta}_G)$ *and its estimate,* $\hat{\boldsymbol{C}}_n^0$ *(see Section 1.4), is nonsingular.*

In Section 1.6.3 we see that $\boldsymbol{C}_n$ is non-singular, but Assumption 19 is a further requirement: that invertibility remains when we look at the composite components individually, in blocks down the main diagonal of $\boldsymbol{C}_S^0(\boldsymbol{\theta}_{\mathsf{G}})$, and in pairs, off the main diagonal. Technically, we require that the product of $\boldsymbol{I}_S^0(\boldsymbol{\theta}_{\mathsf{G}})$ with various other matrices is nonsingular but this assumption ensures that is the case.

Note that we are typically working with classes of estimating functions per Section 1.7 where an estimating function is premultiplied by $\boldsymbol{I}^{-1}(\boldsymbol{\theta}_{\mathsf{G}})$ to give the class representative. We thus premultiply any weighting scheme by a normalising matrix $\boldsymbol{I}_S^0(\boldsymbol{\theta}_{\mathsf{G}})^T$ to adjust for that and allow us to compare any weighting scheme.

We now state and prove our optimality theorem for $\boldsymbol{\psi}_S^0(\boldsymbol{\theta})$.

Theorem 3.3.1. $\boldsymbol{W}_B^0$*, as defined in (3.9), is the most efficient estimating function weighting scheme for* $\boldsymbol{\psi}_S^0(\boldsymbol{\theta})$*, subject to the existence of the appropriate moment functions of* $\boldsymbol{\psi}_S^0(\boldsymbol{\theta})$*.*

Proof Firstly, note that $\boldsymbol{\psi}_B^0(\boldsymbol{\theta}) = \boldsymbol{W}_B^0\boldsymbol{\psi}_S^0(\boldsymbol{\theta})$ is an estimating function. By Assumption 17, each of the elements of $\boldsymbol{\psi}_S^0(\boldsymbol{\theta})$ has expected value zero at $\boldsymbol{\theta}_{\mathbf{G}}$ under **G** and thus any linear combination will have the same characteristic, so that $\boldsymbol{E}[\boldsymbol{\psi}_B^0(\boldsymbol{\theta}_{\mathbf{G}})] = 0$.

For proving efficiency, we adopt a similar approach to that in Hansen (1982). This originally arose in the study of General Method of Moments (GMM) where the dimension of the estimating function (moment function in GMM) is not necessarily the same as that of θ.

We consider the asymptotic parameter variance under BWEF ($\mathsf{Var}_B[\hat{\boldsymbol{\theta}}_\infty]$) which has the usual sandwich form but can also be reduced, all functions being evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathbf{G}}$ (and the suffix 0 being dropped for notational simplicity):

$$\begin{aligned}
\mathsf{Var}_B[\hat{\boldsymbol{\theta}}_\infty] &= (\mathsf{E}[\boldsymbol{\psi}_B']^T)^{-1}\mathsf{Var}[\boldsymbol{\psi}_B]\mathsf{E}[\boldsymbol{\psi}_B']^{-1} \\
&= (\mathsf{E}[\boldsymbol{W}_B\boldsymbol{\psi}_S']^T)^{-1}\mathsf{Var}[\boldsymbol{W}_B\boldsymbol{\psi}_S]\mathsf{E}[\boldsymbol{W}_B\boldsymbol{\psi}_S']^{-1} \\
&= (\mathsf{E}[\boldsymbol{\psi}_S']^T\boldsymbol{W}_B^T)^{-1}\boldsymbol{W}_B\mathsf{Var}[\boldsymbol{\psi}_S]\boldsymbol{W}_B^T(\boldsymbol{W}_B\mathsf{E}[\boldsymbol{\psi}_S'])^{-1} \\
&= (\boldsymbol{I}_S^T\boldsymbol{W}_B^T)^{-1}\boldsymbol{W}_B\boldsymbol{C}_S\boldsymbol{W}_B^T(\boldsymbol{W}_B\boldsymbol{I}_S)^{-1} \\
&= (\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{C}_S\boldsymbol{C}_S^{-1}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1} \\
&= (\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1} \\
&= (\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}.
\end{aligned}$$

If we then consider any other weighting scheme

$$\begin{aligned}
\boldsymbol{\psi}_W &= \boldsymbol{W}\boldsymbol{\psi}_S \\
\boldsymbol{W} &= \boldsymbol{I}_S^T\boldsymbol{V}
\end{aligned}$$

for some $pq \times pq$ symmetric matrix $\boldsymbol{V}$, its asymptotic parameter variance will be the usual sandwich form

$$\begin{aligned}
\mathsf{Var}_W[\hat{\boldsymbol{\theta}}_\infty] &= (\mathsf{E}[\boldsymbol{\psi}'_W]^T)^{-1}\mathsf{Var}[\boldsymbol{\psi}_W]\mathsf{E}[\boldsymbol{\psi}'_W]^{-1} \\
&= (\mathsf{E}[\boldsymbol{W}\boldsymbol{\psi}'_S]^T)^{-1}\mathsf{Var}[\boldsymbol{W}\boldsymbol{\psi}_S]\mathsf{E}[\boldsymbol{W}\boldsymbol{\psi}'_S]^{-1} \\
&= (\mathsf{E}[\boldsymbol{\psi}'_S]^T\boldsymbol{W}^T)^{-1}\boldsymbol{W}\mathsf{Var}[\boldsymbol{\psi}_S]\boldsymbol{W}^T(\boldsymbol{W}\mathsf{E}[\boldsymbol{\psi}'_S])^{-1} \\
&= (\boldsymbol{I}_S^T\boldsymbol{W}^T)^{-1}\boldsymbol{W}\boldsymbol{C}_S\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{I}_S)^{-1} \\
&= (\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{C}_S\boldsymbol{V}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{I}_S)^{-1}
\end{aligned}$$

as above. This expression cannot be simplified further in general as we have made no assumptions about how $\boldsymbol{V}$ and $\boldsymbol{C}_S$ are related. Then:

$$\begin{aligned}
\mathsf{Var}_W[\hat{\boldsymbol{\theta}}] - \mathsf{Var}_B[\hat{\boldsymbol{\theta}}] &= (\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{C}_S\boldsymbol{V}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{I}_S)^{-1} - (\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1} \\
&= (\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{I}_S)^{-1}(\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{C}_S^{\frac{1}{2}})\left(\boldsymbol{I} - \boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}\right) \\
&\quad .(\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{C}_S^{\frac{1}{2}})^T(\boldsymbol{I}_S^T\boldsymbol{V}\boldsymbol{I}_S)^{-1} \\
&= \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}^T
\end{aligned}$$

say, where $\boldsymbol{C}_S^{\frac{1}{2}}$ is a unique real positive definite and symmetric square root (Horn and Johnson, 1987, theorem 7.2.6, page 405, as $\boldsymbol{C}$ is positive definite) and:

$$\begin{aligned}
\boldsymbol{M}\boldsymbol{M}^T &= (\boldsymbol{I} - \boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}})(\boldsymbol{I} - \boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}})^T \\
&= \boldsymbol{I} - (\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}})^T - (\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}) \\
&\quad +(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}})(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}})^T \\
&= \boldsymbol{I} - 2(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}) \\
&\quad +(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}})(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}) \\
&= \boldsymbol{I} - 2(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}) \\
&\quad +(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}) \\
&= \boldsymbol{I} - 2(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}) \\
&\quad +(\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}) \\
&= \boldsymbol{I} - (\boldsymbol{C}_S^{-\frac{1}{2}}\boldsymbol{I}_S(\boldsymbol{I}_S^T\boldsymbol{C}_S^{-1}\boldsymbol{I}_S)^{-1}\boldsymbol{I}_S^T\boldsymbol{C}_S^{-\frac{1}{2}}) \\
&= \boldsymbol{M}
\end{aligned}$$

where the third equality arises as $\boldsymbol{C}_S^{-1}$ and then $\boldsymbol{B}\boldsymbol{C}_S^{-1}\boldsymbol{B}^T$ are symmetric for any $pq \times pq$ matrix $\boldsymbol{B}$ (Horn and Johnson, 1987, Section 4.1). Finally, we can see that:

$$\begin{aligned}
\mathsf{Var}_W[\hat{\boldsymbol{\theta}}] - \mathsf{Var}_B[\hat{\boldsymbol{\theta}}] &= \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}^T \\
&= \boldsymbol{L}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{L}^T \\
&= (\boldsymbol{L}\boldsymbol{M})(\boldsymbol{L}\boldsymbol{M})^T
\end{aligned}$$

which is semi-positive definite and therefore by the definition of efficiency in Section 1.7.2 the BWEF gives rise to the smallest possible parameter variance. □

We can now return to a weighted estimating function with components containing the right number of elements by

$$\boldsymbol{\psi}_B(\boldsymbol{\theta}) = \boldsymbol{B}^T\boldsymbol{\psi}_B^0(\boldsymbol{\theta}),$$

which restores the zeroes in the same locations that they were taken away before Assumption 19, and then destacking the components. One might consider restoring non zero elements to the estimating function. However, by Assumption 17, in order for it to remain as an estimating function, the expected value of these elements would have to be zero. Since, by Assumption 3, the estimating equations give rise to unique parameter estimators, these new non zero elements would just be linear combinations of the existing estimating function elements. As a consequence, they add no new information and the parameter estimators taking them into account would be no more efficient than those without them.

A useful property of our most efficient estimating function follows.

Theorem 3.3.2. *The estimating function* $\boldsymbol{\psi}_B^0(\boldsymbol{\theta}) \equiv \boldsymbol{W}_B^0(\boldsymbol{\theta}_{\boldsymbol{G}})\boldsymbol{\psi}_S^0(\boldsymbol{\theta})$ *satisfies the Information Identity at* $\boldsymbol{\theta}_{\boldsymbol{G}}$*.*

Proof We omit the $\boldsymbol{\theta}_{\mathbf{G}}$s, where appropriate, for ease of notation:

$$\begin{aligned}
\mathsf{Var}[\boldsymbol{\psi}_B^0(\theta)]|_{\theta=\theta_{\mathbf{G}}} &= \mathsf{Var}[\boldsymbol{W}_B^0\boldsymbol{\psi}_S^0] \\
&= \boldsymbol{W}_B^0\mathsf{Var}[\boldsymbol{\psi}_S^0](\boldsymbol{W}_B^0)^T \\
&= (\boldsymbol{I}_S^0)^T(\boldsymbol{C}_S^0)^{-1}\mathsf{Var}[\boldsymbol{\psi}_S^0](\boldsymbol{C}_S^0)^{-1}\boldsymbol{I}_S^0 \\
&= (\boldsymbol{I}_S^0)^T(\boldsymbol{C}_S^0)^{-1}\boldsymbol{C}_S^0(\boldsymbol{C}_S^0)^{-1}\boldsymbol{I}_S^0 \\
&= (\boldsymbol{I}_S^0)^T(\boldsymbol{C}_S^0)^{-1}\boldsymbol{I}_S^0
\end{aligned}$$

and

$$\begin{aligned}
\left.\frac{\mathrm{d}\boldsymbol{\psi}_B^0}{\mathrm{d}\theta}\right|_{\theta=\theta_{\mathbf{G}}} &= \left.\frac{\mathrm{d}\boldsymbol{W}_B^0\boldsymbol{\psi}_S^0}{\mathrm{d}\theta}\right|_{\theta=\theta_{\mathbf{G}}} \\
&= \boldsymbol{W}_B^0\left.\frac{\mathrm{d}\boldsymbol{\psi}_S^0}{\mathrm{d}\theta}\right|_{\theta=\theta_{\mathbf{G}}} \\
&= (\boldsymbol{I}_S^0)^T(\boldsymbol{C}_S^0)^{-1}\left.\frac{\mathrm{d}\boldsymbol{\psi}_S^0}{\mathrm{d}\theta}\right|_{\theta=\theta_{\mathbf{G}}} \\
&= -(\boldsymbol{I}_S^0)^T(\boldsymbol{C}_S^0)^{-1}\boldsymbol{I}_S^0.
\end{aligned}$$

□

This is useful for hypothesis testing, since nested models can be compared using a χ^2 distribution rather than a weighted sum of χ_1^2s for the difference in surrogate loglikelihood as described in Section 1.5.

As well as the very strong assumption made about the non singularity of $\boldsymbol{C}_S^0(\boldsymbol{\theta})$, Theorem 3.3.1 requires knowledge of some moments of $\boldsymbol{\psi}$ and its derivative in the area of of θ_{G}. However, the most significant issue with this approach to optimality in practice is that we are required to invert a $pq \times pq$ (or slightly smaller to account for the zeroes) matrix. The number of computer operations required to do this is $O(p^3q^3)$. One of the reasons for using composite surrogate techniques is the complexity of the model we would like to study, ie the number of parameters (p) and / or the dimension of each cluster (q) is large. The matrix inversion would then be formidable and so this result may not be useful in practice.

3.3.2 A Partially Dependent Weighting Scheme

We saw in Section 3.2.4 that optimal estimating function weighting schemes are computationally expensive. The dependence between elements within and between components of estimating functions ensures that we are required to invert a potentially large matrix. We also saw a scheme that ignores all such dependencies. In this section we propose a scheme that retains the dependence between elements of estimating functions in the same composite component but ignores the dependence between components. We consider the weighted estimating function

$$\begin{aligned} \psi_w^*(\theta) &= \sum_{C \in \mathcal{C}} \boldsymbol{W}_C^* \psi_C(\theta) \\ &= -\sum_{C \in \mathcal{C}} \mathsf{E}[\psi_C'(\theta_{\mathsf{G}})]^T \mathsf{Var}[\psi_C(\theta_{\mathsf{G}})]^{-1} \psi_C(\theta) \qquad (3.10) \\ & \qquad (3.11) \end{aligned}$$

This does require the inversion of q covariance matrices but they have dimensions only $p \times p$ giving a total of $O(qp^3)$ operations compared to the $O(q^3p^3)$ operations required for the BWEF inversion described at (3.6). The scheme is similar to the vertical adjustment described in Section 1.6.3 where an adjustment $(\boldsymbol{I}(\theta_{\mathsf{G}})^T\boldsymbol{C}(\theta_{\mathsf{G}})^{-1})$ is made to the whole of the surrogate estimating function. However, here a similar type of weight $(\boldsymbol{I}_C(\theta_{\mathsf{G}})^T\boldsymbol{C}_C(\theta_{\mathsf{G}})^{-1})$ is applied separately to each component of the composite surrogate.

If we take the same approach as we did in Section 3.3.1, ie make Assumption 19, deal with zero elements in the estimating function and assume appropriate moments exist, then Appendices D and E show that (3.10) defines the most efficient weighting scheme amongst the class of estimating functions where dependence between components is ignored, at $\boldsymbol{\theta}_{\mathsf{G}}$. The proof used in Section 3.3.1 does not seem to work in this case and so we make use of an alternative approach.

In practice this scheme involves the inversion of q $p \times p$ matrices. While that is computationally cheaper than the scheme from Section 3.3.1, many applications, for instance in genetics, have large numbers of parameters and make even this approach unfeasible. We now turn to a completely new scheme based around scalar weights.

3.4 A Weighting Scheme Based Upon Constant of Proportionality

3.4.1 Introduction

In Section 1.2, we established the minimisation of the Kullback-Leibler Divergence (KLD) as a criterion for estimating parameters. We have seen other criteria, such as efficiency, used for calculating weights but it seems reasonable to understand the effect of extending the KLD approach in order to do the same. In essence, we are trying to improve model fit.

A naive approach to calculating weights for a given surrogate composite loglikelihood of the form given in (3.1) would be to minimise the KLD between the surrogate and the density, g, from which the data was generated. As there are no weights present in the latter, however, this is equivalent is maximising the expected value of the loglikelihood of the surrogate over the weights. The value is just a linear combination of the weights and we can thus maximise it by assigning a weight of one to the weight (w_m) with the largest coefficient $(\ln f_m(\theta; y_m))$. A different dataset might assign primacy to a different component. This approach takes no account of either of complexity in the composite model nor of natural variation in the data. It is thus unsatisfactory.

However, if we take into account the effect of adding a constant of proportionality, K^{-1}, to a composite surrogate to give a distribution $\mathbf{F}_K$ with a density, f_K, as described in Section 2.5, this unsatisfactory weighting will no longer be the case. Example IV in that section shows that K is likely to be a function of the weights and we will make that explicit below by using $K(w)$. While this approach takes us away from the simplicity of composite likelihoods, it is not completely unreasonable as we are working with a genuine density that is derived from the multivariate distribution with which we would like to work ($\mathbf{H}$).

The main theorem, giving equations to be solved for optimal weights is stated, proved and discussed in Section 3.4.2. The situation where unique weights may not result is discussed in Section 3.4.3. Use of the equations in practice is analysed in Section 3.4.4 and applied to a simulation in Section 3.4.5.

3.4.2 Theory

A surrogate with a constant of proportionality has density

$$f_K(y) = \frac{1}{K(w)} \prod_{C \in \mathcal{C}} f_C(y_C; \theta)^{w_C}.$$

For a given dataset generated by a possibly unknown distribution $\mathbf{G}$, with density $g(y)$, we can calculate weights by minimising the KLD between $\mathbf{F}_K$ and $\mathbf{G}$. This results in the rather pleasing

Theorem 3.4.1. *For a composite surrogate, $\boldsymbol{F}_K$, that includes a constant of proportionality, the set of weights that minimise $D(\theta)$, the KLD from $\boldsymbol{G}$, at $\theta_{\boldsymbol{G}}$ can be* *found as a solution to*

$$E_{\boldsymbol{F}_K}[\ln f_C(Y_C;\theta_{\boldsymbol{G}})] = E_{\boldsymbol{G}}[\ln f_C(Y_C;\theta_{\boldsymbol{G}})] \quad C \in \mathcal{C} \tag{3.12}$$

where f_C is the component of $\boldsymbol{F}$ acting on Y_C, a subset of Y. In addition, for $C_1, C_2 \in \mathcal{C}$

$$\begin{aligned}\frac{\partial^2 D}{\partial w_{C_i}\partial w_{C_j}} &= E_{\boldsymbol{F}_K}[\ln f_{C_i}(Y_{C_i};\theta_{\boldsymbol{G}}) \ln f_{C_j}(Y_{C_j};\theta_{\boldsymbol{G}})] \\ &\quad - E_{\boldsymbol{F}_K}[\ln f_{C_i}(Y_{C_i};\theta_{\boldsymbol{G}})] E_{\boldsymbol{F}_K}[\ln f_{C_j}(Y_{C_j};\theta_{\boldsymbol{G}})]^T \end{aligned} \tag{3.13}$$

and thus the matrix of second derivatives of D with respect to the weights, $\boldsymbol{J}$ say, is just the covariance matrix of $\{\ln f_C(Y_C;\theta_{\boldsymbol{G}}) : C \in \mathcal{C}\}$ under $\boldsymbol{F}_K$.

