



University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

This paper is made available online in accordance with publisher policies. Please scroll down to view the document itself. Please refer to the repository record for this item and our policy information available from the repository home page for further information.

To see the final version of this paper please visit the publisher's website. Access to the published version may require a subscription.

Author(s): JB Copas

Article Title: Likelihood for Statistically Equivalent Models

Year of publication: 2009

Link to published article:

<http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2009/paper09-03>

Publisher statement: None

# Likelihood for Statistically Equivalent Models

John Copas\*

University of Warwick, UK

and Shinto Eguchi

Institute of Statistical Mathematics, Tokyo, Japan

## SUMMARY

In likelihood inference we usually assume the model is fixed and then base inference on the corresponding likelihood function. Often however the choice of model is rather arbitrary, and there may be other models which fit the data equally well. We study robustness of likelihood inference over such “statistically equivalent” models, and suggest a simple “envelope likelihood” to capture this aspect of model uncertainty. Robustness depends critically on how we specify the parameter of interest. Some asymptotic theory is presented, illustrated by three examples.

**KEY WORDS:** Likelihood inference, robust likelihood, model uncertainty, parameter of interest.

---

\*Corresponding author: [jbc@stats.warwick.ac.uk](mailto:jbc@stats.warwick.ac.uk); phone 44(0)2476523370; fax 44(0)2476524532

# 1 Introduction

Most statistical methods are based directly or indirectly on likelihood for a parametric model. Thus if model  $f$  asserts that the data  $x_1, x_2, \dots, x_n$  are a random sample from  $f(x, \theta)$ , then the likelihood for  $\theta$  is

$$l_f(\theta) = \sum_{i=1}^n \log f(x_i, \theta) . \quad (1)$$

Parameter  $\theta$  is just a convenient way of indexing the model, in practice we will usually be interested in the value of some specific scalar population parameter  $\phi$ . If the population actually is described by  $f$  then  $\phi$  is just a function of  $\theta$ , say  $\phi = \phi(\theta)$ . If  $\theta$  is a scalar and  $\phi(\theta)$  is invertible, then the model likelihood for  $\phi$  is just (1) re-expressed as a function of  $\phi$ . More generally the profile likelihood for  $\phi$  is

$$L_f(\phi) = \sup_{\theta: \phi(\theta)=\phi} l_f(\theta) . \quad (2)$$

As we will only be concerned with first-order asymptotic methods (large  $n$ ), we make no notational distinction between actual and profile likelihoods.

We use suffix  $f$  in this notation to emphasize that likelihood, and hence methods derived from it, depend on the model as well as on the data. Conventional asymptotic sampling theory results first assume that  $f$  is fixed, and then address uncertainty in  $x$  as described by  $f$ . But what about uncertainty in the model itself? When the data result from explicit random sampling or from a designed experiment, the choice of  $f$  is sometimes self-evident from the context. Usually however, and always for observational data, a model is chosen for reasons which are never entirely convincing, such as mathematical convenience (when we can work things out explicitly) or custom and practice (the model has been used by previous researchers). Traditionally, we confirm that our analysis is sensible by checking that model  $f$  gives an acceptable fit to the data, conveniently forgetting that for any given set of data there will always be a multitude of other models which also fit the data just as well. If these models all give the same (or roughly the same) inference about  $\phi$  then this indicates a comforting degree of robustness to modeling assumptions. If, on the other hand, these models lead to different conclusions about  $\phi$  then this suggests a sensitivity to arbitrary modeling assumptions which we need to take into account. The aim of this paper is to suggest how this might be done.

For any given model  $f$  we will be interested in the set of possible distributions  $g$  which are “statistically equivalent” to  $f$  in the sense that if we were to test the hypothesis that the data were sampled from  $g$  rather than  $f$  we would have no significant evidence one way or the other. When  $n$  is large this means that  $f$  and  $g$  must be close together in the sense that if they were substantially different then this would be immediately evident from the data. We are thus only concerned with *local* model mis-specification, and it is this which allows us to base our discussion on relatively simple asymptotic approximations. In this discussion we are assuming that all models being considered are fully identified, so we are not including cases such as missing data when different non-ignorable models might give exactly the same distribution for what is observed.