Proof For simplicity of notation, the proof given here assumes that Y has a continuous density function. The extension to more general settings is straightforward by an appropriate choice of measure for the integration - the proof is otherwise unchanged. We omit the $\theta_{\mathbf{G}}$ for brevity. By the definition of the Kullback-Leibler Divergence

$$\begin{aligned} D &= \mathsf{E}_{\mathbf{G}}[\ln(g(Y)/f_K(Y))] \\ &= \int g(y)\left(\ln g(y) - \ln f_K(y)\right) \mathrm{d}y \\ &= \int g(y)\left(\ln g(y) - \ln\left(K(\omega)^{-1}\prod_{C\in\mathcal{C}} f_C(y_C)^{w_C}\right)\right)\mathrm{d}y \\ &= \ln K(\omega) + \int g(y)\left(\ln g(y) - \sum_{C\in\mathcal{C}} w_C \ln f_C(y_C)\right)\mathrm{d}y. \end{aligned}$$

Differentiating with respect to w_k, the weight for composite component k

$$\begin{aligned}\frac{\partial D}{\partial w_k} &= K(\omega)^{-1}\frac{\partial K(\omega)}{\partial w_k} - \int g(y)\ln f_k(y_k)\mathrm{d}y \\ &= K(\omega)^{-1}\frac{\partial K(\omega)}{\partial w_k} - \mathsf{E}_{\mathbf{G}}[\ln f_k(Y_k)]. \end{aligned}$$

From the definition of $K(\boldsymbol{w})$ at (2.8)

$$\begin{aligned} \frac{\partial K(\omega)}{\partial w_k} &= \frac{\partial}{\partial w_k}\left(\int \prod_{C\in\mathcal{C}} f_C(y_C)^{w_C}\mathrm{d}y\right) \\ &= \int \ln f_k(y_k) \prod_{C\in\mathcal{C}} f_C(y_C)^{w_C}\mathrm{d}y \qquad (3.14) \end{aligned}$$

so that

$$\begin{aligned} K(\omega)^{-1}\frac{\partial K(\omega)}{\partial w_k} &= \int K(\omega)^{-1}\prod_{C\in\mathcal{C}} f_C(y_C)^{w_C}\ln f_k(y_k)\mathrm{d}y \\ &= \mathsf{E}_{\mathbf{F}_K}[\ln f_k(Y_k)]. \qquad (3.15) \end{aligned}$$

In order to minimise D over the weights, we set $\partial D/\partial w_k = 0$ so that

$$\mathsf{E}_{\mathbf{F}_K}[\ln f_C(Y_C)] = \mathsf{E}_{\mathbf{G}}[\ln f_C(Y_C)] \quad C\in\mathcal{C}.$$

This establishes the first part of the theorem.

Next, differentiating D with respect to the weights for any two, possibly equal, composite components C_i and C_j, we obtain

$$\begin{aligned} \frac{\partial^2 D}{\partial w_{C_1}\partial w_{C_2}} &= \frac{\partial}{\partial w_{C_2}}\left(K(\omega)^{-1}\frac{\partial K(\omega)}{\partial w_{C_1}}\right) \\ &= K(\omega)^{-1}\frac{\partial^2 K(\omega)}{\partial w_{C_1}\partial w_{C_2}} - K(\omega)^{-2}\left(\frac{\partial K(\omega)}{\partial w_{C_1}}\frac{\partial K(\omega)^T}{\partial w_{C_2}}\right) \\ &= K(\omega)^{-1}\frac{\partial^2 K(\omega)}{\partial w_{C_1}\partial w_{C_2}} - \mathsf{E}_{\mathbf{F}_K}[\ln f_{C_1}(Y_{C_1})]\mathsf{E}_{\mathbf{F}_K}[\ln f_{C_2}(Y_{C_2})]^T \quad \text{by (3.15)} \\ &= \mathsf{E}_{\mathbf{F}_K}[\ln f_{C_1}(Y_{C_1})\ln f_{C_2}(Y_{C_2})] - \mathsf{E}_{\mathbf{F}_K}[\ln f_{C_1}(Y_{C_1})]\mathsf{E}_{\mathbf{F}_K}[\ln f_{C_2}(Y_{C_2})]^T \\ &= \mathsf{Cov}_{\mathbf{F}_K}[\ln f_{C_1}(Y_{C_1}), \ln f_{C_2}(Y_{C_2})] \end{aligned}$$

where the penultimate line arises by applying (3.14) twice. Thus, the second derivative of D with respect to the weights, $\boldsymbol{J}$, is just a covariance matrix

$$\boldsymbol{J} = \mathsf{Cov}\left[\{\ln f_C(Y_C;\theta_{\mathbf{G}}) : C\in\mathcal{C}\}\right].$$

□

A number of points can be made about this result:

1. A weighted composite loglikelihood varies with both the parameters and the weights. We derive values for each of them separately. Thus, the optimal weights we arrive at through solving (3.12) will vary, typically with the value of θ used. We are aiming for a target value of θ that most closely matches our object of interest from **G** and there will be weights that match that value. Any parameter estimator, $\hat{\theta}$, will depend upon the choice of weights. Since, by Assumption 17, the parameter estimates arising from each component are the same, $\hat{\theta}$ will be consistent for $\theta_{\mathbf{G}}$ irrespective of the choice of weights. How this is implemented in practice is explored in Section 3.4.4.

2. Adding together the different equations from (3.12) we get, for any value of θ

$$\mathsf{E}_{\mathbf{F}_K}[\ell_c(\theta;Y)] = \mathsf{E}_{\mathbf{G}}[\ell_c(\theta;Y)]; \qquad \text{or} \tag{3.16}$$

$$\mathsf{E}_{\mathbf{F}_K}[\ell_K(\theta;Y)] + \ln(K) = \mathsf{E}_{\mathbf{G}}[\ell_c(\theta;Y)], \tag{3.17}$$

   where ℓ_K is the loglikelihood of the composite surrogate including constant of proportionality, so that the optimal weights represent the point where the expected value of the composite surrogate likelihood is the same under the distribution that generated the data and the composite distribution including the constant of proportionality, (3.16).

3. The simplicity of (3.12) may turn into a very messy set of q equations to solve for the weights. The left hand side (LHS) requires knowledge of the form of $K(\boldsymbol{w})$, as shown, for example, in Example IV in Section 2.5, which may not always be possible. In addition, these equations may be expensive to solve - $O(q^3)$ for linear equations in the weights, the same order for every iteration of a numerical approximation such as Newton-Raphson. However, see Section 4.7 for an approximation that is cheaper - $O(m^3)$ rather than $O(q^3)$, where $m < q$: for bivariate composites $q = m(m-1)/2$.

4. No assumption is made about the form of or our knowledge of **G** on the right

hand side (RHS). However, if the components of **F** are marginal for **G**, then $\mathsf{E}_{\mathbf{G}}[\ln f_C(Y_C))] = \mathsf{E}_{\mathbf{F}_C}[\ln f_C(Y_C))]$ (ie we can take expectations over the appropriate marginal distribution) and an analytical form may be available. An example of this case is considered in Section 4.3.

5. As mentioned in Section 3.2.5 there are two sorts of clustered data one might encounter and each gives rise to a different approach to calculating the weights:

   (a) Clusters are iid and will therefore necessarily be of fixed length. This might be the case for some controlled trials within treatment groups or, for weather readings as described in Example II where summary information is being analysed ,ie where we have disposed of issues around short term dependence, per Section 2.6, seasonality and missing data. In that case, we can estimate the RHS of (3.12) simply from the data by using

   $$\sum_{i=1}^{n} \frac{\ln f_C(y_C; \hat{\theta}_n)}{n},$$

   by the law of large numbers, since $\hat{\theta}_n$ is consistent for $\theta_{\mathbf{G}}$ (Section 1.3). We can also calculate a single set of weights that is common to all clusters.

   (b) The clusters vary in length and distribution, as in longitudinal datasets. Here, weights will need to be calculated separately for each cluster. There will thus be only one item of data for the RHS of (3.12) in each calculation and so no averaging is possible. Some assumptions will need to be made about the form of **G** in order for the RHS of (3.12) to be calculated.

6. A reason that we might be interested in the second derivative of D, $\boldsymbol{J}$, is if we calculate the weights as the solution to (3.12) using a numerical approximation scheme such as Newton-Raphson. Defining the vector $\boldsymbol{j}$ as the derivative of D with respect to the weights, then under such an iterative scheme we move from one estimate of the weights, $\boldsymbol{w}^{(i)}$, to the next, $\boldsymbol{w}^{(i+1)}$, by

$$\boldsymbol{w}^{(i+1)} = \boldsymbol{w}^{(i)} - \boldsymbol{J}^{-1}\boldsymbol{j}. \tag{3.18}$$

7. As $\boldsymbol{J}$ is a covariance matrix it is positive semidefinite. If however, it is not positive definite, there will be a linear combination of the $\{\ln f_C(Y_C; \theta_{\mathbf{G}}) : C \in \mathcal{C}\}$ that will equal zero. Consequently, the same linear combination of the equations (3.12) will be zero, leading to more than one set of weights that solve these equations, as we would have more weights than equations. Alternatively, the iteration in (3.18) will not work as $\boldsymbol{J}$ will not be invertible. Examples where this occurs are explored in the next section.

8. Where the composite elements of $\mathbf{F}_K$ are univariate, so that $y_C = y_j$ and

$$K_j = \int_{y_j} f_j(y_j)^{w_j} \mathrm{d}y_j \quad 1 \leq j \leq q$$

say, the LHS of (3.12) can be expanded for all $1 \leq j \leq q$:

$$\begin{aligned} \mathsf{E}_{\mathbf{F}}[lnf_j(y_j)] &= \frac{1}{K}\int_{y_q} \ldots \int_{y_1} lnf_j(y_j) f_1(y_1)^{w_1} \ldots f_q(y_q)^{w_q}\, \mathrm{d}y_1 \ldots \mathrm{d}y_q \\ &= \frac{\prod_{j=1}^{q} K_j}{KK_j} \int_{y_j} lnf_j(y_j) f_j(y_j) w_j \mathrm{d}y_j \end{aligned} \tag{3.19}$$

where all the constants of proportionality, K_j, involve weights. If, in addition, the components of $\mathbf{F}$ are the univariate margins of $\mathbf{G}$, then the RHS of (3.12) can be expanded similarly:

$$\begin{aligned} \mathsf{E}_{\mathbf{G}}[lnf_j(y_j)] &= \int_{y_q} \ldots \int_{y_1} lnf_j(y_j) g(y) \mathrm{d}y_1 \ldots \mathrm{d}y_q \\ &= \int_{y_j} lnf_j(y_j) f_j(y_j) \mathrm{d}y_j. \end{aligned} \tag{3.20}$$

If we select all the weights to be 1 then all the K_js and K will also be 1, and (3.19) and (3.20) will be identical. Thus, in the univariate case, the KLD criterion, (3.12), has a solution with uniform weights. Furthermore, if the covariance matrix of Y is of full rank then so will $\boldsymbol{J}$ be and this solution will be unique.

Components present	$\boldsymbol{J}$ singular
$(y_1, y_2), (y_1, y_3), (y_1, y_4), (y_2, y_3), (y_2, y_4), (y_3, y_4)$	100%
$(y_1, y_3), (y_1, y_4), (y_2, y_3), (y_2, y_4), (y_3, y_4)$	100%
$(y_1, y_3), (y_1, y_4), (y_2, y_3), (y_2, y_4)$	100%
$(y_1, y_4), (y_2, y_3), (y_2, y_4), (y_3, y_4)$	100%
$(y_1, y_4), (y_2, y_3), (y_2, y_4)$	0%
$(y_1, y_4), (y_2, y_4), (y_3, y_4)$	0%
$(y_2, y_3), (y_2, y_4), (y_3, y_4)$	0%

Table 3.1: Percentage of simulations for which $\boldsymbol{J}$ is not invertible in a range of bivariate composite surrogates for data from a multivariate normal distribution with zero means and exchangeable correlation, of order 4.

3.4.3 Uniqueness of Weights

Following on from Note 7 in Section 3.4.2 it is of interest to see whether the non uniqueness of optimal weights, arising from the singularity of $\boldsymbol{J}$, is likely to be an issue in practice. We now examine this issue in more detail. A natural starting point is the multivariate normal distribution.

We work with a more general version of Example IV. Data, (y_1, y_2, y_3, y_4), are generated from a multivariate normal distribution, $\mathbf{G}$, with zero means, dimension $m = 4$ and covariance matrix with variances and exchangeable correlation, resulting in five parameters. We examine a number of bivariate composite surrogates, all unweighted and having the same mean and covariance matrix assumptions as were used to generate the data, in which we vary the number of component pairs. We ran 1000 simulations, each with a randomly generated covariance matrix, of 1000 data items and tested whether the resulting $\boldsymbol{J}$ was singular. The results are given in Table 3.1.

We have used unweighted composite surrogates built from $\{f_C : C \in \mathcal{C}\}$ say. There would be no change if we were to use non zero weighted composite surrogates built from $\{\tilde{f}_C \equiv f_C^{w_C} : C \in \mathcal{C}\}$. The $\boldsymbol{J}$s in the weighted and unweighted cases would be the covariance matrices of $\{\ln f_C\}$ and $\{\ln \tilde{f}_C\}$ respectively and the second is just a full rank transformation of the first (by a diagonal matrix of the weights). The singularity or otherwise of $\boldsymbol{J}$ is thus unaffected by weighting.

The results of the simulations would seem to indicate that, in many cases, such as when all pairs are present, the standard bivariate composite surrogate for the multivariate normal distribution has components which are linear combinations of other components. These correlations become fewer as the number of component pairs are reduced and disappear once we have reached three pairs. Varying the components to be eliminated, for instance the final simulation eliminates all pairs containing y_1, has no effect on the results - it is the number of distinct components that is important.

Define f_{ij} to be the density derived from the distribution that is marginal for **G** for y_i, y_j. Correlation between the $\ln f_{ij}(y_i, y_j)$s is not immediately clear algebraically as, for the full covariance matrix with ith variance σ_i^2 and correlation coefficient ρ,

$$\begin{aligned}\ln f_{ij}(y_i, y_j) &= -ln(2\pi) \\ &\quad -\frac{1}{2}\left(\ln \sigma_i^2 + \ln \sigma_j^2 + \ln(1-\rho^2) + \frac{1}{1-\rho^2}\left(\frac{y_i^2}{\sigma_i^2} + \frac{y_j^2}{\sigma_j^2} - \frac{2\rho y_i y_j}{\sigma_i \sigma_j}\right)\right),\end{aligned}$$

and there is no obvious linear combination of these, as i and j vary, that is zero, A lengthier analysis shows otherwise, and in Appendix F, for example, we prove that the covariance matrix of $\{\ln f_{ij}(Y_i, Y_j; \theta_{\mathbf{G}}) : 1 \leq i, j \leq q\}$, ie for all pairs, under $\mathbf{F}_K$ is indeed singular. Thus, the simulations for $m = 4$, seem to indicate more general results.

The effect of eliminating components in the composite surrogate on parameter estimates and their variances is interesting. Continuing the example described at the start of this section, we ran 1000 simulations, each of 1000 data elements, with data generated from a multivariate normal distribution with means zero, variances $(2.907, 7.719, 5.707, 4.723)$, common correlation 0.244, all chosen randomly (given to 3dp), and calculated the Relative Mean Squared Error (*RMSE*) of individual parameters for bivariate composite surrogates with 6, 5, 4 and 3 components, against parameter estimates from a distribution of the form that generated the data. RMSE is defined analogously to efficiency - it is the MSE of data generating distribution over that of the composite surrogate. For 54 of the 1000 simulations, the nonlinear minimisation routine for **G** did not complete and those results were not taken into account. The pairs eliminated in turn were (y_1, y_3), (y_3, y_4), (y_1, y_2) and (y_2, y_4). A wider range of methods for comparing variances of multiple parameter estimates is described in Section 1.7.2 and explored in practice in

	Number of components				
	6	5	4	3	2
σ_1^2	0.9931	0.9943	0.9900	0.9804	0.9749
σ_2^2	0.9915	0.9868	0.9792	0.9806	0.9733
σ_3^2	0.9943	0.9954	0.9859	0.9841	0.9800
σ_4^2	0.9959	0.9906	0.9928	0.9854	0.9787
ρ	1.0001	0.8626	0.7031	0.5766	0.4327

Table 3.2: Comparison of relative mean squared error for bivariate normal composites with data dimension 4 for decreasing numbers of components.

Section 4.4. The results of the simulations are given in Table 3.2.

Note firstly, that the RMSE with all components present are very close to one. See Section 4.4 for a discussion of this phenomenon, the fact that they are not exactly one resulting from two separate numerical minimisations. As we eliminate components the RMSE for the variances decrease slightly. This presumably results from the fact that the information about the variance parameters mostly exists in the remaining components. However, for the correlation parameter, the RMSE reduces considerably as each component, each of which will contain unique information about the correlation parameter, is eliminated.

3.4.4 Practice

Given the intertwined nature of parameter estimates and weights described in Note 1 in Section 3.4.2, the obvious approach to implementing a weighting scheme is to iterate as follows:

1. With all the weights set to 1, calculate $\hat{\theta}^{(0)}$ in the usual way by solving the estimating equations.
2. Derive a set of weights, $\boldsymbol{w}^{(1)}$, at $\hat{\theta}^{(0)}$, for instance by solving the weights equations (3.12).
3. Calculate $\hat{\theta}^{(1)}$ by solving the estimating equations with weights $\boldsymbol{w}^{(1)}$.

Since both $\hat{\theta}^{(0)}$ and $\hat{\theta}^{(1)}$ are consistent for θ_{G}, the target parameter value, their difference is asymptotically zero. As a consequence, if we consider the Kullback-Leibler difference that we are minimising, $D(\theta, \boldsymbol{w})$ per Theorem 3.4.1, as a function of both the parameters and the weights then, by the continuous mapping theorem described in Section 1.4 (we make the not unreasonable assumption that D is continuous as a function of θ in the area around θ_{G}), $D(\hat{\theta}^{(0)}, \boldsymbol{w}^{(1)}) \to D(\hat{\theta}^{(1)}, \boldsymbol{w}^{(1)})$. Asymptotically, with respect to the weights, $D(\hat{\theta}^{(0)}, \boldsymbol{w})$ and $D(\hat{\theta}^{(1)}, \boldsymbol{w})$ will have the same minimum points (ie at $\boldsymbol{w}^{(1)}$) and thus repeating the iteration described in this Section is unnecessary. Clearly, this will require unique solutions to (3.12), as discussed in Note 7 in Section 3.4.2.

As discussed in Note 5 in Section 3.4.2, if we have iid clusters then we can calculate common weights for all the data whereas if the distributions vary, then we will require separate sets of weights for each cluster.

3.4.5 Simulation II

In order to examine the effect of the optimal weights described in Section 3.4.2 we introduce a simulation in which we compare the power of unweighted and weighted bivariate composite surrogates.

Consider a longitudinal study where each patient has a measurement taken at time zero and then some or no repeat measurements over a period, t_{ij} denoting the jth measurement time for the ith patient. We introduce a random effect which is more closely correlated the nearer in time any two measurements for a particular patient occur. The fixed effects are the same as described in Simulation I in Section 2.7.1. We choose to describe this situation, statistically, as follows.

Data are generated by a probit regression model with a random effect

$$\mathsf{probit}(\mu_{ij}) = \Phi^{-1}(\mu_{ij}) = \eta_{ij} + \Xi_{ij} = \beta_0 + \beta_1 x_{ij\,1} + \beta_2 x_{ij\,2} + \Xi_{ij} \tag{3.21}$$

with binary responses Y_{ij}, $\mu_{ij} = \mathsf{E}[Y_{ij}]$. The first suffix, i, represents a cluster of variable length, $m_i \sim 1 + \mathsf{Poi}(4)$, the second, j, the position in the cluster. The first covariate,

$x_{ij\,1}$, consists of independent realisations of a Bernoulli random variable with mean $0.2 + 0.6t_{ij}$. The second covariate, $x_{ij\,2} = t_{ij}$, is a linear trend with values from U$[0, 1]$, the value for the first item in every cluster being zero, so that we have normalised the period over which patients might be measured to $[0, 1]$. Instance of the random effects, Ξ_{ij}, for cluster i of length m_i, $(\xi_{i1}, \ldots, \xi_{im_i})$, are generated from a multivariate normal distribution with means $\mathbf{0}$, exchangeable variance σ^2 and correlation $\exp(-\alpha|t_{ij} - t_{ik}|)$ for members j and k of cluster i.

We had five parameters and set their values for generating the data as follows

- $\beta_0 = 0.25$ for consistency with Simulation I.
- $\beta_1 = \beta_2$ for 25 values between -0.6 and 0.6 in increments of 0.05 for consistency with Simulation I.
- $\sigma = 0.5$ so that the random effect is large enough to be noticeable but does not swamp the data.
- $\alpha = 2$ so that the correlation belongs to $(0.135, 1]$ and is significant but shows ample variation.