It makes no sense comparing likelihood functions for  $\theta$  directly, since a model parameter  $\theta$  only has meaning within the assumptions of that particular model. Hence the need to calibrate each likelihood in terms of a common population parameter  $\phi$ . The nature of  $\phi$  turns out to be crucial for the study of robustness. For example if  $f$  is normal and  $\phi$  is the population mean, then the maximum likelihood estimate (MLE) of  $\phi$  is the sample mean, which remains a sensible estimate whatever distribution actually generated the data. By contrast, if  $f$  is normal and  $\phi$  is skewness (standardized third moment) then the MLE of  $\phi$  is identically zero (as the model asserts *a priori* that the distribution is symmetrical). In this case zero is clearly not a sensible estimate if we want to entertain the possibility that the distribution is actually skewed. Later in the paper we will define a correlation quantity  $\rho$  which shows that these two examples are extreme cases. In the first example  $\rho = 1$  means that undetectable mis-specification is unimportant, in the second  $\rho = 0$  means that undetectable mis-specification is very important.

In section 2 of the paper we describe the basic set-up. For any given model  $f$  we define  $\mathcal{G}$  to be the set of distributions which can be considered statistically equivalent to  $f$ . We then look at estimates and likelihoods for  $\phi$  for models within  $\mathcal{G}$ . By looking at appropriate bounds we suggest a “worst case” likelihood which allows for uncertainty of  $g \in \mathcal{G}$  as well as uncertainty through sampling variation in the data. We show how the underlying geometry of the distributions involved gives further insight and motivation for many of our developments. For clarity of exposition we restrict Section 2 to a particular prescription for  $\phi$ . This is generalized in Section 3, which shows that the same idea in fact works under a much more general setting. In Section 4 we look more closely at discrete distributions and show that tighter bounds for the worst-case likelihood are possible in this case. Further comments and discussion are included in the final Section 5.

Three examples are used to illustrate the paper. In Section 2 we discuss the use of the log-normal model for estimating the mean of a right-skewed distribution. In Section 3 we look at a parametric survival model for estimating and comparing survival functions, using data from a cancer clinical trial. Then in Section 4 we discuss estimating the average causal effect of a binary treatment from an observational study with discrete covariates. Further generalizations are mentioned in Section 5.

The literature on likelihood inference is extensive: many fundamental aspects have recently been reviewed in the masterful text by Cox (2006). A useful account of wider aspects of likelihood methods is Severini (2000). Important aspects of robustness are discussed in the classic text by Huber (1981). Several papers discuss sensitivity aspects of likelihood inference, in particular two papers published in this Journal, Gustafson (2001) and Royall and Tsou (2003). Gustafson (2001) highlights similar distinctions between the object of inference ( $\theta$ ) and the object of interest ( $\phi$ ), and between the true distribution ( $g$ ) and the model distribution ( $f$ ). Gustafson’s paper suggests how the inferential effects of mis-specifying  $L_f$  can be related to the informational distance between  $f$  and  $g$ . Royall and Tsou (2003) study the behaviour of mis-specified likelihoods as  $n \rightarrow \infty$ . They suggest that  $L_f$  should be scaled in such a way that the good asymptotic properties expected of likelihood functions are retained even when the data are sampled from some other distribution. Links with Royal and Tsou’s paper are discussed in more detail in Section 5.

There is nothing new in pointing out inherent uncertainty in models as well as in data. A recent on-line discussion on one of the statistical web-sites has debated whether we fit data

to models (the textbook approach: the model is fixed, the data are a random realization from the model), or fit models to data (first the data, then we explore models to help us understand the data). Our paper is an attempt to explore one possible approach to this dilemma. Other approaches, also discussed extensively in the literature, include model selection (where we allow the data to select from a predefined set of possible models), and Bayesian averaging (based on a prior distribution specifying prior uncertainty both within and between such a predefined set of models).

## 2 Basic theory and example

### 2.1 Notation and set-up

Let  $f(x, \theta)$  be the working model for a large sample of observations  $x_1, x_2, \dots, x_n$ . Familiar quantities for likelihood inference are the score and information functions

$$s(x, \theta) = \frac{\partial \log f(x, \theta)}{\partial \theta} \quad \text{and} \quad I(\theta) = -E_{\theta} \left\{ \frac{\partial^2 \log f(x, \theta)}{\partial \theta \partial \theta^T} \right\} = \text{Var}_{\theta} s(x, \theta) ,$$

with the MLE  $\hat{\theta}$  and its asymptotic variance given by

$$\bar{E} s(x, \hat{\theta}) = 0 \quad \text{and} \quad \text{Var}_{\theta}(\hat{\theta}) = \{nI(\theta)\}^{-1} . \quad (3)$$

The suffix  $\theta$  indicates expectation and variance with respect to  $f(x, \theta)$ , and  $\bar{E}$  indicates sample average over the data values  $x_1$  to  $x_n$  (expectation over the empirical distribution).