For each of the values of $\beta_1 = \beta_2$ we generated 1000 datasets, each with 100 clusters.

Two surrogates were studied and then compared using power curves over the range of values for $\beta_1 = \beta_2$, against the null hypotheses, $H_0 : \beta_1 = \beta_2 = 0$.

Bivariate unweighted A standard bivariate composite surrogate per (3.1) with all the weights set to 1.

Bivariate weighted A standard bivariate composite surrogate per (3.1) iterated once with weights calculations as described in Section 3.4.4. As the clusters are of variable length, weights were calculated separately for each cluster. The longest cluster for which weights were calculated had length 18. See Note 3 for a discussion on the computational cost of solving the weights equations.

Inference for the surrogates was carried out by a process described in, for instance, Cox and Snell (1989, Section 1.3) and customised for this particular situation. As the outcomes studied here are binary, one could view the probit function in terms of latent variables. Specifically, for element j in cluster i we introduce a latent variable $Z_{ij} \sim \mathsf{N}(0,1)$ and set $Y_{ij} = 1$ if $Z_{ij} < \eta_{ij} + \Xi_{ij}$ etc. This means that

$$\begin{aligned} \Pr(Y_{ij} = 1) &= \Pr(Z_{ij} < \eta_{ij} + \Xi_{ij}) & (3.22) \\ &= \Phi(\eta_{ij} + \Xi_{ij}). & (3.23) \end{aligned}$$

which is equivalent to the description set out in (3.21). When we examine the probability of a bivariate outcome we find that the distribution of the pair of random effects variables (Ξ_{ij}, Ξ_{ik}) is bivariate normal with zero means, common variance σ^2 and correlation $\exp(-\alpha|t_{ij} - t_{ik}|)$. Then, considering a particular outcome:

$$\begin{aligned} &\Pr(Y_{ij} = 1, Y_{ik} = 1) \\ &= \Pr(Z_{ij} < \eta_{ij} + \Xi_{ij}, Z_{ij} < \eta_{ik} + \Xi_{ik}) \\ &= \Pr(Z_{ij} - \Xi_{ij} < \eta_{ij}, Z_{ij} - \Xi_{ik} < \eta_{ik}) \\ &= Pr(Z^*_{ij} < \eta_{ij}, Z^*_{ik} < \eta_{ik})) \end{aligned}$$

where we define $Z^*_{ij} = Z_{ij} - \Xi_{ij}$ etc, so that

$$\begin{pmatrix} Z^*_{ij} \\ Z^*_{ik} \end{pmatrix} \sim \mathsf{BVN}(\mathbf{0}, \mathbf{\Sigma}_B)$$

and

$$\mathbf{\Sigma}_B = \begin{pmatrix} 1+\sigma^2 & \sigma^2 \exp(-\alpha|t_{ij} - t_{ik}|) \\ \sigma^2 \exp(-\alpha|t_{ij} - t_{ik}|) & 1+\sigma^2 \end{pmatrix}$$

where $Z^*_{ij}/(1+\sigma^2)^{\frac{1}{2}}$ has a standard normal distribution. In a similar fashion, as partially described in, say Ashford and Sowden (1970),

$$\begin{aligned}
\Pr(Y_{ij}=1, Y_{ik}=0) &= \Pr(Z^*_{ij} < \eta_{ij}) - \Pr(Z^*_{ij} < \eta_{ij}, Z^*_{ik} < \eta_{ik}) \\
\Pr(Y_{ij}=0, Y_{ik}=1) &= \Pr(Z^*_{ik} < \eta_{ik}) - \Pr(Z^*_{ij} < \eta_{ij}, Z^*_{ik} < \eta_{ik}) \\
\Pr(Y_{ij}=0, Y_{ik}=0) &= \Pr(Z^*_{ij} < -\eta_{ij}, Z^*_{ik} < -\eta_{ik}).
\end{aligned}$$

The bivariate joint probabilities are all thus expressed in terms of cumulative bivariate normal densities which are easily calculable. We used a routine based on an algorithm in Donnelly (1973).

Maximisation of loglikelihoods for estimating parameters and minimisation of the KLD for calculating weights were carried out using the nonlinear minimisation function 'nlm' in R (R Development Core Team, 2012). As the objective function for parameter estimation appeared to be very unstable, initial estimates were calculated using Nelder-Mead optimisation via the function 'optim' in R (R Development Core Team, 2012).

Univariate only surrogates were not examined as the parameters are not all identifiable. For a univariate outcome

$$\begin{aligned}
\Pr(Y_{ij}=1) &= \Pr(Z_{ij} < \eta_{ij} + \xi_{ij}) \quad \text{from (3.22)} \\
&= \Pr(Z^*_{ij} < \eta_{ij}) \\
&= \Phi(\eta_{ij}/(1-\sigma^2)^{\frac{1}{2}}) \\
&= \Phi\left(\frac{\beta_0 + \beta_1 x_{ij\,1} + \beta_2 x_{ij\,2}}{(1-\sigma^2)^{\frac{1}{2}}}\right),
\end{aligned}$$

the penultimate step arising from the distribution of $Z^*_{ij}/(1+\sigma^2)^{\frac{1}{2}}$. Minimising a loglikelihood consisting of sums of these types of terms, one cannot differentiate between changes in, for instance, σ^2 and $\{\beta_i : i = 0, 1, 2\}$ for a given set of data, and so the full set of parameters is not identifiable.

The resulting power curves are shown in Figure 3.1. Examining the plot reveals no significant differences between the two surrogates as at different points on the steeper parts of the curves, each of them has the greatest power. At the minimum points of

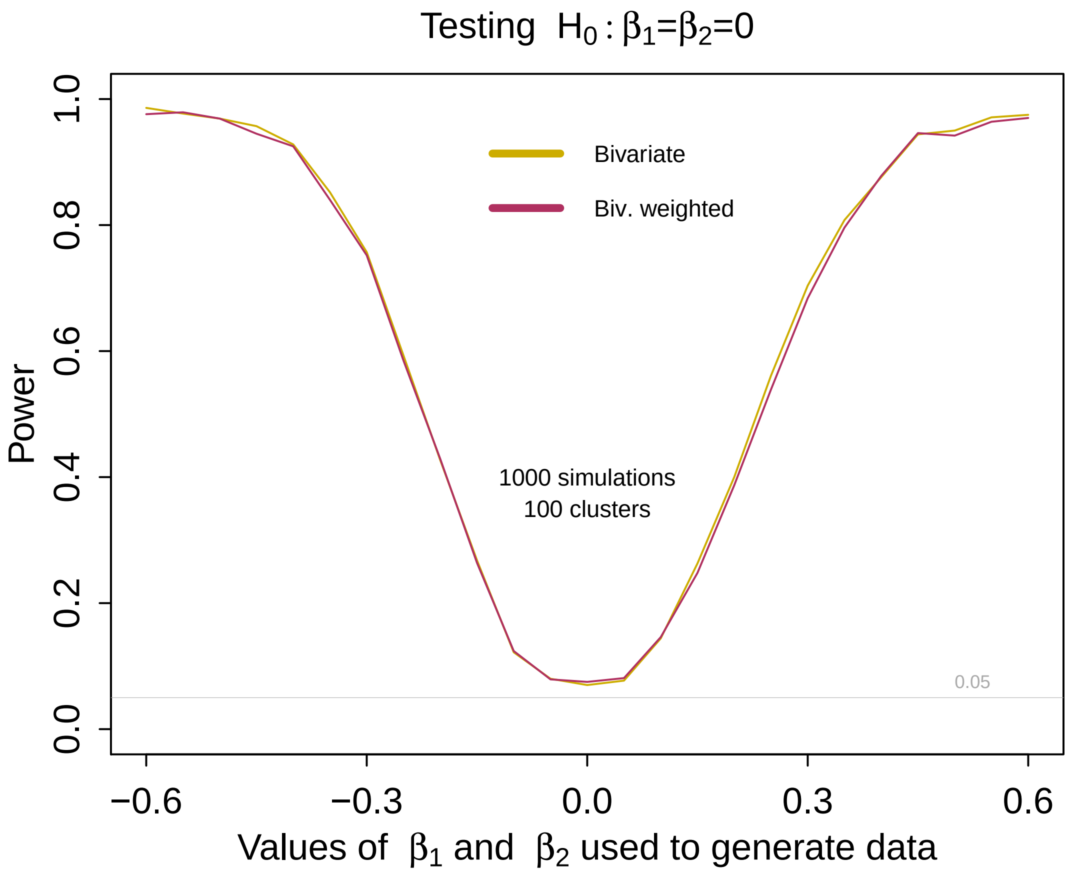

Figure 3.1: Simulation II. Power plots at level 0.05 for a bivariate and weighted bivariate surrogates.

the curves, where one would like values as close as possible to 0.05, the unweighted surrogate has value 0.080 and the weighted, 0.079.

The data which have been used to generate the power curves are the p-values of the hypothesis that $H_0 : \beta_1 = \beta_2 = 0$ for 1000 simulations of each of the 25 values of $\beta_1 = \beta_2$ used to generate the data. For a range of those values $(-0.35, -0.2, 0.0.2, 0.35)$, the resulting p-values were compared to see whether any significant differences emerged. The data were transformed to be approximately normal and then compared using paired t-tests. Again, no consistent patterns emerged and the results reflected the patterns of the power curves in Figure 3.1.

Finally, unlike the multivariate normal example described in Section 3.4.3, the weights that are calculated for the bivariate weighted surrogate are unique, ie the matrix $\boldsymbol{J}$ discussed in Point 7 of Section 3.4.2 is invertible. Further surrogates are examined in Section 4.7.1.

3.5 Summary

In this chapter we have reviewed the existing literature on weighting components of composite surrogates. We have suggested a more generally applicable version of one such optimal scheme but it is still expensive computationally to implement. We have also suggested a restricted version of that scheme that is optimal within its class. We have then derived, analytically, equations to be solved to give optimal scalar weights, on the basis of minimising the Kullback-Leibler discrepancy between $\mathbf{G}$ and the weighted composite surrogate, taking into account the constant of proportionality for the surrogate. The resulting weights may not be unique. We have explored how to calculate these weights in practice and then used them in a simulation. The resulting power shows no significant improvement once the weights are taken into account. In order to better understand the effect and structure of the weights, in Chapter 4 we analyse the scalar weighting scheme for multivariate normal composite surrogates.

Chapter 4

Multivariate Normal Composite Surrogates

4.1 Introduction

In Section 3.4 we presented some equations, derived from minimising the Kullback-Leibler Divergence (KLD) between the distribution that generated the data, $\mathbf{G}$, and a composite surrogate with constant of proportionality, $\mathbf{F}_K$, that can be solved to calculate the scalar weights for the surrogate, namely

$$\mathsf{E}_{\mathbf{F}_K}[\ln f_C(Y_C;\theta_{\mathbf{G}})] = \mathsf{E}_{\mathbf{G}}[\ln f_C(Y_C;\theta_{\mathbf{G}})] \qquad C \in \mathcal{C} \tag{4.1}$$

where f_C is the component of $\mathbf{F}$, the surrogate without the constant of proportionality, acting on Y_C, a subset of Y and $\mathcal{C}$ is the set of all subsets of $1, \ldots .m$ used in the composite. The analytical form of those equations is generally complex and not always tractable. In this section we present an example, the multivariate normal (MVN), where (4.1) can be simply expressed. We begin, in Section 4.2, with a review of the use of weighted multivariate normal composite surrogates from the literature. In Section 4.3 we derive the MVN version of (4.1) for calculating weights. One of the consequences of the theory set out in Section 4.3 is that MVN composite surrogates can be considered as data transforms and this aspect is considered in Section 4.4. We examine examples of normal composite surrogates - univariate, bivariate and combined univariate and bivariate - in Sections 4.5 to 4.7. We show how in the combined case, one can recover the original data generating distribution for particular choices of weights. We apply that lesson to the simulation from Chapter 3 in Section 4.7.1 and the use of MVN composites as surrogate for data generated from autoregressive distributions in Section 4.8.

In a number of places in this Chapter we shall derive a distribution whose density is proportional to the exponential of a multivariate quadratic form. Clearly, the distribution will then be multivariate normal with the constant of proportionality calculated according to the standard formulation. A specific condition for finiteness of the CoP forms part of the main result - Theorem 4.3.1.

4.2 Literature Review

As we have seen, there are very few published papers that examine the use of weighted composite surrogates, although many of those go on to consider multivariate normal distributions explicitly. Lindsay et al. (2011) examine two simple examples where the data arise from a multivariate normal distribution with zero means, and covariance matrices with exchangeable variances, σ^2, and correlations, ρ, for a variety of data dimensions, d. Each of the two examples treats one of the parameters as known and one as unknown. The five methods compared in each case are:

1. Unweighted bivariate composite consisting of all pairs.
2. The composite surrogate formed from all conditional densities between pairs of observations. We discussed this in Section 3.2.3 where it is reformulated as a weighted sum of univariate and bivariate marginals.
3. The second order Hoeffding score (ie bivariate and univariate) described in Section 3.2.2.
4. The modified second order Hoeffding score described in Section 3.2.4.
5. MLE.

The simplicity of the parameter schemes mean that we just have to consider component type weights, as opposed to weights for every component, as set out in Section 3.2.3. Investigations are carried out, theoretically, to see whether the surrogate can recover the distribution that generated the data (MLE) and by simulation to investigate efficiency.

For the case where the variance is the unknown parameter, only the conditional and modified Hoeffding methods recover the MLE for specific, albeit different, relationships between ρ and d. For large d, a large ρ is also required. Simulations show that as d increases, the conditionals method is the most efficient (50% for $d = 50$) followed by the two Hoeffding approaches.

For the case where the correlation is the unknown parameter, recovering the MLE is not considered in Lindsay et al. (2011) due to algebraic complexity. Simulations show that as d increases, the modified Hoeffding method is most efficient (30% for $d = 50$) followed by the conditionals approach.

Obviously, these examples are fairly rudimentary and, in particular, they only address scalar unknown parameters. For large d, efficiency is not high. More sophisticated techniques set out in Joe and Lee (2009) are discussed in Section 3.2.5.

4.3 Constant of Proportionality Weights

As we remarked in Note 3 to Theorem 3.4.1, the KLD minimising equations, (3.12), that result in weights can be messy and will usually have to be solved using some form of Newton-Raphson approach, such as 'nlm' in R (R Development Core Team, 2012). However, for the multivariate normal case, elegant analytical solutions exist. The core theorem proved in this Section drives the results from the rest of the chapter. In addition, in contrast to some of the results described in Section 4.2, the weights can be derived analytically.

Theorem 4.3.1. *Consider a distribution $\boldsymbol{G}$ of random variables $\boldsymbol{Y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, of dimension m, where $\boldsymbol{\Sigma}$ is positive definite. Let $f_C(y_C)$, μ_C and Σ_C represent the density and components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively for the marginal distribution of Y_C,* *of dimension c, corresponding to subset C of* $1, \ldots, m$*. Define a weighted composite surrogate constructed from the densities* $\{f_C(y_C) : C \in \mathcal{C}\}$ *and denote by* $\boldsymbol{F}_K$ *the distribution corresponding to that surrogate with a constant of proportionality as described in Section 2.5. Then* $\boldsymbol{F}_K$ *is a multivariate normal distribution with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma_F}$ *say where*

$$\boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1} = \sum_{C \in \mathcal{C}} w_C \boldsymbol{A}_C^T (\boldsymbol{A}_C \boldsymbol{\Sigma} \boldsymbol{A}_C^T)^{-1} \boldsymbol{A}_C$$

and $\boldsymbol{A}_C$ *is a* $c \times m$ *matrix with a single 1 per row, corresponding to the marginal locations, and 0s elsewhere.*

Moreover, denoting by $\Sigma_{\boldsymbol{F}_C}$ *the covariance matrix for index subset C under* $\boldsymbol{F}_K$*, equations (3.12) in the statement of Theorem 3.4.1 become*

$$tr(\Sigma_C^{-1} \Sigma_{\boldsymbol{F}_C}) = c, \tag{4.2}$$

where Σ_C^{-1} *is the matrix inverse of* Σ_C*, and the second derivative matrix defined in equations (3.13) is*

$$Cov_{\boldsymbol{F}_K}[\ln(f_{C_i}(Y_{C_i}), \ln(f_{C_j}(Y_{C_j})] = \frac{1}{2} tr((\Sigma^{-1})_{C_i} \Sigma_{\boldsymbol{F}_{C_i}} (\Sigma^{-1})_{C_j} \Sigma_{\boldsymbol{F}_{C_j}}). \tag{4.3}$$

Proof From the standard properties of the multivariate normal distribution, we have for any subset C of $\boldsymbol{Y}$ that

$$f_C(y_C) = (2\pi)^{-\frac{c}{2}} |\Sigma_C|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y_C - \mu_C)^T \Sigma_C^{-1} (-\frac{1}{2}(y_C - \mu_C)\right). \tag{4.4}$$

Note that we can write $Y_C = \boldsymbol{A}_C \boldsymbol{Y}$, $\mu_C = \boldsymbol{A}_C \boldsymbol{\mu}$ and $\Sigma_C = \boldsymbol{A}_C \boldsymbol{\Sigma} \boldsymbol{A}_C^T$. The surrogate distribution F_K, as defined in Section 2.5, then has density

$$\begin{aligned} f(\boldsymbol{y}) &\propto \prod_{C\in\mathcal{C}} f_C(y_C) \\ &\propto \exp\left(-\frac{1}{2}\sum_{C\in\mathcal{C}} w_C(y_C-\mu_C)^T\Sigma_C^{-1}(y_C-\mu_C)\right) \\ &= \exp\left(-\frac{1}{2}\sum_{C\in\mathcal{C}} w_C\left(\boldsymbol{A}_C(\boldsymbol{y}-\boldsymbol{\mu})\right)^T\left(\boldsymbol{A}_C\boldsymbol{\Sigma}\boldsymbol{A}_C^T\right)^{-1}\left(\boldsymbol{A}_C(\boldsymbol{y}-\boldsymbol{\mu})\right)\right) \\ &= \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\left(\sum_{C\in\mathcal{C}} w_C\boldsymbol{A}_C^T(\boldsymbol{A}_C\boldsymbol{\Sigma}\boldsymbol{A}_C^T)^{-1}\boldsymbol{A}_C\right)(\boldsymbol{y}-\boldsymbol{\mu})\right) \end{aligned}$$

so that once we take into account the constant of proportionality

$$\mathsf{F}_K \equiv \mathsf{MVN}\left(\mu, \left(\sum_{C\in\mathcal{C}} w_C\boldsymbol{A}_C^T(\boldsymbol{A}_C\boldsymbol{\Sigma}\boldsymbol{A}_C^T)^{-1}\boldsymbol{A}_C\right)^{-1}\right) \equiv \mathsf{MVN}(\mu, \boldsymbol{\Sigma}_{\mathsf{F}}). \tag{4.5}$$

Since, we have specified that $\boldsymbol{\Sigma}$ is positive definite then so must $\boldsymbol{\Sigma}_{\mathsf{F}}$ be. This proves the first part of the theorem. Next, from (4.4) we have that

$$\begin{aligned} \mathsf{E}_{\mathsf{F}_K}\left[\ln(f_C(Y_C)\right] &= \mathsf{E}_{\mathsf{F}_K}\left[-\frac{1}{2}(Y_C-\mu_C)^T\Sigma_C^{-1}(Y_C-\mu_C)\right] \\ &- \frac{c}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_C|. \end{aligned} \tag{4.6}$$