Although we use  $f$  as our working model, we want to entertain the possibility that the data were in fact generated from some other distribution  $g = g(x)$ . Specifying  $g(x)$  is equivalent to specifying the log likelihood ratio function  $\log\{g(x)/f(x, \theta)\}$ , which we express in terms of a scalar  $\epsilon$  and a function  $u = u(x, \theta)$  so that we can write  $g$  as

$$g_u = g_u(x; \epsilon, \theta) = \exp\{\epsilon u(x, \theta) - K(\epsilon, \theta)\} f(x, \theta) , \quad (4)$$

where the integrating factor  $K$  (assumed finite) is the cumulant generating function

$$K(\epsilon, \theta) = \log E_{\theta} \exp\{\epsilon u(x, \theta)\} .$$

In (4) there is a redundancy in notation for the location and scale of  $u$ ,  $\epsilon$  and  $K$ , so we can assume without loss of generality that  $u$  is standardized to have zero mean and unit variance under  $f$ . There is also a redundancy in notation between  $u$  and  $\theta$  in the sense that for any small  $\delta$  we can write

$$f(x, \theta + \delta) = \exp\{\delta^T s(x, \theta)\} f(x, \theta) + O(\delta^2) . \quad (5)$$

This means that in (4) we can compensate for a small shift in  $\theta$  by adjusting  $u$  by the addition or subtraction of a small multiple of the score function  $s$ . This redundancy is removed by insisting that  $u$  is orthogonal to  $s$ . The three constraints we assume on  $u$  are therefore

$$E_{\theta} u(x, \theta) = 0 \quad , \quad E_{\theta} u^2(x, \theta) = 1 \quad \text{and} \quad E_{\theta} \{u(x, \theta) s(x, \theta)\} = 0 . \quad (6)$$

For any function  $u$  satisfying (6), (4) defines  $g_u$  as a two parameter family of distributions which we interpret as an alternative model to  $f$ . The orthogonality constraint in (6) means that  $(\epsilon, \theta)$  are orthogonal model parameters in the sense that their joint information matrix (for fixed  $u$ ) is diagonal. Following Copas and Eguchi (2005) we think of  $u$  as the *direction* of mis-specification and  $\epsilon$  as the *magnitude* of mis-specification: when  $\epsilon = 0$  then  $g = f$ , but as  $\epsilon$  moves away from 0,  $g$  moves away from  $f$  in the direction  $u$ . The complete family of distributions defined by (4) can be thought of as a *tubular neighbourhood* surrounding  $f$  with “radius”  $\epsilon$ . See Section 2.5 below for further discussion of geometrical aspects.

We now ask whether the difference between the models  $f$  and  $g_u$  is sufficiently large to be detectable from the data. To do this, assume the data are generated by  $g_u$  (for some fixed direction  $u$ ) and use the data to test the null hypothesis that  $\epsilon = 0$ . The log likelihood ratio statistic divided by  $n$  is

$$\bar{E} \left\{ \log \frac{g_u(x; \epsilon, \theta)}{g_u(x; 0, \theta)} \right\} = \epsilon \bar{E}u(x, \theta) - K(\epsilon, \theta) .$$

For  $\epsilon$  near zero this is locally linear in  $\bar{E}u(x, \theta)$  since the moment constraints on  $u$  in (6) mean that  $K(\epsilon, \theta) = \frac{1}{2}\epsilon^2 + O(\epsilon^3)$ . From the orthogonality of  $\epsilon$  and  $\theta$ , the asymptotic locally most powerful test therefore replaces  $\theta$  by  $\hat{\theta}$  to give the asymptotic standardized test statistic

$$S_u = n^{\frac{1}{2}} \bar{E}u(x, \hat{\theta}) \sim_{\epsilon=0} N(0, 1) .$$

For a two sided level  $\alpha$  test we accept the null hypothesis if  $|S_u| \leq z_\alpha = \Phi^{-1}(1 - \alpha/2)$ , where  $\Phi$  is the standard normal distribution function. We therefore define the set  $\mathcal{G}$  of statistically equivalent models to be

$$\mathcal{G} = \{g_u : |S_u| \leq z_\alpha\} . \tag{7}$$

It is easy to see that  $E_{g_u}u(x, \theta) = \epsilon + O(\epsilon^2)$ , so when  $\epsilon$  is small but non-zero, the asymptotic power function of this test is

$$\Phi(-z_\alpha - n^{\frac{1}{2}}\epsilon) + \Phi(-z_\alpha + n^{\frac{1}{2}}\epsilon) .$$

For this to be bounded away from one as  $n \rightarrow \infty$ , so that even for large  $n$  we can expect  $\mathcal{G}$  not to be empty,  $\epsilon$  must be small, at most  $O(n^{-\frac{1}{2}})$ . This will be important in controlling the accuracy of the approximations discussed in the next subsection. We emphasize that the assumption that  $\epsilon = O(n^{-\frac{1}{2}})$  is merely a mathematical device for obtaining useful approximations: we are not assuming in any literal sense that as we obtain more data somehow our working model becomes more nearly correct. In studying model robustness of inference our aim is to focus just on those alternative models which we would not be able to distinguish empirically from  $f$ .

## 2.2 The parameter of interest

For any distribution  $g$ , let  $\phi$  be the solution of the equation

$$E_g a(x, \phi) = 0 \tag{8}$$

for some given estimating function  $a(x, \phi)$ . For the moment we assume that both  $\phi$  and  $a(x, \phi)$  are scalar. We take  $\phi$  to be the parameter of interest, a functional of  $g$  specified by the particular choice of  $a$ . Three examples of  $a(x, \phi)$  are

$$a(x, \phi) = \phi - x^k \quad , \quad a(x, \phi) = \begin{cases} 1 - p & \text{if } x \leq \phi \\ -p & \text{if } x > \phi \end{cases} \quad , \quad a(x, \phi) = \begin{cases} \phi & \text{if } x < t \\ \phi - 1 & \text{if } x \geq t \end{cases} \quad . \quad (9)$$

Respectively, these define  $\phi$  to be the  $k$ th moment ( $\phi = E_g x^k$ ), the  $p$ th quantile ( $P_g(x \leq \phi) = p$ ) and the survivor function ( $\phi = P_g(x \geq t)$ ). For further examples and discussion of estimating functions, see the review article by Desmond and Godambe (1998).

Under the working model (when  $g = f$  or  $\epsilon = 0$ ),  $\phi$  is just a scalar function  $\phi(\theta)$  of the parameter vector  $\theta$  defined by

$$E_\theta a\{x, \phi(\theta)\} = 0 \quad . \quad (10)$$

Note that in (8)  $\phi$  is invariant to the scaling of  $a$  by any arbitrary factor which does not depend on  $x$ , so we can assume from now on that  $a(x, \phi)$  is scaled so that

$$E_\theta \frac{\partial a\{x, \phi(\theta)\}}{\partial \phi} = 1 \quad . \quad (11)$$

Of the three examples of  $a(x, \phi)$  in (9), the first and third already satisfy (11). For the second,  $E_\theta a(x, \phi) = P_\theta(x \leq \phi) - p$ , so in this case  $a(x, \phi)$  needs to be scaled by dividing by  $f(\phi(\theta), \theta)$ , the model probability density function at  $x = \phi(\theta)$  (quantiles are only uniquely defined in the continuous case).

Differentiating (10) with respect to  $\theta$  and using (11) gives

$$\frac{d\phi(\theta)}{d\theta} = -E_\theta \{s(x, \theta) a(x, \phi(\theta))\} \quad .$$

Under model  $f$ , the MLE of  $\theta$  and its variance are given in (3). The corresponding MLE of  $\phi$  is therefore  $\hat{\phi}$  with asymptotic variance  $n^{-1}\sigma_\phi^2$ , where

$$\hat{\phi} = \phi(\hat{\theta}) \quad , \quad \sigma_\phi^2 = [E_\theta \{a(x, \phi(\theta))s(x, \theta)\}]^T I^{-1}(\theta) E_\theta \{a(x, \phi(\theta))s(x, \theta)\} \quad . \quad (12)$$

A natural comparison for parametric estimates of  $\phi$  is the non-parametric estimate  $\tilde{\phi}$  given by

$$\bar{E}a(x, \tilde{\phi}) = 0 \quad . \quad (13)$$

If the asymptotic variance (under model  $f$ ) of  $\tilde{\phi}$  is  $n^{-1}\sigma_a^2$ , and its relative efficiency is  $\rho^2$ , then using (11) we have

$$\sigma_a^2 = \text{Var}_\theta a\{x, \phi(\theta)\} \quad , \quad \rho^2 = \frac{\sigma_\phi^2}{\sigma_a^2} = \frac{\{E_\theta(as)\}^T I^{-1} E_\theta(as)}{\sigma_a^2} \quad . \quad (14)$$

Equation (14) shows that the efficiency parameter  $\rho$  is just the multiple correlation coefficient between  $a(x, \phi(\theta))$  and  $s(x, \theta)$ .