There is a general result that

$$\mathsf{E}[\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}] = \mathsf{tr}(\boldsymbol{\Lambda}\boldsymbol{A}) \tag{4.7}$$

for any $\boldsymbol{Z} \sim \mathsf{MVN}(0, \boldsymbol{\Lambda})$ and appropriately dimensioned $\boldsymbol{A}$ (see, for instance, Schott, 1997, Theorem 9.18(a)). Applying this to (4.6) gives

$$\begin{aligned} \mathsf{E}_{\mathsf{F}_K}\left[\ln(f_C(Y_C)\right] &= -\left(\frac{1}{2}\mathsf{tr}(\Sigma_{\mathsf{F}_C}\Sigma_C^{-1}) + \frac{c}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma_C|\right) \\ &= -\left(\frac{1}{2}\mathsf{tr}(\Sigma_C^{-1}\Sigma_{\mathsf{F}_C}) + \frac{c}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma_C|\right) \end{aligned}$$

as tr$(\boldsymbol{AB})$ = tr$(\boldsymbol{BA})$. Now

$$\begin{aligned}
\mathsf{E}_{\mathbf{G}}[\ln(f_C(Y_C)] &= \mathsf{E}_{\mathbf{G}}\left[-\frac{1}{2}(Y_C-\mu_C)^T\Sigma_C^{-1}(Y_C-\mu_C)\right]-\frac{c}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma_C| \\
&= -\left(\frac{1}{2}\mathsf{tr}(\Sigma_C\Sigma_C^{-1})+\frac{c}{2}\ln(2\pi)+\frac{1}{2}\ln|\Sigma_C|\right) \qquad \text{by (4.7)} \\
&= -\left(\frac{c}{2}+\frac{c}{2}\ln(2\pi)+\frac{1}{2}\ln|\Sigma_C|\right)
\end{aligned}$$

and we have proved the second part of the theorem, namely

$$\mathsf{tr}(\Sigma_C^{-1}\Sigma_{\mathbf{F}_C}) = c. \tag{4.8}$$

Finally, we note that Cov$[\boldsymbol{Z}^T\boldsymbol{AZ},\boldsymbol{Z}^T\boldsymbol{BZ}] = 2tr(\boldsymbol{A\Lambda B\Lambda})$ where $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric and $\boldsymbol{Z}\sim$ MVN$(\mathbf{0},\boldsymbol{\Lambda})$ (Schott, 1997, Theorem 9.21(b)). The proof of that result is easily extended to distinct random variables, $\boldsymbol{Z}_i\sim$ MVN$(\mathbf{0},\boldsymbol{\Lambda}_i)$ and $\boldsymbol{Z}_j\sim$ MVN$(\mathbf{0},\boldsymbol{\Lambda}_j)$, so that Cov$[\boldsymbol{Z}_i^T\boldsymbol{AZ}_i,\boldsymbol{Z}_j^T\boldsymbol{BZ}_j] = 2tr(\boldsymbol{A\Lambda}_i\boldsymbol{B\Lambda}_j)$. Therefore

$$\begin{aligned}
&\mathsf{Cov}_{\mathbf{F}_K}[\ln(f_C(Y_{C_i}),\ln(f_C(Y_{C_j})] \\
&= \mathsf{Cov}_{\mathbf{F}_K}\left[-\frac{1}{2}(Y_{C_i}-\mu_{C_i})^T\Sigma_{C_i}^{-1}(Y_{C_i}-\mu_{C_i}),-\frac{1}{2}(Y_{C_j}-\mu_{C_j})^T\Sigma_{C_j}^{-1}(Y_{C_j}-\mu_{C_j})\right] \\
&= \frac{1}{2}\mathsf{tr}(\Sigma_{C_i}^{-1}\Sigma_{\mathbf{F}_{C_i}}\Sigma_{C_j}^{-1}\Sigma_{\mathbf{F}_{C_j}})
\end{aligned}$$

completing our theorem. □

There are a number of points worth making about this result:

1. To evaluate our weights we solve equations involving the ratio of covariance matrices from the distribution that generated the data and our composite surrogate. This makes some sort of sense as with our surrogate we are trying to recover features of the distribution that generated the data. For an example, see Section 4.7.

2. It requires knowledge of **G**. For an example of where we use an MVN composite as a surrogate for a distribution that is not MVN, see Section 4.8.

3. In practice, the result has limitations. We need to calculate $\Sigma_{\mathbf{F}_C}$ for all $C \in \mathcal{C}$. From (4.5), we can see that $\Sigma_{\mathbf{F}}^{-1}$ is easily derived, but it is its inverse that we require and so, to calculate optimal weights, in general, we need to invert an $m \times m$ matrix.

4. $\boldsymbol{J}$, the second derivative of the KLD, D, with respect to the weights is the covariance matrix of $\{\ln f_C(Y_C; \theta_{\mathbf{G}}) : C \in \mathcal{C}\}$ under $\mathbf{F}_K$ not $\mathbf{G}$. The covariance matrix under the latter is temptingly easy to calculate (the diagonals are all 1 for instance), this is not the case under $\mathbf{F}_K$ - the distribution is multivariate normal but its covariance matrix is complex. We explore this, amongst other things, for particular types of components in Sections 4.5 to 4.7.

In Section 4.4 we explore a consequence of part of Theorem 4.3.1. In Sections 4.5 to 4.7 we review some specific examples of composite surrogates - univariate, bivariate and combined univariate and bivariate in order to examine the optimal weights arising from solving (4.2), and to understand whether, by a suitable choice of weights we can recover $\mathbf{G}$. For simplicity we assume zero means, and take $\boldsymbol{\Sigma}$ to have variances σ_i^2 and correlations ρ_{ij} for $1 \leq i, j \leq m$. We define $\boldsymbol{R}$ to be the corresponding correlation matrix.

4.4 Composite Surrogates are Transforms

This section examines certain features of composite surrogates, irrespective of whether they are weighted. We shall work with multivariate normal distributions with zero means to simplify notation, but the results extend relatively easily to uncentred distributions. A corollary of Theorem 4.3.1 where we saw that if $\mathbf{G}$ is multivariate normal, then so is $\mathbf{F}_K$, is the following

Corollary 4.4.1. *Let $\boldsymbol{G}$ be a distribution of random variables $\boldsymbol{Y} \sim MVN(\mathbf{0}, \Sigma)$, with dimension $m > 1$, and $\boldsymbol{F}_K$ a composite surrogate, with constant of proportionality, whose components are marginal for $\boldsymbol{G}$, describing random variables $\boldsymbol{Z}$.*

Then, if $\mathbf{\Sigma}$ and $\mathbf{\Sigma_F}$, as defined in Theorem 4.3.1, are positive definite, $\boldsymbol{Z}$ is a linear transformation of $\boldsymbol{Y}$, ie $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{Y}$ for some $m \times m$ matrix $\boldsymbol{A}$ of full rank.

Proof We can see from Theorem 4.3.1 that the distribution of $\boldsymbol{Z}$ will be multivariate normal, say MVN$(\mathbf{0}, \mathbf{\Sigma}_{\mathsf{F}})$. Define $\boldsymbol{M}$ and $\boldsymbol{M}_{\mathsf{F}}$ such that $\boldsymbol{M}\boldsymbol{M}^T = \mathbf{\Sigma}$ and $\boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}_{\mathsf{F}}^T = \mathbf{\Sigma}_{\mathsf{F}}$ and consider the distribution of $\boldsymbol{X} = \boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}^{-1}\boldsymbol{Y}$. Since $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{\mathsf{F}}$ are positive definite, $\boldsymbol{M}$, $\boldsymbol{M}_{\mathsf{F}}$ and $\boldsymbol{M}^{-1}$ must all exist (there is, for instance, a unique positive definite square root per Horn and Johnson, 1987, Theorem 7.2.6). Also, we can see that $\boldsymbol{X} \sim$ MVN$(\mathbf{0}, \boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}^{-1}\mathbf{\Sigma}(\boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}^{-1})^T)$ (eg Krzanowski, 2000, page 205). But

$$\begin{aligned}
\boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}^{-1}\mathbf{\Sigma}(\boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}^{-1})^T) &= \boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}^{-1}\mathbf{\Sigma}(\boldsymbol{M}^{-1})^T\boldsymbol{M}_{\mathsf{F}}^T \\
&= \boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}^{-1}\boldsymbol{M}\boldsymbol{M}^T(\boldsymbol{M}^{-1})^T\boldsymbol{M}_{\mathsf{F}}^T \\
&= \boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}_{\mathsf{F}}^T \\
&= \mathbf{\Sigma}_{\mathsf{F}}
\end{aligned}$$

so that $\boldsymbol{X} = \boldsymbol{Z}$ and we have shown that the distribution of the composite surrogate is just that of a transformation of the random variables under consideration, with the transformation matrix

$$\boldsymbol{A} = \boldsymbol{M}_{\mathsf{F}}\boldsymbol{M}^{-1}$$

where $\boldsymbol{A}$ is of full rank as $\boldsymbol{M}_{\mathsf{F}}$ and $\boldsymbol{M}^{-1}$ are positive definite. □

$\boldsymbol{M}$ and $\boldsymbol{M}_z$ could also be Cholesky roots. We now compare inference about the parameters contained in $\mathbf{\Sigma}$ for the distributions of $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Rather than describing the parameters as a vector, we shall use the matrix $\mathbf{\Sigma}$ and for parameter estimates, $\hat{\mathbf{\Sigma}}$.

Theorem 4.4.1. *With the terminology of Corollary 4.4.1, if there is a full set of distinct parameters in $\mathbf{\Sigma}$, then the parameter estimator $\hat{\mathbf{\Sigma}}$, under maximum likelihood estimation, is identical whether one analyses the distributions of $\boldsymbol{Y}$ or $\boldsymbol{Z}$.*

Proof For data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, so that $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{y}_i$, $1 \leq i \leq n$, we define

$$\begin{aligned} \boldsymbol{S} &= \sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T \quad \text{and} \\ \boldsymbol{T} &= \sum_{i=1}^{n} (\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})^T \end{aligned}$$

where $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{z}}$ are the means of the respective datasets. It is a standard result (see, for instance, Kotz et al. (2000, page 161)) that the maximum likelihood estimator, $\hat{\boldsymbol{\Sigma}}$, is $\boldsymbol{S}/n$. Let $\boldsymbol{\Lambda} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$ (for a full set of parameters) and we saw in Corollary 4.4.1 that $\boldsymbol{Z} \sim$ MVN$(\boldsymbol{0}, \boldsymbol{\Lambda})$. Again, we can carry out maximum likelihood estimation for the parameters in $\boldsymbol{\Lambda}$ with data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ so that the estimator, $\hat{\boldsymbol{\Lambda}}$, is $\boldsymbol{T}/n$.

From the definition of $\boldsymbol{\Lambda}$, an estimator for $\boldsymbol{\Sigma}$, $\tilde{\boldsymbol{\Sigma}}$, can be calculated from $\hat{\boldsymbol{\Lambda}} = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}\boldsymbol{A}^T$ so that, as $\boldsymbol{A}$ is defined as being of full rank

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}} &= \boldsymbol{A}^{-1}\hat{\boldsymbol{\Lambda}}(\boldsymbol{A}^T)^{-1} \\ &= \boldsymbol{A}^{-1}\boldsymbol{T}(\boldsymbol{A}^T)^{-1}/n \\ &= \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T(\boldsymbol{A}^T)^{-1}/n \\ &= \boldsymbol{S}/n \\ &= \hat{\boldsymbol{\Sigma}} \end{aligned}$$

and inference about the parameters in $\boldsymbol{\Sigma}$ is identical whether one analyses the distributions of $\boldsymbol{Y}$ or $\boldsymbol{Z}$. □

Thus, we can treat a composite surrogate of a multivariate normal distribution as if it were a transformation of the random variables. This might provide a hint as to why composite surrogates are effective in general (ie not just for multivariate normal distributions) and will be the subject of further work. It is worth noting that

1. The fact that $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}/n$ as the maximum likelihood estimator only holds as $\boldsymbol{Y} \sim$ MVN$(\boldsymbol{0}, \boldsymbol{\Sigma})$. A similar argument applies to Λ, $\boldsymbol{T}$ and $\boldsymbol{Z}$.

2. Taking Corollary 4.4.1 and Theorem 4.4.1 together, we see that, subject to the

conditions outlined there, for data generated from a multivariate normal distribution, equivalent inference about the parameters of that distribution can be gained by carrying out maximum likelihood estimation either on the originating distribution itself or on a composite surrogate with constant of proportionality. We can only do the latter estimation as, from the previous point, the composite surrogate is the distribution of known transformed data, $\boldsymbol{Z}$

3. In Section 3.4.3 we saw that for multivariate normal data with an exchangeable correlation coefficient, the first and second moments of the parameter estimators from the bivariate composite surrogate do indeed appear to match those from the data generating distribution. Note there, however, that no constant of proportionality was involved. Also, there was not a full set of parameters as required in Theorem 4.4.1.

4. In the worked example in Section 2.5, we noted that Cox and Reid (2004) gave the example of a bivariate standard multivariate normal surrogate with an exchangeable correlation coefficient where the parameter estimates became less efficient as the number of composite components (ie, the length of the vector $\boldsymbol{Y}$) grew. This example does not contravene Theorem 4.4.1 as there is not a full set of distinct parameters. As a consequence, in this case $\hat{\boldsymbol{\Sigma}} \neq \boldsymbol{S}/n$ as, for instance, $\boldsymbol{S}/n$ will not have 1s down the diagonal. Also, as with the previous point, no constant of proportionality was involved, but unlike that point, inference is not maintained with the composite surrogate.

Resolution of the apparent tension between these points and the extent to which Theorem 4.4.1 might explain the effectiveness of the composite approach will be the subject of future work.

4.5 Univariate Margins

We first consider the case when all the composite components are univariate margins of **G**. So, $\mathcal{C}$ will consist of sets each containing an individual element from $\{1, \ldots, m\}$.

For instance, if we take $C = \{i\}$, some i, then using the terminology of Theorem 4.3.1

$$(\boldsymbol{A}_i \boldsymbol{\Sigma} \boldsymbol{A}_i^T)^{-1} = \sigma_i^{-2}$$

and $\boldsymbol{A}_i^T (\boldsymbol{A}_i \boldsymbol{\Sigma} \boldsymbol{A}_i^T)^{-1} \boldsymbol{A}_i$ has one non zero element, σ_i^{-2}, in the ith entry of the main diagonal and zeroes elsewhere. $\mathbf{F}_K$ is the resulting weighted univariate composite surrogate with constant of proportionality, whose covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{F}}^{-1} = \sum_{C \in \mathcal{C}} w_C \boldsymbol{A}_C^T (\boldsymbol{A}_C \boldsymbol{\Sigma} \boldsymbol{A}_C^T)^{-1} \boldsymbol{A}_C,$$

has ith diagonal element

$$\frac{w_{ii}}{\sigma_i^2}$$

where w_{ii} denotes the weight for the ith margin (the double subscript is used for consistency with material in Section 4.7) and zeroes elsewhere.

The KLD weights equations, (4.2) become

$$\Sigma_C^{-1} \Sigma_{\mathbf{F}_C} = 1 \qquad C \in \mathcal{C},$$

where $\Sigma_C = \Sigma_{(i,i)}$ etc, which results in:

$$\begin{aligned} w_{ii} &= \sigma_i^2 \boldsymbol{\Sigma}_{(i,i)}^{-1} \\ &= 1 \end{aligned} \tag{4.9}$$

in agreement with Note 8 of Section 3.4.2. Thus, for these optimal weights

$$\boldsymbol{\Sigma}_{\mathbf{F}}^{-1} = \begin{pmatrix} \sigma_1^{-2} & \ldots & 0 & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & & & \vdots \\ 0 & \ldots & 0 & \ldots & \sigma_m^{-2} \end{pmatrix}$$

which is $\mathbf{\Sigma}^{-1}$ with zeroes replacing all the off diagonal elements, and

$$\mathbf{\Sigma}_{\mathsf{F}} = \begin{pmatrix} \sigma_1^2 & \dots & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & 0 & \dots & \sigma_m^2 \end{pmatrix}$$

which is $\mathbf{\Sigma}^{-1}$ with zeroes replacing all the off diagonal elements. We have not recovered $\mathbf{\Sigma}$ as we have only considered univariate margins, but we have made the most of the information in those margins.

We now examine $\boldsymbol{J}$, the second derivative with respect to the weights of the KLD between **G** and $\mathbf{F}_K$. The composite surrogate has m components and weights so $\boldsymbol{J}$ will be an $m \times m$ matrix with (i,j)th entry, by (4.3),

$$\begin{aligned} \boldsymbol{J}_{(i,j)} &= \frac{1}{2}\mathrm{tr}(\Sigma_{C_i}^{-1}\Sigma_{\mathsf{F}_{C_i}}\Sigma_{C_j}^{-1}\Sigma_{\mathsf{F}_{C_j}}) \\ &= \frac{1}{2}\Sigma_{(i,i)}^{-1}\frac{\sigma_i^2}{w_{ii}}\Sigma_{(j,j)}^{-1}\frac{\sigma_j^2}{w_{jj}} \\ &= \frac{\sigma_i^2\sigma_j^2\Sigma_{(i,i)}^{-1}\Sigma_{(j,j)}^{-1}}{2w_{ii}w_{jj}} \end{aligned}$$

and substituting our optimal weights we have

$$\boldsymbol{J}_{(i,j)} = \frac{1}{2}.$$

$\boldsymbol{J}$ is thus singular and so, as described in Note 7 to Theorem 3.4.1, numerical approximation schemes for calculating the weights will not work. Fortunately, they are unnecessary as we have an analytical solution, (4.9).

4.6 Bivariate Margins

4.6.1 Main Results

We now review the case when all our composite components are bivariate marginals of **G**. $\mathcal{C}$ will then consist of the $q = m(m-1)/2$ pairs of distinct elements in $\{1, \dots, m\}$. For instance, if $C = \{i, j\} \quad (i < j)$, then, using the terminology of Theorem 4.3.1

$$(\boldsymbol{A}_{ij}\boldsymbol{\Sigma}\boldsymbol{A}_{ij}^T)^{-1} = \frac{1}{1-\rho_{ij}^2}\begin{pmatrix} \sigma_i^{-2} & -\rho_{ij}/\sigma_i\sigma_j \\ -\rho_{ij}/\sigma_i\sigma_j & \sigma_j^{-2} \end{pmatrix}$$

so that

$$\boldsymbol{A}_{ij}^T(\boldsymbol{A}_{ij}\boldsymbol{\Sigma}\boldsymbol{A}_{ij}^T)^{-1}\boldsymbol{A}_{ij} = \frac{1}{1-\rho_{ij}^2}\begin{pmatrix} 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & \sigma_i^{-2} & \dots & -\rho_{ij}/\sigma_i\sigma_j & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & -\rho_{ij}/\sigma_i\sigma_j & \dots & \sigma_j^{-2} & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 0 \end{pmatrix}$$

with non zero entries only in the ith and jth rows and columns. As a consequence, for $\mathbf{F}_K$, the bivariate composite surrogate with constant of proportionality, $\boldsymbol{\Sigma}_{\mathbf{F}}^{-1} = \sum_{C \in \mathcal{C}} w_C \boldsymbol{A}_C^T(\boldsymbol{A}_C\boldsymbol{\Sigma}\boldsymbol{A}_C^T)^{-1}\boldsymbol{A}_C$ will have off diagonal (i, j) element

$$\frac{-w_{ij}\rho_{ij}}{\sigma_i\sigma_j(1-\rho_{ij}^2)}$$

and ith diagonal element

$$\frac{1}{\sigma_i^2}\sum_{k \neq i}\frac{w_{ik}}{1-\rho_{ik}^2}.$$

Note that

$$\mathbf{\Sigma}_{\mathsf{F}}^{-1} = \boldsymbol{D}(\sigma)^{-1}\boldsymbol{A}^{-1}\boldsymbol{D}(\sigma)^{-1} \tag{4.10}$$

where $\boldsymbol{D}(\sigma)$ is the $m \times m$ diagonal matrix whose entries are $\sigma_1, \ldots, \sigma_m$ and $\boldsymbol{A}^{-1}$ is an $m \times m$ matrix with all entries being functions of the ρ_{ij}s. For the exchangeable correlation case with equal weights (ie unweighted), $\boldsymbol{A}$ will have diagonal entries that are identical and off diagonal entries that are also identical, all entries being functions of ρ but not the σ_i^2s. This formulation is used in Appendix F.

If we use the weights equations (4.2) to calculate weights, we have

$$\mathrm{tr}(\Sigma_C^{-1}\mathbf{\Sigma}_{\mathsf{F}_C}) \;=\; 2 \qquad C \in \mathcal{C}.$$

At first sight, unlike the univariate case, there appears to be no simple algebraic form for any solution to these equations. However, if

$$\begin{aligned} \mathbf{\Sigma}_{\mathsf{F}_C}^{-1}\mathbf{\Sigma}_C &= \boldsymbol{I}_{2\times2} && \text{; or, equivalently} \\ \mathbf{\Sigma}_{\mathsf{F}_C}^{-1} &= (\Sigma^{-1})_C \end{aligned}$$

for all $C \in \mathcal{C}$ then (4.11) would hold. This would mean that

$$\begin{aligned} \mathbf{\Sigma}_{\mathsf{F}}^{-1}\mathbf{\Sigma} &= \boldsymbol{I}_m && \text{; or, equivalently} && (4.11) \\ \mathbf{\Sigma}_{\mathsf{F}}^{-1} &= \mathbf{\Sigma}^{-1} && && (4.12) \end{aligned}$$

and we would have recovered the covariance matrix from **G**, so that **G** and $\mathbf{F}_K$ represent the same distribution. There may be other solutions but this would be ideal and, indeed, would be a good aim for any surrogate.

Unfortunately, in general, there are no consistent solutions to (4.11) and (4.12). For instance, if we take data dimension, m, as four, examining the equations at $(1, k)$ for

$k = 2, 3, 4$ in (4.12) we have

$$w_{1k} = \frac{-(\boldsymbol{R}^{-1})_{(1,k)}(1-\rho_{1k}^2)}{\rho_{1k}} \tag{4.13}$$

where w_{1k} represents the weight for the composite component for (y_1, y_k). Similarly, the equation at $(1, 1)$ in (4.11) results in

$$\begin{aligned} \sum_{k=2}^{4} \frac{w_{1k}}{1-\rho_{1k}^2} - \sum_{k=2}^{4} \frac{w_{1k}\rho_{1k}}{1-\rho_{1k}^2}\rho_{1k} &= 1 \qquad \text{or} \\ \sum_{k=2}^{4} w_{1k} &= 1. \end{aligned} \tag{4.14}$$

It is easy to construct a positive definite $\boldsymbol{\Sigma}$ (or $\boldsymbol{R}$ as the σ_i^2s are irrelevant in this case) whereby the weights in (4.13) do not satisfy (4.14). For instance, with a positive definite

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0.1 & 0.2 & 0.3 \\ 0.1 & 1 & 0.4 & 0.5 \\ 0.2 & 0.4 & 1 & 0.6 \\ 0.3 & 0.5 & 0.6 & 1 \end{pmatrix}$$

we find from (4.13) that $w_{12} = 0.7909$, $w_{13} = -0.2256$ and $w_{14} = -1.0406$ and they certainly do not sum to one as required by (4.14).

Thus, for bivariate composites, it is not, in general, possible to use the normalising constant and weights to recover the covariance matrix and so the distribution for **G**. That is because there are $m(m+1)/2$ incompatible equations for the $m(m-1)/2$ weights. Also, as we saw in Section 3.4.3 there are multiple sets of weights that do satisfy the weights equations (4.2), ie that minimise the KLD we are considering. We propose a method to resolve both these problems in Section 4.7.

4.6.2 Alternative Derivation

Rather than using the weights equation, (4.1), to calculate the weights, some elegant results arise if we minimise the Kullback-Leibler Divergence (KLD) directly, ie we minimise

$$\mathsf{E}_{\mathsf{G}}[\ln(g)-\ln(f_K)]$$

over $\boldsymbol{w}$, the vector of weights of dimension q, where $q=\binom{m}{2}$, and g and f_K are the densities corresponding to the two distributions under consideration. Now

$$\begin{aligned}
\ln(g)-\ln(f_K) &= -\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})-\frac{m}{2}\ln(2\pi)-\frac{1}{2}\ln|\boldsymbol{\Sigma}| \\
&\quad -(-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})-\frac{m}{2}\ln(2\pi)-\frac{1}{2}\ln|\boldsymbol{\Sigma}_{\mathsf{F}}|) \\
&= -((\boldsymbol{Y}-\boldsymbol{\mu})^T(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}_{\mathsf{F}}^{-1})(\boldsymbol{Y}-\boldsymbol{\mu})+\ln|\boldsymbol{\Sigma}|-\ln|\boldsymbol{\Sigma}_{\mathsf{F}}|)/2
\end{aligned}$$

so that

$$\begin{aligned}
\mathsf{E}_{\mathsf{G}}[\ln(g)-\ln(f_K)] &= -(\mathsf{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1})-\mathsf{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_{\mathsf{F}}^{-1})+\ln|\boldsymbol{\Sigma}|-\ln|\boldsymbol{\Sigma}_{\mathsf{F}}|)/2 \qquad \text{by (4.7)} \\
&= \left(-\ln|\boldsymbol{\Sigma}|-m-\ln\left|\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}\right|+\mathsf{tr}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}\right)\right)/2. \qquad (4.15)
\end{aligned}$$

The first step in the minimisation is to differentiate $\ln\left|\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}\right|$. We use (for instance Schott, 1997, Theorem 8.1 (b))

$$\frac{\partial|\boldsymbol{A}|}{\partial x}=\mathsf{tr}\left(\mathsf{adj}\left(\boldsymbol{A}\right)\frac{\partial\boldsymbol{A}}{\partial x}\right)$$

for any matrix $\boldsymbol{A}$ and variable x and where $\mathsf{adj}\left(\boldsymbol{A}\right)$ is the adjugate matrix (ie $\boldsymbol{A}^{-1}=|\boldsymbol{A}|^{-1}\mathsf{adj}(\boldsymbol{A})$), so that, if $\mathsf{C}=\{i,j\}$

$$\begin{aligned}
\frac{\partial\ln\left|\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}\right|}{\partial w_{\mathsf{C}}} &= \frac{1}{\left|\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}\right|}\mathsf{tr}\left(\mathsf{adj}\left(\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}\right)\frac{\partial\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}}{\partial w_{\mathsf{C}}}\right) \\
&= \mathsf{tr}\left(\frac{\mathsf{adj}\left(\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}\right)}{\left|\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}\right|}\frac{\partial\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}}{\partial w_{\mathsf{C}}}\right) \\
&= \mathsf{tr}\left(\boldsymbol{\Sigma}_{\mathsf{F}}\frac{\partial\boldsymbol{\Sigma}_{\mathsf{F}}^{-1}}{\partial w_{\mathsf{C}}}\right). \qquad (4.16)
\end{aligned}$$

We are just considering one of the derivatives, that with respect to w_{C}, here but the results extend to the rest of the weights.

It is clear that for $n \times n$ matrices $\boldsymbol{A} = \{a_{kl}\}$ and $\boldsymbol{B} = \{b_{kl}\}$

$$\text{tr}(\boldsymbol{AB}) = \sum_{k=1}^{n} \sum_{l=1}^{n} a_{kl} b_{lk}. \tag{4.17}$$

Now, from the contribution of component C to $\mathbf{F}^{-1}$ given at (4.10), $\frac{\partial \boldsymbol{\Sigma}_{\mathbf{F}}^{-1}}{\partial w_C}$ has zero entries everywhere except for the submatrix consisting of the intersections of the ith and jth rows and columns:

$$\frac{1}{(1-\rho_{ij}^2)} \begin{pmatrix} \sigma_i^{-2} & -\rho_{ij}/\sigma_i\sigma_j \\ -\rho_{ji}/\sigma_i\sigma_j & \sigma_j^{-2} \end{pmatrix}.$$

which is $\boldsymbol{\Sigma}_C^{-1}$. So

$$\begin{aligned} \frac{\partial \ln \left|\boldsymbol{\Sigma}_{\mathbf{F}}^{-1}\right|}{\partial w_C} &= \text{tr}\left(\boldsymbol{\Sigma}_{\mathbf{F}} \frac{\partial \boldsymbol{\Sigma}_{\mathbf{F}}^{-1}}{\partial w_C}\right) \qquad \text{by } (4.16) \\ &= \sum_{k,l=1}^{m} \boldsymbol{\Sigma}_{\mathbf{F}(k,l)} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{F}}^{-1}}{\partial w_C}\right)_{(l,k)} \qquad \text{from (4.17)} \\ &= \sum_{k,l \in C} \boldsymbol{\Sigma}_{\mathbf{F}(k,l)} \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{F}}^{-1}}{\partial w_C}\right)_{(l,k)} \qquad \text{as other elements have a zero second term} \\ &= \text{tr}\left(\boldsymbol{\Sigma}_{\mathbf{F}_C} \boldsymbol{\Sigma}_C^{-1}\right) \qquad \text{from (4.17)} \\ &= \text{tr}\left(\boldsymbol{\Sigma}_C^{-1} \boldsymbol{\Sigma}_{\mathbf{F}_C}\right). \end{aligned} \tag{4.18}$$

Also, by calculating individual components of the matrix product,

$$\begin{aligned} \text{tr}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_{\mathbf{F}}^{-1}\right) &= \sum_{i=1}^{m} \left(\sigma_i^2 \sum_{k \neq i} \frac{w_{ik}}{\sigma_i^2(1-\rho_{ik}^2)} - \sum_{k \neq i} \frac{\rho_{ik}\sigma_i\sigma_k\rho_{ik}w_{ik}}{\sigma_i\sigma_k(1-\rho_{ik}^2)}\right) \\ &= \sum_{i=1}^{m} \sum_{k \neq i} w_{ik} \\ &= 2 \sum_{C \in \mathcal{C}} w_C \end{aligned} \tag{4.19}$$

which is extremely elegant. The derivative of tr $\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_{\mathbf{F}}^{-1}\right)$ with respect to any weight will thus be 2.

Combining (4.18) and (4.19), we differentiate (4.15) to get

$$\left(-\mathrm{tr}\left(\boldsymbol{\Sigma}_{C}^{-1}\boldsymbol{\Sigma}_{\mathbf{F}_C}\right)+2\right)/2 \tag{4.20}$$

and by setting (4.20) equal to zero for the minimum value, it matches the result at (4.2).

4.7 Combined Bivariate and Univariate Margins

We have seen how the use of our KLD criterion produces easily calculable weights in the univariate but not the bivariate case. We now examine what happens if we combine the two, ie our composite components consist of all univariate and all bivariate marginals of the distribution that generated the data, **G**.

Clearly, from its definition at (4.5), $\boldsymbol{\Sigma}_{\mathbf{F}}^{-1}$ for the combined case is just the sum of its values for the univariate and bivariate cases so that, from (4.10), it will have off diagonal (i,j) element

$$\frac{-w_{ij}\rho_{ij}}{\sigma_i\sigma_j(1-\rho_{ij}^2)}$$

and, from (4.9) and (4.10), ith diagonal element

$$\frac{1}{\sigma_i^2}\left(w_{ii}+\sum_{k\neq i}\frac{w_{ik}}{1-\rho_{ik}^2}\right).$$

We follow the bivariate case and examine equations (4.11) and (4.12) to see if there are values of the weights for which $\boldsymbol{\Sigma}_{\mathbf{F}}=\boldsymbol{\Sigma}$. The advantage in the combined case is that we have the same number of weights and equations, namely $m(m+1)/2$. The equations resulting from the off diagonal comparisons in (4.12) are all linear in individual weights giving, for position (i,j)

$$-\frac{\omega_{ij}\rho_{ij}}{\sigma_i\sigma_j(1-\rho_{ij}^2)}=(\boldsymbol{\Sigma}^{-1})_{(i,j)}$$

or

$$\begin{aligned} w_{ij} &= -\frac{(\boldsymbol{R}^{-1})_{(i,j)}(1-\rho_{ij}^2)}{\rho_{ij}} \\ &= -\frac{(\boldsymbol{R}^{-1})_{(i,j)}(1-(\boldsymbol{R}_{(i,j)})^2)}{\boldsymbol{R}_{(i,j)}}. \end{aligned} \tag{4.21}$$

Taking the ith diagonal comparison in (4.11) we see that

$$\begin{aligned} \left(\frac{w_{ii}}{\sigma_i^2}+\sum_{k\neq i}\frac{w_{ik}}{\sigma_i^2(1-\rho_{ik}^2)}\right)\sigma_i^2-\sum_{k\neq i}\frac{w_{ik}\rho_{ik}}{\sigma_i\sigma_k(1-\rho_{ik}^2)}\rho_{ik}\sigma_i\sigma_k &= 1 \qquad (4.22)\\ w_{ii}+\sum_{k\neq i}\frac{w_{ik}(1-\rho_{ik}^2)}{1-\rho_{ik}^2} &= 1 \\ \sum_{k=1}^{n} w_{ik} &= 1 \qquad (4.23) \end{aligned}$$

which is elegant. Note that (4.22) is equivalent to combining the m equations from row i of (4.12), but the derivation we have used is simpler algebraically.

We have thus derived a set of weights uniquely from (4.11) and (4.12) such that $\boldsymbol{\Sigma}_{\mathsf{F}} = \boldsymbol{\Sigma}$ and we have recovered the original distribution **G**. These weights do not involve the variances from $\boldsymbol{\Sigma}$ - they are derived solely from the corresponding correlation matrix, $\boldsymbol{R}$.

These weights uniquely solve (4.11) and (4.12) and so, following the argument in Note 7 of Section 3.4.2, the second derivative of the KLD with respect to the weights must be nonsingular.

Interpretation of the optimal value for the off diagonal weights, (4.21), relies on understanding $\boldsymbol{R}^{-1}$. Elements of the inverse of a covariance matrix, sometimes known as the *concentration matrix*, can be understood in the context of partial correlation. For instance, for $\boldsymbol{Y} \equiv (Y_1,\ldots,Y_m)$ distributed with zero mean and covariance matrix $\boldsymbol{\Sigma}$, Cox and Wermuth (1996, Section 3.4), show that the correlation between Y_i and Y_j, conditional upon the other elements of $\boldsymbol{Y}$, (ie the *partial correlation*) is

$$-\frac{(\boldsymbol{\Sigma}^{-1})_{(i,j)}}{((\boldsymbol{\Sigma}^{-1})_{(i,i)}(\boldsymbol{\Sigma}^{-1})_{(j,j)})^{0.5}}. \tag{4.24}$$

Now, it is simple to show that (4.24) equals

$$\frac{(\boldsymbol{R}^{-1})_{(i,j)}}{((\boldsymbol{R}^{-1})_{(i,i)}(\boldsymbol{R}^{-1})_{(j,j)})^{0.5}}.$$

and so our weight, w_{ij}, is zero whenever the partial correlation is zero.

If one is going to use a bivariate normal composite surrogate, then we have seen that by adding in univariate elements and then applying optimal weights one can recover the equivalent multivariate normal distribution. In practice, by using (4.21) and (4.23) to calculate the optimal weights requires knowledge of and then inversion of $\boldsymbol{\Sigma}$. It is more computationally efficient than solving the weights equations, (4.2), through some iterative process such as nonlinear minimisation, as we just have to invert $\boldsymbol{\Sigma}$ once rather than additionally having to invert $\boldsymbol{J}$ for every iteration. It also replaces any parameter estimation through solving the estimating equations, which, for numerical approaches, involves repeated inversion of the matrix of derivatives of the estimating equations.

Having derived our weights equations, (4.2), for the multivariate normal distribution, we have applied them to some generic cases. We have seen that we need to consider composite surrogates consisting of univariate and bivariate components in order to arrive at unique optimal weights, which we have derived analytically. We first apply these insights to Simulation II and then to a specific case, where we find that there is some value in using the weighted approach as the optimal weights bear interpretation.

4.7.1 Simulation II

Having seen the effect of combining bivariate with univariate components in Section 4.7, it is of interest to see whether that insight has any effect on Simulation II from Section 3.4.5. To that end, it was updated to include the following two further surrogates

Bivariate and univariate unweighted A standard combined univariate and bivariate composite surrogate per (3.1) with all the weights set to 1.

Bivariate and univariate weighted A standard univariate and bivariate composite surrogate per (3.1) iterated with weights calculations as in Section 3.4.4. As the clustered datasets are of variable length, weights were calculated separately for each cluster.

The resulting power curves are shown in Figure 4.1 where they are compared to the bivariate surrogate curve. Examining the plot, again reveals no significant differences between the surrogates as at different points on the steeper parts of the curves, each of them has the greatest power. At the minimum points of the curves, where one would like values as close as possible to 0.05, the bivariate surrogate has value 0.08 while the unweighted and weighted combined surrogates have values 0.057 and 0.063 respectively.

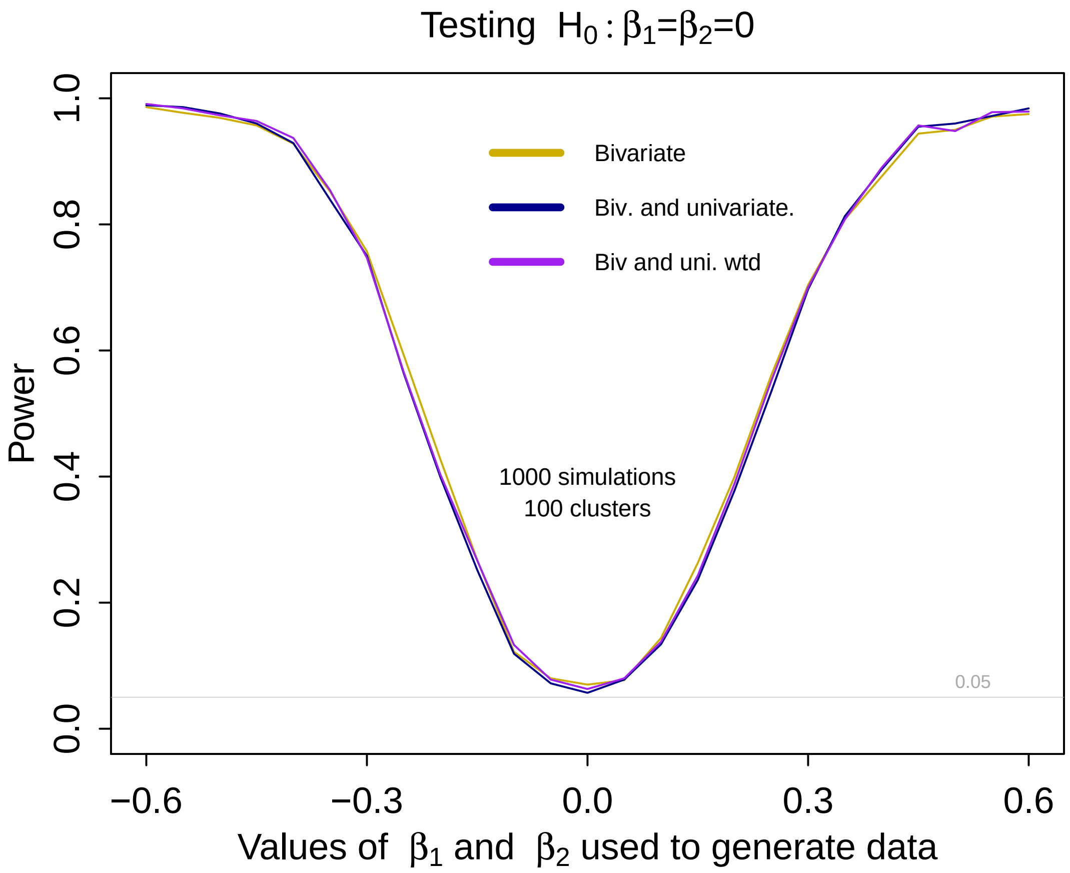

Figure 4.1: Simulation II. Power plots at level 0.05 for a range of surrogates.

Once again, as described in Section 3.4.5, the raw data which were used to generate the power curves were tested against the null hypothesis that the means of the p-values of the three surrogates were identical. The results showed no consistent pattern but it is worth noting that at data generated with $\beta_1 = \beta_2 = 0$, the mean for the unweighted combined univariate and bivariate surrogate (but not for the weighted) was significantly lower than that for the bivariate surrogate (p-value 0.0017).

Despite the minor advantage that the unweighted combined surrogate has over the other surrogates at the minimum of the power curves, there is no overall significant difference between any of the four surrogates studied. This suggests that, in practice, there may be no advantage in applying weights to composite surrogates and is discussed, together with a possible explanation, in Chapter 5.

4.8 Autoregressive Models

We now examine combined weighted univariate and bivariate normal composites, $\mathbf{F}_K$, as surrogates for the distribution of data (of length m) generated from stationary autoregressive models, $\mathbf{G}$. We define Y_t, $t \in \mathbb{N}$ as an autoregressive process of order $l < m$, AR(l), with mean zero if

$$Y_t = \sum_{i=1}^{l} \phi_i Y_{t-i} + \epsilon_t$$

where the $\{\epsilon_t : t \in \mathbb{N}\}$ form a white noise sequence (ie a sequence of iid random variables with zero mean), uncorrelated with the Y_ts, and with variance σ^2. Stationarity requires that all the roots of the characteristic equation

$$1 - \sum_{i=1}^{l} \phi_i x^i = 0$$

lie outside the unit circle.

We now attempt to recover the covariance matrix, Σ, that was involved in generating the data by using the combined bivariate and univariate composite surrogate results from Section 4.7. Standard results (see, eg, Brockwell and Davis, 1991, Section 5.1.1) show that in the AR(1) case,

$$\boldsymbol{\Sigma} = \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{m-1} \\ \phi & 1 & \phi & \dots & \phi^{m-2} \\ \phi^2 & \phi & 1 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \phi \\ \phi^{m-1} & \phi^{m-2} & \dots & \phi & 1 \end{pmatrix}$$

with the correlation matrix, $\boldsymbol{R}$, found by omitting the $\sigma^2/(1-\phi^2)$ factor. It is then easy to verify that

$$\boldsymbol{R}^{-1} = \frac{1}{1-\phi^2}\begin{pmatrix} 1 & -\phi & 0 & \dots & 0 \\ -\phi & 1+\phi^2 & -\phi & \dots & 0 \\ 0 & -\phi & 1+\phi^2 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & -\phi \\ 0 & 0 & \dots & -\phi & 1 \end{pmatrix}$$

Working with a combined univariate and bivariate normal weighted composite surrogate as in Section 4.7, (4.21) gives us, for $|i-j| = 1$

$$\begin{aligned} w_{ij} &= -\frac{-\phi(1-\phi^2)}{(1-\phi^2)\phi} \\ &= 1 \end{aligned}$$

and for $|i-j| > 1$, zero. Then, from (4.23), w_{ii} will be -1 for all i except $i = 1, m$ when it will be zero. The simplest way to lay that out is in an $m \times m$ weights matrix, $\boldsymbol{W}$ where $\boldsymbol{W}_{(i,j)} = w_{ij}$:

$$\boldsymbol{W} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & -1 & 1 & \dots & 0 \\ 0 & 1 & -1 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}. \tag{4.25}$$

Note that

1. $\boldsymbol{W}$ is a convenient way of representing the weights, particularly in light of (4.23). Note that the weights for all the bivariate pairs appear twice in the matrix but only once in the composite likelihood.

2. The optimal weights are not functions of the parameters. We have seen from Section 4.7 that σ^2 is not involved in the weights but, here, the correlation coefficient is also excluded.

3. (4.25) makes some sort of intuitive sense in that

   - we learn about ρ and σ^2 from the lag one pairs, (y_i, y_{i+1}). Adding more distant pairs does not give us any more critical information which is consistent with the zeroes in $\boldsymbol{W}$; and
   - all but the outer two random variables, y_1 and y_m, appear in two pairs in our composite surrogate. The univariate information about them has thus been duplicated and the -1s in $\boldsymbol{W}$ remedy that.

4. Davis and Yau (2011) note that the exact loglikelihood for an AR(1) sequence is

$$\ell(\phi, \sigma^2; y_m) = \sum_{i=1}^{m-1} \ln f_{i+1|i}(y_{i+1}|y_i) + \ln f_1(y_1) \tag{4.26}$$

   where the subscripts to the densities f denote appropriate marginality from the density corresponding to **G**. If we then apply the weights we have just calculated for normal composite surrogates to a combined univariate and bivariate loglikelihood based on those marginal densities, we have a composite surrogate loglikelihood

$$\begin{aligned} \ell_c(\phi, \sigma^2; \boldsymbol{y}) &= \sum_{i=1}^{m-1} \ln f_{ij}(y_i, y_{i+1}) - \sum_{i=2}^{m-1} \ln f_i(y_i) \\ &= \sum_{i=1}^{m-1} \ln f_{i+1|i}(y_{i+1}|y_i) + \ln f_1(y_1) \end{aligned} \tag{4.27}$$

   where $\boldsymbol{y} = (y_1, \ldots, y_m)$, and we have recovered the exact loglikelihood from (4.26). Davis and Yau (2011) point out the similarity between the exact and bivariate composite loglikelihoods but take no account of the univariate components. Joe and Lee (2009) make a similar comment but on the assumption that the variances (and means if appropriate) are known. We do not make that assumption here.

5. If we define $B(Y_i, Y_{i+1})$ to be the bivariate distribution for Y_i and Y_{i+1} (and $U(Y_i)$ similarly for univariate), then our composite surrogate is (omitting the parameters, for simplicity, and any constant of proportionality)

$$B(Y_1, Y_2)B(Y_2, Y_3|Y_2) \dots B(Y_{m-1}, Y_m|Y_{m-1})$$

   or, the more usual Markov chain result

$$U(Y_1)B(Y_2|Y_1)B(Y_3|Y_2) \dots B(Y_m|Y_{m-1}) \tag{4.28}$$

   equivalent to that at (4.27).

6. (4.25) is consistent with Lindsay et al. (2011)'s general result given at (3.5) for the set of parameters such that the Y_t are mutually independent, $\theta \in \mathbf{\Theta}_{ind}$. Here, we are working with a particular example, but the result holds not just in $\mathbf{\Theta}_{ind}$.

7. Returning to the partial correlation interpretation of the weights described in Section 4.7, we see from $\boldsymbol{W}$ that the partial correlation between Y_i and Y_{i+1} is the same for all i and that the partial correlation between elements that are not adjacent is zero. This is consistent with the AR(1) model that we are examining.

Unfortunately, the analogous results for higher order AR models are more difficult to interpret. For the AR(2) case, the inverse of the covariance matrix that was involved in generating the data, $\mathbf{\Sigma}$, is given in Barry et al. (1997) in the form of a Cholesky decomposition whereby $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\boldsymbol{Q}^T\boldsymbol{Q}$. Note however, that the $(2, 2)$ entry for $\boldsymbol{Q}$ should be $(1 - \phi_2^2)^{\frac{1}{2}}$ not $(1 - \phi_2)^{\frac{1}{2}}$ as given in Barry et al. (1997). As one would expect, the matrix has zeroes everywhere apart from a strip up to five entries wide down the main diagonal. We can solve for the weights in exactly the same way as for AR(1) and the resulting weights matrix is symmetric along both main diagonals. The upper left hand corner of $(1+\phi_2)\boldsymbol{W}$ is

$$\begin{pmatrix} -\frac{(1-\phi_2)\phi_2}{\phi_1^2-\phi_2^2+\phi_2} & 1 & \frac{\phi_2(1+\phi_1^2-\phi_2^2)}{\phi_1^2-\phi_2^2+\phi_2} & 0 & \cdots \\ 1 & -\frac{(1-\phi_2)(\phi_1^2-\phi_2^2+2\phi_2)}{\phi_1^2-\phi_2^2+\phi_2} & 1-\phi_2 & \frac{\phi_2(1+\phi_1^2-\phi_2^2)}{\phi_1^2-\phi_2^2+\phi_2} & \cdots \\ \frac{\phi_2(1+\phi_1^2-\phi_2^2)}{\phi_1^2-\phi_2^2+\phi_2} & 1-\phi_2 & -\frac{(1-\phi_2)(\phi_1^2-\phi_2^2+3\phi_2)}{\phi_1^2-\phi_2^2+\phi_2} & 1-\phi_2 & \cdots \\ 0 & \frac{\phi_2(1+\phi_1^2-\phi_2^2)}{\phi_1^2-\phi_2^2+\phi_2} & 1-\phi_2 & -\frac{(1-\phi_2)(\phi_1^2-\phi_2^2+3\phi_2)}{\phi_1^2-\phi_2^2+\phi_2} & \cdots \\ 0 & 0 & \frac{\phi_2(1+\phi_1^2-\phi_2^2)}{\phi_1^2-\phi_2^2+\phi_2} & 1-\phi_2 & \cdots \\ 0 & 0 & 0 & \frac{\phi_2(1+\phi_1^2-\phi_2^2)}{\phi_1^2-\phi_2^2+\phi_2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

which is, however, hard to interpret directly.

The use of univariate and bivariate normal composites as surrogates for distributions more generally will be the subject of future work.

4.9 Summary

In this chapter we have derived an elegant form of the optimal scalar weighting scheme for the multivariate normal case. One of the consequences of that form is that a composite surrogate consisting of multivariate normal components has a related density that is also multivariate normal, and that if the data under consideration is itself generated from a multivariate normal distribution then the composite surrogate is just the distribution of a linear transformation of the data. We have then applied the multivariate normal form to univariate, bivariate and combined univariate and bivariate composite surrogates. Only in the latter case can a weighting scheme recover the data generating distribution. We apply that knowledge to the simulation from Chapter 3 but, again, there is no significant improvement in power. Finally, we create a combined multivariate surrogate for autoregressive processes. In the case of data from an AR(1) distribution, it recovers a standard presentation of that distribution.

Chapter 5

Discussion

We have explored a number of options for applying weights to the components of composite surrogates. We have done that by understanding the theory behind surrogate distributions (Chapter 1), applying that theory to composite surrogates (Chapter 2) and then focusing on the weights for components, in theory (Chapter 3) and as applied to specific examples (Chapters 3 and 4). Our most significant contributions are, in the order in which they were introduced

1. Assumption 17 in Section 2.3, which allows us to work with unbiased composite estimating functions when the data generating mechanism (DGM) is unknown.

2. The introduction of the constant of proportionality (or normalising constant) into the study of composite likelihoods in Section 2.5, resulting in a genuine density function.

3. Applying the Bartlett correction to adjusted composite surrogates, in Section 2.7.2, for small samples results in rejection levels that are closer to the Type I error rate than that for the uncorrected surrogates, when the parameters that are the coefficients of the covariate data are set to 0. This was carried out in a simulation where the correction is relatively easy to calculate. In a real example, that calculation is more complex.

4. A more general version of an optimal weighting scheme for composite estimating

functions described in Lindsay et al. (2011), proved in Section 3.3.1.

5. A version of that scheme, in Section 3.3.2, that is optimal in its restricted class, and is computationally cheaper.

6. In Section 3.4 a completely new weighting scheme for composite likelihoods based on minimising the Kullback-Leibler discrepancy (KLD) over the weights between the weighted distribution, taking into account the constant of proportionality (CoP), and the DGM, resulting in a set of equations to be solved to generate, not necessarily unique, weights.

7. Application of that new scheme to the multivariate normal distribution in Section 4.3 to give elegant equations for deriving the optimal weights.

8. Demonstration that multivariate normal composite surrogates have a distribution that is that of a transformation of the random variables that generated the data, in Section 4.4.

9. Exploration of multivariate normal composites consisting of univariate and / or bivariate components. In Section 4.7 it is shown that only when all bivariate and univariate components are combined can the DGM's covariance matrix be recovered.

10. A simulation based on probit regression with an autoregressive random effect to which weights from the weights equations and the combined composite surrogate lesson are applied in Sections 3.4.5 and 4.7.1. They show no significant improvement on the standard unweighted bivariate composite surrogate.

11. Application of that combined bivariate and univariate principle to multivariate composite surrogates for autoregressive data and, in Section 4.8, it is shown that in the AR(1) case, the covariance matrix involved in generating the data is recovered.

The most significant of these contributions is the new scalar weighting scheme in Point 6 and we now analyse that and its application, with particular reference to the simulation in Point 10.

There are two key features to the derivation of the new weighting scheme The first is the minimisation of the KLD over the weights between the DGM and the weighted composite surrogate loglikelihood in order to derive the weights equations. This is similar to our approach to parameter estimation as described in Section 1.2, where we minimise the KLD over the parameters between the DGM and a surrogate loglikelihood. A strength of this new scheme is thus the commonality of approach with parameter estimation. However, there is a difference, which is that in the former we take into account the CoP, our second key feature, while in the latter we do not. There are good reasons in each case. For weights, if we use the KLD without the CoP, we end up with all the weight attached to one component, as described in Section 3.4.1. For parameters, if we take the CoP as a feature of the surrogate, then the estimates can be biased (see Section 2.5).

Having derived the weights equations, we need to solve them to calculate optimal weights. We saw in Section 3.4.3 an example where the solutions are not unique. However, that is not always the case as in Simulation II in Section 3.4.5. Usually, the equations will not be linear in the weights, and their number and complexity will require a numerical minimisation routine.

Associated with each set of weights are a set of parameter estimators which will need to be calculated in the usual way for the weighted composite surrogate. Repeating the process by calculating further weights (based on the weighted estimates) and estimators is unnecessary as described in Section 3.4.4. It is the estimators in which we are interested as, generally, the weights will not have a plausible interpretation.

We can now assess the effectiveness of weighted composite surrogates by whichever criterion we deem appropriate (see Section 1.7 for a review). For Simulation II, we have chosen power, and in Sections 3.4.5 and 4.7.1, we see that applying weights to bivariate and combined univariate and bivariate composite surrogates has no significant effect. It might be thought that this is because there is no further power to be had, ie the bivariate composite is as powerful as the DGM. However, Simulation I in Section 2.7 is a case where the DGM is clearly more powerful and, again, applying our optimal weights to the bivariate composite surrogate results in no significant improvement.

An informal explanation for the lack of improvement in power might be that the information contained within the bivariate surrogate is repeated many times as we consider all bivariate pairs. In that case, any weighting scheme will just be shuffling around the multiple instances of information and is unlikely to show any significant improvement. Lindsay et al. (2011) suggest that there may be such redundant information, as described in Sections 3.2.3 and 3.2.4. To test this possibility, one would need to reduce the number of bivariate components to the point that there is little redundant information and then assess the effect of optimal weights. However, this would be a time consuming process and would not improve the results from the all pair bivariate surrogate. In simple examples such as AR(1) in Section 4.8 or the multivariate normal considered in Section 3.4.3, it is relatively simple to interpret the weights as emphasising useful or downgrading redundant information. However, the complexity of the structure of Simulation II with a link function and partially correlated covariates, means that it is very difficult to interpret each weight directly and anticipate the relative size of the weights with respect to useful marginal information.

It is worth considering, then, the circumstances in which further exploration of weighting schemes might be worthwhile. The first is in the case of clustered data of varying lengths as discussed in Section 3.2.5. Various formulae, dependent upon the cluster length, for the common cluster weight have been suggested. Contrast that with Simulation II, where we have calculated different weights for each component of each cluster, each cluster treated separately from one other. One possibility for further research might be to combine the two ideas so that the individual weights for each cluster component would undergo some form of normalisation across the whole of the data.

A second possibility might be to explore further the optimally efficient weighting schemes described in Points 4 and 5. While the former is likely to be too computationally expensive for interesting situations, ie large numbers of parameters or composite components, the latter is only expensive for large numbers of parameters and might be worth understanding better for composite surrogates with large numbers of components, such as can arise from long time series.

Further work, not related to weighting schemes, resulting from ideas in this thesis could be undertaken in

- Understanding whether the form of multivariate normal composites described in Point 8 has more general applicability to the understanding of any composite surrogate.
- Analysing the effect of using combined bivariate and univariate composite surrogates as opposed to bivariate alone, as set out in Point 9.

Overall, we have shown that there may be no advantage to scalar weighting schemes for the components of composite surrogates. Optimal weighting schemes may be either computationally very expensive or show no significant improvement over their unweighted, relatively cheap to calculate, equivalents.

Appendix A

Equivalence of the Three Usual Statistics

We show that, asymptotically, the Wald, score and likelihood ratio statistics for surrogates, as defined in Section 1.5, are equivalent. Our null hypothesis is that $\theta_{\mathsf{G}} = \theta_*$. The proof begins with a modified version of that from Kent (1982). By the Lagrange form of Taylor's theorem, for any n:

$$\begin{aligned}
\ell(\theta) &= \ell(\hat{\theta}_n) + \psi_n(\hat{\theta}_n)^T(\theta - \hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^T \psi_n'(\breve{\theta}_n)(\theta - \hat{\theta}_n) \\
&\quad \text{for some } \breve{\theta}_n \text{ `between' } \theta \text{ and } \hat{\theta}_n \\
&= \ell(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^T \psi_n'(\breve{\theta}_n)(\theta - \hat{\theta}_n) \quad \text{by the definition of } \hat{\theta}_n \qquad \text{(A.1)}
\end{aligned}$$

where 'between' is used in the sense described after (1.13). Similarly:

$$\begin{aligned}
\ell(\theta) &= \ell(\theta_{\mathsf{G}}) + \psi_n(\theta_{\mathsf{G}})^T(\theta - \theta_{\mathsf{G}}) + \frac{1}{2}(\theta - \theta_{\mathsf{G}})^T \psi_n'(\tilde{\theta}_n)(\theta - \theta_{\mathsf{G}}) \\
&\quad \text{for some } \tilde{\theta}_n \text{ `between' } \theta \text{ and } \theta_{\mathsf{G}} \\
&= \ell(\theta_{\mathsf{G}}) + \psi_n(\hat{\theta}_n)^T(\theta - \theta_{\mathsf{G}})(1 + o_p(1)) + \frac{1}{2}(\theta - \theta_{\mathsf{G}})^T \psi_n'(\tilde{\theta}_n)(\theta - \theta_{\mathsf{G}}) \\
&\quad \text{by Assumption 12} \\
&= \ell(\theta_{\mathsf{G}}) + \frac{1}{2}(\theta - \theta_{\mathsf{G}})^T \psi_n'(\tilde{\theta}_n)(\theta - \theta_{\mathsf{G}}) \quad \text{by the definition of } \hat{\theta}_n. \qquad \text{(A.2)}
\end{aligned}$$

Evaluating (A.1) at $\theta = \theta_{\mathsf{G}}$ we find for the likelihood ratio statistic as defined in (1.24):

$$\begin{aligned}
W_l &= -(\theta_{\mathsf{G}} - \hat{\theta}_n)^T \psi_n'(\breve{\theta}_n)(\theta_{\mathsf{G}} - \hat{\theta}_n) \\
&= -(\theta_{\mathsf{G}} - \hat{\theta}_n)^T \psi_n'(\theta_{\mathsf{G}})(\theta_{\mathsf{G}} - \hat{\theta}_n)(1 + o_p(1)) \\
&\qquad \text{asymptotically, as } \hat{\theta}_n \text{ is consistent for } \theta_{\mathsf{G}}\text{, } \breve{\theta}_n \text{ is 'between' the two and using} \\
&\qquad \text{Assumption 12} \\
&= -(\theta_{\mathsf{G}} - \hat{\theta}_n)^T \boldsymbol{A}_n^{-1} \bar{\psi}_n'(\theta_{\mathsf{G}})(\theta_{\mathsf{G}} - \hat{\theta}_n)(1 + o_p(1)) \quad \text{by the definition of } \bar{\psi}_n'(\theta_{\mathsf{G}}) \\
&= -(\theta_{\mathsf{G}} - \hat{\theta}_n)^T \boldsymbol{A}_n^{-1} \psi_\infty'(\theta_{\mathsf{G}})(\theta_{\mathsf{G}} - \hat{\theta}_n)(1 + o_p(1)) \quad \text{by Assumption 7} \\
&= (\theta_{\mathsf{G}} - \hat{\theta}_n)^T \boldsymbol{A}_n^{-1} \boldsymbol{I}_{\mathsf{G}}(\theta_{\mathsf{G}})(\theta_{\mathsf{G}} - \hat{\theta}_n)(1 + o_p(1)) \quad \text{by the definition of } \boldsymbol{I}_{\mathsf{G}}(\theta_{\mathsf{G}}) \\
&= (\theta_{\mathsf{G}} - \hat{\theta}_n)^T \boldsymbol{A}_n^{-\frac{1}{2}} \boldsymbol{I}_{\mathsf{G}}(\theta_{\mathsf{G}}) \boldsymbol{A}_n^{-\frac{1}{2}} (\theta_{\mathsf{G}} - \hat{\theta}_n)(1 + o_p(1)) \quad \text{by Assumption 7} \qquad \text{(A.3)}
\end{aligned}$$

and so under the null hypothesis, $H_0 : \theta_{\mathsf{G}} = \theta_*$, the likelihood ratio and Wald statistics are asymptotically equivalent :

$$W_l = W_w(1 + o_p(1)) \qquad \text{by (1.25).}$$

Similarly, evaluating (A.2) at $\theta = \hat{\theta}_n$ we have from (1.24):

$$\begin{aligned}
W_l &= (\hat{\theta}_n - \theta_{\mathsf{G}})^T \psi_n'(\tilde{\theta}_n)(\hat{\theta}_n - \theta_{\mathsf{G}}) \\
&= -(\hat{\theta}_n - \theta_{\mathsf{G}})^T \boldsymbol{A}_n^{-\frac{1}{2}} \boldsymbol{I}_{\mathsf{G}}(\theta_{\mathsf{G}}) \boldsymbol{A}_n^{-\frac{1}{2}} (\hat{\theta}_n - \theta_{\mathsf{G}})(1 + o_p(1)) \\
&\qquad \text{asymptotically, by a similar argument to (A.3)} \\
&= -(\boldsymbol{I}_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}})\bar{\psi}_n(\theta_{\mathsf{G}}))^T \boldsymbol{A}_n^{-\frac{1}{2}} \boldsymbol{I}_{\mathsf{G}}(\theta_{\mathsf{G}}) \boldsymbol{A}_n^{-\frac{1}{2}} (\boldsymbol{I}_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}})\bar{\psi}_n(\theta_{\mathsf{G}}))(1 + o_p(1)) \quad \text{by (1.14)} \\
&= -(\boldsymbol{I}_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}})\boldsymbol{A}_n\psi_n(\theta_{\mathsf{G}}))^T (\boldsymbol{A}_n^{-\frac{1}{2}})^T \boldsymbol{I}_{\mathsf{G}}(\theta_{\mathsf{G}}) \boldsymbol{A}_n^{-\frac{1}{2}} \boldsymbol{I}_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}})\boldsymbol{A}_n\psi_n(\theta_{\mathsf{G}})(1 + o_p(1)) \\
&\qquad \text{by the definition of } \bar{\psi}_n(\theta_{\mathsf{G}}) \\
&= -\psi_n(\theta_{\mathsf{G}})^T \boldsymbol{A}_n^{\frac{1}{2}} \boldsymbol{I}_{\mathsf{G}}^{-1}(\theta_{\mathsf{G}}) \boldsymbol{A}_n^{\frac{1}{2}} \psi_n(\theta_{\mathsf{G}})(1 + o_p(1))
\end{aligned}$$

by Assumption 7 and the symmetric condition from Assumption 5.

Thus under the null hypothesis, $H_0 : \theta_{\mathsf{G}} = \theta_*$, the likelihood ratio and score statistics are asymptotically equivalent:

$$W_l = W_s(1 + o_p(1)) \qquad \text{by (1.26)}$$

the required change in sign being noted in, for instance, Cox (2006, page 105).

Appendix B

Cluster Estimating Functions Are Uncorrelated

We consider clustered data where there is an order (eg time, space) to those elements and where the marginal components of our composite likelihood are marginal for **G**, as described in Section 2.6. We show that if the elements are independent (eg Y_i), conditional upon earlier elements (eg $\mathcal{D}_i$), then the univariate composite surrogate estimating functions components are uncorrelated. We define those elements of θ that parameterise the marginal structure as α ($\alpha_{\mathbf{G}}$ having the obvious meaning), the rest of the parameters providing intra cluster dependence. Let $i < i'$ be cluster indices. Then:

$$
\begin{aligned}
&\mathsf{Cov}[\psi_i(\boldsymbol{\theta}_{\mathbf{G}}; Y), \psi_{i'}(\theta_{\mathbf{G}}; Y)] \\
&\quad = \mathsf{Cov}[\psi_i(\alpha_{\mathbf{G}}; Y), \psi_{i'}(\alpha_{\mathbf{G}}; Y)] \\
&\quad = \mathsf{E}[\psi_i(\alpha_{\mathbf{G}}; Y)^T \psi_{i'}(\alpha_{\mathbf{G}}; Y)] \\
&\quad = \int_y \left.\frac{\partial \ln f_i}{\partial \alpha}\right|_{\alpha_{\mathbf{G}}} \left.\frac{\partial \ln f_{i'}}{\partial \alpha}\right|_{\alpha_{\mathbf{G}}} f_c(y|\mathcal{D}) \, \mathrm{d}y \\
&\quad = \int_y \frac{1}{f_i(y|\mathcal{D}_i; \alpha_{\mathbf{G}})} \left.\frac{\partial f_i}{\partial \alpha}\right|_{\alpha_{\mathbf{G}}} \frac{1}{f_{i'}(y|\mathcal{D}_{i'}; \alpha_{\mathbf{G}})} \left.\frac{\partial f_{i'}}{\partial \alpha}\right|_{\alpha_{\mathbf{G}}} \prod_{j=1}^{n} f_j(y|\mathcal{D}_j; \alpha) \, \mathrm{d}y \\
&\quad = \int_{y_1} f_1(y|\mathcal{D}_1; \alpha) \ldots \int_{y_i} \frac{f_i(y|\mathcal{D}_i; \alpha)}{f_i(y|\mathcal{D}_i; \alpha_{\mathbf{G}})} \left.\frac{\partial f_i}{\partial \alpha}\right|_{\alpha_{\mathbf{G}}} \\
&\qquad \ldots \int_{y_{i'}} \frac{f_{i'}(y|\mathcal{D}_{i'}; \alpha)}{f_{i'}(y|\mathcal{D}_{i'}; \alpha_{\mathbf{G}})} \left.\frac{\partial f_{i'}}{\partial \alpha}\right|_{\alpha_{\mathbf{G}}} \ldots \int_{y_n} f_n(y|\mathcal{D}_n; \alpha) \, \mathrm{d}y
\end{aligned}
$$

where the $\{f_j\}$ are conditional densities which integrate to 1. Working from right to left each term integrates to 1 until the term with index i'. For that:

$$\begin{aligned}
\int_{y_{i'}} \frac{f_{i'}(y|\mathcal{D}_{i'};\alpha)}{f_{i'}(y|\mathcal{D}_{i'};\alpha_{\mathsf{G}})} \frac{\partial f_{i'}}{\partial \alpha}\bigg|_{\alpha_{\mathsf{G}}} \mathrm{d}y_i &= \frac{\partial}{\partial \alpha} \int_{y_{i'}} \frac{f_{i'}(y|\mathcal{D}_{i'};\alpha)}{f_{i'}(y|\mathcal{D}_{i'};\alpha_{\mathsf{G}})} f_{i'}(y|\mathcal{D}_{i'};\alpha_{\mathsf{G}})\,\mathrm{d}y_i \\
&= \frac{\partial}{\partial \alpha} \int_{y_{i'}} f_{i'}(y|\mathcal{D}_{i'};\alpha)\,\mathrm{d}y_i \\
&= \frac{\partial 1}{\partial \alpha} \\
&= 0
\end{aligned}$$

leading to the overall covariance being 0.

Appendix C

Integrating Out the Random Effect in Simulation I Bivariate Surrogates

As part of our composite surrogate simulation in Section 2.7 we create a further surrogate for the bivariate case. The advantage of this further surrogate is that we can integrate out the random effect, as we demonstrate here. For cluster i, and any two elements in that cluster, y_{j_1}, y_{j_2}, the surrogate for our bivariate component has Bernoulli density, π_b, such that

$$\pi_b(y_{ij_1}, y_{ij_2}; \boldsymbol{\beta}, a, u_{ij_{12}}) = p_{ij_1}^{y_{ij_1}}(1-p_{ij_1})^{1-y_{ij_1}} p_{ij_2}^{y_{ij_2}}(1-p_{ij_2})^{1-y_{ij_2}}$$

where $\boldsymbol{\beta}$ is the vector of parameters in which we are interested, $u_{ij_{12}} \sim U[0,1]$, the random effect, and

$$p_{ij_l} = \frac{1}{(1+\exp(-a_n(u_{ij_{12}} - 1 + \mu_{ij_l}))} \quad l = 1, 2$$

where a_n is a nuisance parameter in place of a. Assuming that all clusters have more than one member, the overall composite loglikelihood for all pairs is then

$$\sum_{i=1}^{n} \sum_{j_1 \neq j_2} \ln(\pi_b(y_{ij_1}, y_{ij_2}; \boldsymbol{\beta}, a, u_{ij_{12}})).$$

We wish to integrate out the random effects as described in, for instance, Fahrmeir and Tutz (2001, Section 7.4.1), to give a marginal loglikelihood

$$\sum_{i=1}^{n} \sum_{j_1 \neq j_2} \ln \left(\int_0^1 \pi_b(y_{ij_1}, y_{ij_2} | u_{ij_{12}}; \boldsymbol{\beta}, a_n) p(u_{ij_{12}}) \mathrm{d}u_{ij_{12}} \right)$$

as, in the composite, each random effect is unique to each bivariate pair. For a pair (j_1, j_2) in cluster i, integration results in one of two loglikelihoods, omitting the gory details

1. $\mu_{ij_1} \neq \mu_{ij_2}$:

$$\begin{aligned} \ln \Bigg(& y_{ij_1} y_{ij_2} + \frac{1}{a_n(b_2 - b_1)} \\ \cdot & \left(\ln \left(\frac{1 + \exp(a_n) b_1}{1 + b_1} \right) (b_2(1 - 2y_{ij_1})(1 - y_{ij_2}) - b_1(1 - 2y_{ij_1}) y_{ij_2}) \right. \\ & \left. + \ln \left(\frac{1 + \exp(a_n) b_2}{1 + b_2} \right) (-b_1(1 - 2y_{ij_2})(1 - y_{ij_1}) + b_2(1 - 2y_{ij_2}) y_{ij_1}) \right) \Bigg) \end{aligned}$$

   where $b_l = \exp(-25\mu_{ij_l})$, $l = 1, 2$.

2. $\mu_{ij_1} = \mu_{ij_2}$:

$$\begin{aligned} \ln \Bigg(y_{ij_1} y_{ij_2} + \frac{1}{a_n} \Bigg(& \ln \left(\frac{1 + \exp(a_n) b}{1 + b} \right) (1 - y_{ij_1} - y_{ij_2}) \\ & + \frac{b(1 - \exp(a_n))}{(1 + \exp(a_n) b)(1 + b)} (1 - 2y_{ij_1} - 2y_{ij_2} + 4 y_{ij_1} y_{ij_2}) \Bigg) \Bigg) \end{aligned}$$

   where $b = \exp(-25\mu_{ij_1})$.

There will be occasions when cluster i has just one member y_1. In that case, we clearly cannot take a bivariate loglikelihood but just replace it with the standard univariate density, π_u

$$\pi_u(y_{i1}; \boldsymbol{\beta}, a_n, u_i) = p_{i1}^{y_{i1}}(1 - p_{i1})^{1-y_{i1}}$$

where u_i is the random effect for the component and cluster and

$$p_{i1} = \frac{1}{(1 + \exp(-a_n(u_i - 1 + \mu_{i1})))} \quad l = 1, 2 :$$

for nuisance parameter a. In that case the loglikelihood for the marginal component is:

$$\ln\left(y_{i1} + \frac{1 - 2y_{i1}}{a_n} \ln\left(\frac{1 + \exp(a_n)b}{1 + b}\right)\right)$$

where $b = \exp(-25\mu_{i1})$.

All integrations in this Appendix have been checked using Mathematica (Wolfram Research, Inc, 2008). One can then differentiate the loglikelihoods analytically to get estimating functions etc. However, numerical differentiation as described in Section 2.7.1 performs equally well.

Appendix D

The Partially Dependent Weighting Scheme Minimises Parameter Variance

We consider the class of estimating functions whose components are independent at $\boldsymbol{\theta}_{\mathsf{G}}$. We show that the weighting scheme described in Section 3.3.2

$$\begin{aligned} \psi^*_w(\theta) &= \sum_{j=1}^{q} W^*_{\mathcal{C}} \psi_{\mathcal{C}}(\theta) \\ &= -\sum_{\mathcal{C} \in \mathcal{C}} \mathsf{E}[\psi'_{\mathcal{C}}(\theta_{\mathsf{G}})]^T \mathsf{Var}[\psi_{\mathcal{C}}(\theta_{\mathsf{G}})]^{-1} \psi_{\mathcal{C}}(\theta) \end{aligned} \tag{D.1}$$

is the most efficient in that class at $\boldsymbol{\theta}_{\mathsf{G}}$, ie the parameter estimator has minimal variance (or maximised sandwich information) over the class. We follow the approach outlined in Crowder (1986, Theorem 4.1) and extended as suggested but not worked through in Crowder (1987, Section 5). We adopt the same strategy with respect to zero elements in component estimating function as we did for the BWEF weighting scheme in Section 3.3.1 - namely, we delete them before inversion of any matrix derived from those components and restore them afterwards. We also need to make Assumption 19 - in this case $\boldsymbol{C}^0_S(\theta_{\mathsf{G}})$ (ie after we have removed the zero rows and columns) will be block diagonal and so the assumption will force each of those blocks to be nonsingular which is what we require. Finally, we assume that the appropriate moments of ψ_C for $C \in \mathcal{C}$ exist.

Consider any estimating function (*EF*) weighting scheme, ψ_w

$$\psi_w(\theta) \quad = \quad \sum_{j=1}^{q} W_C \psi_C(\theta)$$

and let

$$\boldsymbol{L} = \begin{pmatrix} \mathsf{Var}[\psi_w(\theta_{\mathbf{G}})] & \mathsf{E}[\psi'_w(\theta_{\mathbf{G}})]^T \\ \mathsf{E}[\psi'_w(\theta_{\mathbf{G}})] & -\mathsf{E}[\psi^{*'}_w(\theta_{\mathbf{G}})] \end{pmatrix}.$$

where $'$ represents the derivative. Note that for any weighted component of ψ^*_w, $W^*_C\psi_C$, we have

$$W^*_C\psi_C(\theta_{\mathbf{G}}) \quad = \quad -\mathsf{E}[\psi'_C(\theta_{\mathbf{G}})]^T \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{-1}\psi_C(\theta_{\mathbf{G}})$$

from (D.1) so that

$$\begin{aligned} \mathsf{Var}[W^*_C\psi_C(\theta_{\mathbf{G}})] &= \mathsf{Var}\left[\mathsf{E}[\psi'_C(\theta_{\mathbf{G}})]^T \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{-1}\psi_C(\theta_{\mathbf{G}})\right] \\ &= \mathsf{E}[\psi'_C(\theta_{\mathbf{G}})]^T \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{-1}\mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]\mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{-1}\mathsf{E}[\psi'_C(\theta_{\mathbf{G}})] \\ &= \mathsf{E}[\psi'_C(\theta_{\mathbf{G}})]^T \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{-1}\mathsf{E}[\psi'_C(\theta_{\mathbf{G}})] \\ &= -\mathsf{E}[W^*_C\psi'_C(\theta_{\mathbf{G}})] \end{aligned} \tag{D.2}$$

and therefore

$$\begin{aligned} \mathsf{E}[\psi^{*'}_w(\theta_{\mathbf{G}})] &= \sum_{C\in\mathcal{C}} \mathsf{E}[W^*_C\psi'_C(\theta_{\mathbf{G}})] \quad \text{from (D.1)} \\ &= -\sum_{C\in\mathcal{C}} \mathsf{Var}[W^*_C\psi_C(\theta_{\mathbf{G}})] \quad \text{from (D.2).} \end{aligned} \tag{D.3}$$

We take $\boldsymbol{z} = (z_1, z_2)^T$ a $2p \times 1$ vector with each element having dimension p and consider

$$\begin{aligned}
\boldsymbol{z}^T \boldsymbol{L} \boldsymbol{z} &= z_1^T \mathsf{Var}[\psi_w(\theta_{\mathbf{G}})] z_1 + 2 z_1^T \mathsf{E}[\psi_w'(\theta_{\mathbf{G}})]^T z_2 - z_2^T \mathsf{E}[\psi_w^{*'}(\theta_{\mathbf{G}})] z_2 \\
&= z_1^T \mathsf{Var}[\psi_w(\theta_{\mathbf{G}})] z_1 + 2 z_1^T \mathsf{E}[\psi_w'(\theta_{\mathbf{G}})]^T z_2 + z_2^T \sum_{C \in \mathcal{C}} \mathsf{Var}[W_C^* \psi_C(\theta_{\mathbf{G}})] z_2 \\
&\quad \text{from (D.3)} \\
&= \sum_{C \in \mathcal{C}} (z_1^T W_C \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})] W_C^T z_1 + 2 z_1^T W_C \mathsf{E}[\psi_C'(\theta_{\mathbf{G}})] z_2 + \\
&\quad z_2^T \mathsf{E}[\psi_C'(\theta_{\mathbf{G}})]^T \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{-1} \mathsf{E}[\psi_C'(\theta_{\mathbf{G}})] z_2) \\
&\quad \text{as EF components taken to be independent} \\
&= \sum_{C \in \mathcal{C}} \big((z_1^T W_C \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{1/2} + z_2^T \mathsf{E}[\psi_C'(\theta_{\mathbf{G}})]^T \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{-1/2}) \,. \\
&\quad (z_1^T W_C \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{1/2} + z_2^T \mathsf{E}[\psi_C'(\theta_{\mathbf{G}})]^T \mathsf{Var}[\psi_C(\theta_{\mathbf{G}})]^{-1/2})^T \big) \\
&\geq 0
\end{aligned}$$

where matrix square roots are taken so as to be symmetric (Horn and Johnson, 1987, theorem 7.2.6, page 405), so that $\boldsymbol{L}$ is semi-positive definite.

Now, by the process known as sweeping, if $\boldsymbol{L}$ is positive semidefinite (*PSD*) then so is:

$$\begin{pmatrix} \mathsf{Var}[\psi_w(\theta_{\mathbf{G}})]^{-1} & \mathsf{Var}[\psi_w(\theta_{\mathbf{G}})]^{-1} \mathsf{E}[\psi_w'(\theta_{\mathbf{G}})]^T \\ -\mathsf{E}[\psi_w'(\theta_{\mathbf{G}})] \mathsf{Var}[\psi_w(\theta_{\mathbf{G}})]^{-1} & -\mathsf{E}[\psi_w^{*'}(\theta_{\mathbf{G}})] - \mathsf{E}[\psi_w'(\theta_{\mathbf{G}})] \mathsf{Var}[\psi_w(\theta_{\mathbf{G}})]^{-1} \mathsf{E}[\psi_w'(\theta_{\mathbf{G}})]^T \end{pmatrix}$$

(see Appendix E) or $\boldsymbol{M}$, say. Sweeping is a technique developed for finding matrix inverses and determinants, for instance, in an incremental manner, usually on a computer (see for instance Beaton, 1964). The partially swept matrix $\boldsymbol{M}$ has had the sweeping process applied to the upper left block of $\boldsymbol{L}$. The use of sweeping replaces the use of the Cauchy-Schwarz inequality typically found in proofs such as this, see for instance (1.49), but allows us to work with an unknown **G**. Note that in Song (2007) a weighting scheme similar to the one under consideration here is compared with the data generating mechanism, **G**, and found to be as efficient. Unfortunately, it makes the component independence assumption tacitly and so the proof is not valid.

Clearly, if $\boldsymbol{M}$ is PSD then so must the bottom right hand corner of $\boldsymbol{M}$

$$-\mathsf{E}[\psi_w^{*\prime}(\theta_{\mathbf{G}})] - \mathsf{E}[\psi_w'(\theta_{\mathbf{G}})]^T \mathsf{Var}[\psi_w(\theta_{\mathbf{G}})]^{-1} \mathsf{E}[\psi_w'(\theta_{\mathbf{G}})]$$

be (for any vector z_2 for the latter, use $(0, z_2)$ for the former). Thus the sandwich information is greater in ψ_w^* than in any other estimating function weighting scheme. Asymptotically, after the usual normalisation, observed quantities converge to their expected values, as discussed in Section 1.4, and so we can use this proof for asymptotic optimality.

Appendix E

Swept PSD Matrix Is Still PSD

Let

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{S} \\ \boldsymbol{S}^T & \boldsymbol{U} \end{pmatrix}$$

where $\boldsymbol{R}$ is symmetric, be SPD so that for any $\boldsymbol{z}^* = (\boldsymbol{z}_1^*, \boldsymbol{z}_2^*)$

$$\boldsymbol{z}_1^{*^T} \boldsymbol{R} \boldsymbol{z}_1^* + 2\boldsymbol{z}_1^{*^T} \boldsymbol{S} \boldsymbol{z}_2^* + \boldsymbol{z}_2^{*^T} \boldsymbol{U} \boldsymbol{z}_2^* \geq 0.$$

Consider

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{R}^{-1} & \boldsymbol{R}^{-1}\boldsymbol{S} \\ -\boldsymbol{S}^T\boldsymbol{R}^{-1} & \boldsymbol{U} - \boldsymbol{S}^T\boldsymbol{R}^{-1}\boldsymbol{S} \end{pmatrix}$$

which to be SPD requires that for any $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2)$

$$\boldsymbol{z}_1^T \boldsymbol{R}^{-1} \boldsymbol{z}_1 + \boldsymbol{z}_2^T (\boldsymbol{U} - \boldsymbol{S}^T \boldsymbol{R}^{-1} \boldsymbol{S}) \boldsymbol{z}_2 \geq 0.$$

If we now transform by

$$\begin{aligned} \boldsymbol{z}_1 &= \boldsymbol{R}\boldsymbol{z}_1^* + \boldsymbol{S}\boldsymbol{z}_2^* \\ \boldsymbol{z}_2 &= \boldsymbol{z}_2^* \end{aligned}$$

then for SPD we require that

$$\begin{aligned} & \boldsymbol{z}_1^{*^T}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{z}_1^* + 2\boldsymbol{z}_1^{*^T}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{S}\boldsymbol{z}_2^* + \boldsymbol{z}_2^{*^T}(\boldsymbol{U} - \boldsymbol{S}^T\boldsymbol{R}^{-1}\boldsymbol{S} + \boldsymbol{S}^T\boldsymbol{R}^{-1}\boldsymbol{S})\boldsymbol{z}_2^* \\ = \; & \boldsymbol{z}_1^{*^T}\boldsymbol{R}\boldsymbol{z}_1^* + 2\boldsymbol{z}_1^{*^T}\boldsymbol{S}\boldsymbol{z}_2^* + \boldsymbol{z}_2^{*^T}\boldsymbol{U}\boldsymbol{z}_2^* \\ \geq \; & 0 \end{aligned}$$

which is true for any $\boldsymbol{z}^* = (\boldsymbol{z}_1^*, \boldsymbol{z}_2^*)$.

Appendix F

Example Where $\boldsymbol{J}$ Is Singular

We consider the example where data of dimension m are generated from a multivariate normal distribution, $\mathbf{G}$, with zero means, variances $\sigma_1^2, \ldots, \sigma_m^2$ and exchangeable correlation ρ. Our composite surrogate, $\mathbf{F}_K$, is unweighted and has components that are bivariate and marginal for $\mathbf{G}$, and thus bivariate normal with zero means, appropriate σ_i^2s and exchangeable correlation ρ. We will show that the second derivative with respect to the weights of the KLD between $\mathbf{G}$ and $\mathbf{F}_K$, $\boldsymbol{J}$, at $\theta_{\mathbf{G}}$ (omitted subsequently to simplify notation) is always singular, indeed it is a matrix with equal entries throughout.

From (4.10), we see that the inverse of the covariance matrix for $\mathbf{F}_K$ has the form

$$\boldsymbol{\Sigma}_{\mathbf{F}}^{-1} = \boldsymbol{D}(\sigma)^{-1}\boldsymbol{A}^{-1}\boldsymbol{D}(\sigma)^{-1}$$

where $\boldsymbol{D}(\sigma)$ is the $m \times m$ diagonal matrix whose entries are $\sigma_1, \ldots, \sigma_m$ and $\boldsymbol{A}^{-1}$ is a matrix with diagonal entries that are identical and off diagonal entries that are also identical, all entries being functions of ρ but not the σ_i^2s. As a consequence

$$\boldsymbol{\Sigma}_{\mathbf{F}} = \boldsymbol{D}(\sigma)\boldsymbol{A}\boldsymbol{D}(\sigma)$$

where, similarly, $\boldsymbol{A}$ is a matrix with diagonal entries that are identical and off diagonal entries that are also identical, all entries being functions of ρ but not the σ_i^2s. For any pair of distinct entries from the data vector, say $\boldsymbol{y}^{ij} = (y_i, y_j)$, there is a component of the composite whose likelihood function is

$$\ln f_{ij}(\boldsymbol{y}^{ij}) = -ln(2\pi) - \frac{\ln|\boldsymbol{\Sigma}_{\mathsf{F}_{ij}}|}{2} - \frac{(\boldsymbol{y}^{ij})^T\boldsymbol{\Sigma}_{\mathsf{F}_{ij}}^{-1}\boldsymbol{y}^{ij}}{2} \tag{F.1}$$

where $\boldsymbol{\Sigma}_{\mathsf{F}_{ij}}$ is the matrix consisting of the ith and jth rows and columns of $\boldsymbol{\Sigma}_{\mathsf{F}}$ so that

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathsf{F}_{ij}} &= \boldsymbol{D}(\sigma_i,\sigma_j)\boldsymbol{A}_{ij}\boldsymbol{D}(\sigma_i,\sigma_j)\\ \boldsymbol{\Sigma}_{\mathsf{F}_{ij}}^{-1} &= \boldsymbol{D}(\sigma_i,\sigma_j)^{-1}\boldsymbol{A}_{ij}^{-1}\boldsymbol{D}(\sigma_i,\sigma_j)^{-1} \end{aligned}\tag{F.2}$$

where $\boldsymbol{D}(\sigma_i,\sigma_j)$ is the diagonal matrix with entries σ_i,σ_j and $\boldsymbol{A}_{ij}$ is the matrix consisting of the ith and jth rows and columns of $\boldsymbol{A}$, which will still have the same features as $\boldsymbol{A}$ in miniature. Note that $\boldsymbol{A}_{ij}$ will be the same for all pairs as it is just a function of ρ with the same entries for all pairs (i,j) and so we shall refer to it as $\boldsymbol{A}_B$.

As we have seen in Theorem 3.4.1, $\boldsymbol{J}$ is the covariance matrix of the $\{\ln f_{ij}(Y^{ij})\}$s under F_K. So, for not necessarily distinct pairs (i,j) and (k,l)

$$\begin{aligned}
&\mathsf{Cov}_{\mathsf{F}_K}[\ln f_{ij}(Y^{ij}),\ln f_{kl}(Y^{kl})]\\
&= \mathsf{Cov}_{\mathsf{F}_K}\left[-ln(2\pi) - \frac{\ln|\boldsymbol{\Sigma}_{\mathsf{F}_{ij}}|}{2} - \frac{(\boldsymbol{y}^{ij})^T\boldsymbol{\Sigma}_{\mathsf{F}_{ij}}^{-1}\boldsymbol{y}^{ij}}{2}\right.\\
&\qquad \left., -ln(2\pi) - \frac{\ln|\boldsymbol{\Sigma}_{\mathsf{F}_{kl}}|}{2} - \frac{(\boldsymbol{y}^{kl})^T\boldsymbol{\Sigma}_{\mathsf{F}_{kl}}^{-1}\boldsymbol{y}^{kl}}{2}\right] \quad \text{by (F.1)}\\
&= \mathsf{Cov}_{\mathsf{F}_K}\left[-\frac{(\boldsymbol{y}^{ij})^T\boldsymbol{\Sigma}_{\mathsf{F}_{ij}}^{-1}\boldsymbol{y}^{ij}}{2}, -\frac{(\boldsymbol{y}^{kl})^T\boldsymbol{\Sigma}_{\mathsf{F}_{kl}}^{-1}\boldsymbol{y}^{kl}}{2}\right]\\
&= \frac{1}{4}\mathsf{Cov}_{\mathsf{F}_K}\left[(\boldsymbol{y}^{ij})^T\boldsymbol{D}(\sigma_i,\sigma_j)^{-1}\boldsymbol{A}_B^{-1}\boldsymbol{D}(\sigma_i,\sigma_j)^{-1}\boldsymbol{y}^{ij}\right.\\
&\qquad \left., (\boldsymbol{y}^{kl})^T\boldsymbol{D}(\sigma_k,\sigma_l)^{-1}\boldsymbol{A}_B^{-1}\boldsymbol{D}(\sigma_k,\sigma_l)^{-1}\boldsymbol{y}^{kl}\right] \quad \text{by (F.2)}\\
&= \frac{1}{4}\mathsf{Cov}_{\mathsf{F}_K}[(\boldsymbol{z}^{ij})^T\boldsymbol{A}_B^{-1}\boldsymbol{z}^{ij}, (\boldsymbol{z}^{kl})^T\boldsymbol{A}_B^{-1}\boldsymbol{z}^{kl}]
\end{aligned}$$

say, where $z_i = y_i/\sigma_i$ for $1 \le i \le m$, etc. As we have seen in Section 4.3,

$$\mathsf{Cov}[\boldsymbol{Z}_a^T\boldsymbol{A}\boldsymbol{Z}_a, \boldsymbol{Z}_b^T\boldsymbol{B}\boldsymbol{Z}_b] = 2\mathsf{tr}(\boldsymbol{A}\boldsymbol{\Lambda}_a\boldsymbol{B}\boldsymbol{\Lambda}_b)$$

for symmetric $\boldsymbol{A}$ and $\boldsymbol{B}$, and $Z_a \sim \mathsf{MVN}(\boldsymbol{0},\boldsymbol{\Lambda}_a)$ and $Z_b \sim \mathsf{MVN}(\boldsymbol{0},\boldsymbol{\Lambda}_b)$. Therefore, we have

$$\begin{aligned}\mathsf{Cov}_{\mathbf{F}_K}[\ln f_{ij}(Y^{ij}), \ln f_{kl}(Y^{kl})] &= \frac{1}{4}\mathsf{Cov}_{\mathbf{F}_K}[(\boldsymbol{z}^{ij})^T \boldsymbol{A}_B^{-1} \boldsymbol{z}^{ij}, (\boldsymbol{z}^{kl})^T \boldsymbol{A}_B^{-1} \boldsymbol{z}^{kl}] \\ &= \frac{1}{2}\mathsf{tr}(\boldsymbol{A}_B \boldsymbol{\Lambda} \boldsymbol{A}_B \boldsymbol{\Lambda})\end{aligned}$$

where $\boldsymbol{\Lambda} = \mathsf{Cov}_{\mathbf{F}_K}[\boldsymbol{z}_{ij}] = \mathsf{Cov}_{\mathbf{F}_K}[\boldsymbol{z}_{kl}]$ is a 2×2 matrix, whose entries are just rows and columns of $\boldsymbol{A}$. Thus, every entry in the covariance matrix of the $\{\ln f_{ij}(Y^{ij})\}$s under $\mathbf{F}_K$ is identical and the matrix is singular. This confirms the numerical results we found in Section 3.4.3.

Bibliography

Aerts, M. and Claeskens, G. (1999). Bootstrapping pseudolikelihood models for clustered binary data. *Biometrika*, 77(3):485–497.

Ashford, J. R. and Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics*, 26(3):535–546.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Number 52 in Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Barry, A., Burney, S., and Bhatti, M. (1997). Optimum influence of initial observations in regression models with AR(2) errors. *Applied Mathematics and Computation*, 82:57–65.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College, London.

Beaton, A. E. (1964). The use of special matrix operators in statistical calculus. Technical report, Educational Testing Service, University of Harvard.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society Series B*, 36:192–236.

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association*, 107(479):268–280.

Bhat, B. R. (1974). On the method of maximum-likelihood for dependent observations. *Journal of the Royal Statistical Society Series B*, 36:48–53.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, New York.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis - the Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer-Verlag, New York, 2nd edition.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. Springer-Verlag, New York, revised edition.

Chandler, R. (2004). Moment-based inference for stochastic-mechanistic models. Internal Report 7, DEFRA. Project: improved methods for national spatial-temporal rainfall and evaporation modelling for BSM.

Chandler, R. E. and Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, 94:167–183.

Chandler, R. E., Isham, V., Bellone, E., Yang, C., and Northrop, P. (2007). Space-time modelling of rainfall for continuous simulation. In Finkenstädt, B., Held, L., and Isham, V., editors, *Statistical Methods for Spatio-Temporal Systems*, number 107 in Monographs on Statistics and Applied Probability. CRC, Boca Raton.

Chandrasekar, B. and Kale, B. K. (1984). Unbiased statistical estimation functions for parameters in presence of nuisance parameters. *Journal of Statistical Planning and Inference*, 9(1):45–54.

Chappell, M. A., Groves, A., and Woolrich, M. W. (2007). The FMRIB variational Bayes tutorial: variational Bayesian inference for a non-linear forward model. Internal report, Oxford Centre for Functional MRI of the Brain. Available at http://www.fmrib.ox.ac.uk/analysis/techrep/.

Copas, J. and Eguchi, S. (2010). Likelihood for statistically equivalent models. *Journal of the Royal Statistical Society Series B*, 72(2):193–217.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society Series B*, 24(2):406–424.

Cox, D. R. (2006). *Principles of Statistical Inference*. CUP, Cambridge.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Number 32 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2nd edition.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies - Models, analysis and interpretation*. Number 67 in Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Crowder, M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory*, 2(3):305–330.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika*, 74(3):591–597.

Crowder, M. (2001). On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society Series B*, 63(1):55–62.

Davidson, R. and McKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press, New York.

Davis, R. A. and Yau, C. Y. (2011). Comments on pairwise likelihood in time series models. *Statistica Sinica*, 21:255–277.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.

Donnelly, T. G. (1973). Algorithm 462: Bivariate normal distribution. *Communications of the ACM*, 16(10):638.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, 2nd edition.

Foutz, R. V. and Srivastava, R. C. (1977). The performance of the likelihood ratio test when the model is incorrect. *The Annals of Statistics*, 5(6):1183–1194.

Garthwaite, P. H., Jolliffe, I. T., and Jones, B. (2006). *Statistical Inference*. Oxford University Press, Oxford, 2nd edition.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8(Supplement: Special Issue on Econometric Inference Using Simulation Techniques):S85–S118.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Heagerty, P. J. and Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, 95(449):197–211.

Horn, R. A. and Johnson, C. R. (1987). *Matrix Analysis*. Cambridge University Press, Cambridge, 1st with corrections edition.

Huber, P. J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*. University of California Press, Berkeley.

Jesus, J. and Chandler, R. E. (2011). Estimating functions and the generalized method of moments. *Interface Focus*, 1:871–885.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Number 73 in Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, 100:670–685.

Joseph, B. and Durairajan, T. (1991). Equivalence of various optimality criteria for estimating functions. *Journal of Statistical Planning and Inference*, 27:355–360.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B*, 63(3):425–464.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69(1):19–27.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions*, volume 1. Wiley, New York.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis*. Oxford University Press, Oxford, revised edition.

Kuk, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika*, 94(4):939–952.

le Cessie, S. and van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, 43(1):95–108.

Lee, P. M. (2004). *Bayesian Statistics An Introduction*. Hodder Arnold, London, third edition.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221–239.

Lindsay, B. G., Yi, G. Y., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21:71–105.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. Wiley, New York.

Lourmas, G. and Chandler, R. (2006). Routines for fitting Poisson cluster rainfall models using a method of moments. Software copyright University College London, available from http://www.homepages.ucl.ac.uk/ ucakarc/work/momfit.html.

Lu, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I., and Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63:935–941.

Mardia, K. V., Kent, J. T., Hughes, G., and Taylor, C. C. (2009). Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika*, 96(4):975–982.

Maronna, R. R., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics - Theory and Methods*. Wiley, Chichester.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic Forms in Random Variables*. Number 126 in Statistics: Textbooks and Monographs. Dekker, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman and Hall, London.

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society Series B*, 52:325–344.

McLeish, D. L. and Small, C. G. (1980). *The Theory and Applications of Statistical Inference Functions*. Number 44 in Lecture Notes in Statistics. Springer, New York.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):325–344. 449-465.

Pace, L., Salvan, A., and Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21:129–148.

Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010). Likelihood based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489):263–277.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Royall, R. and Tsou, T.-S. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society Series B*, 62(2):391–404.

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, 54(2):221–226.

Schott, J. R. (1997). *Matrix Analysis for Statistics*. Wiley, New York.

Smith, R. L. (1989). A survey of nonregular problems. *Bulletin of the International Statistical Institute*, 53:353–372.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer, New York.

Stafford, J. E. (1996). A robust adjustment of the profile likelihood. *Annals of Statistics*, 24(1):336–352.

Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Annals of Statistics*, 8(6):1375–1381.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187–202.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92:1–28.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92:519–528.

Vasdekis, V. G. S., Cagnone, S., and Moustaki, I. (2012). A composite likelihood inference in latent variable models for ordinal longitudinal reponsess. *Psychometrika*, 77(3):425–441.

Viraswami, K. and Reid, N. (1998a). Higher-order asymptotics under model misspecification. *The Canadian Journal of Statistics*, 24(2):263–278.

Viraswami, K. and Reid, N. (1998b). A note on the likelihood-ratio statistic under model misspecification. *The Canadian Journal of Statistics*, 26(1):161–168.

Wei, B.-C. (1997). *Exponential Family Nonlinear Models*, volume 130 of *Lecture Notes in Statistics*. Springer, Singapore.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Wolfram Research, Inc (2008). *Mathematica*. Wolfram Research, Inc, Champaign, Illinois, version 7.0 edition.

Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference*, 141(9):3047–3054.

Yan, Z., Bate, S., Chandler, R. E., Isham, V., and Wheater, H. (2002). An analysis of daily maximum wind speed in Northwestern Europe using generalized linear models. *Journal of Climate*, 15:2072–2088.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge University Press, Cambridge.