Note that many of the parameters in these and following expressions are functions of  $\theta$ , even if this is not made explicit in the notation. However, it turns out that for the asymptotic approximations to be discussed below it is only values of  $\theta$  within a local

neighbourhood of  $\hat{\theta}$  which concern us, so all we need are the values of these quantities at  $\theta = \hat{\theta}$ . Similarly, it is sufficient that the constraints (6) and (11) hold at  $\theta = \hat{\theta}$ .

The correlation  $\rho$  plays a key role in robustness. If  $\rho = \pm 1$ ,  $a(x, \phi)$  can be written as a linear function of the components of  $s(x, \theta)$ , so equation (13) is simply reproducing the usual likelihood equations in (3), hence  $\hat{\phi} = \tilde{\phi}$ . If  $\rho = 0$  then  $\sigma_\phi = 0$ , so the model is useless for estimating  $\phi$  — model  $f$  is essentially assigning the value of  $\phi$  in advance. Examples of these extremes were mentioned earlier in the Introduction.

Now suppose the data are generated by distribution  $g_u$  in (4). Then (8) gives  $\phi$  as

$$0 = E_{g_u} a(x, \phi) = E_\theta e^{\epsilon u} \left\{ a(x, \phi(\theta)) + (\phi - \phi(\theta)) \frac{\partial a}{\partial \phi} \right\} + O(\epsilon^2) .$$

Hence, if we let

$$\lambda_u = \sigma_\phi^{-1} E_\theta \{ a(x, \phi(\theta)) u(x, \theta) \} \quad (15)$$

then

$$\phi = \phi(\theta) - \epsilon \sigma_\phi \lambda_u + O(\epsilon^2) . \quad (16)$$

Note that the first-order bias in  $\phi$  is the product of three terms: the natural scale parameter  $\sigma_\phi$ , the mis-specification magnitude  $\epsilon$ , and the directional component  $\lambda_u$  depending on the mis-specification direction  $u$ . If  $\rho^2 = 1$  then  $\lambda_u = 0$  so there is no first-order bias. This is because  $a$  is then a linear function of  $s$  so the third constraint in (6), that  $u$  is orthogonal to  $s$ , means that  $u$  must also be orthogonal to  $a$ . The directional component of the bias is largest when

$$u(x, \theta) = u_a = \frac{a - \{E(as)\}^T I^{-1} s}{\sigma_a (1 - \rho^2)^{\frac{1}{2}}} , \quad (17)$$

which gives the bound

$$\lambda_u^2 \leq \lambda_M^2 = \frac{1 - \rho^2}{\rho^2} . \quad (18)$$

Proving this bound is an exercise in Lagrange multipliers: maximize  $E(au)$  over  $u$  subject to the constraints (6).

### 2.3 Likelihoods for locally mis-specified models

The (profile) likelihood for  $\phi$  under model  $f$  is  $L_f$  as already given in (2). This is most easily expressed in terms of the standardized parameter

$$\omega = \frac{n^{\frac{1}{2}} \{ \phi - \hat{\phi} \}}{\sigma_\phi} ,$$

the denominator  $\sigma_\phi$  and similar quantities being assumed from now on to be evaluated at  $\theta = \hat{\theta}$ . Then by standard likelihood asymptotics, the asymptotic likelihood for  $\phi$  under model  $f$  is simply (omitting irrelevant additive constants)

$$L_f(\phi) = -\frac{1}{2} \omega^2 . \quad (19)$$

Likelihood (19) is a consequence of the asymptotic standard normal sampling distribution of  $n^{\frac{1}{2}}(\hat{\phi} - \phi)/\sigma_\phi$ . Similarly, the corresponding standardized deviate  $n^{\frac{1}{2}}(\tilde{\phi} - \phi)/\sigma_a$  for the non-parametric estimate  $\tilde{\phi}$  is also asymptotically standard normal from (13). This inverts to a non-parametric (psuedo-) likelihood for  $\phi$ , re-written in terms of  $\omega$  as

$$L_{NP}(\phi) = -\frac{1}{2}\rho^2(\omega - \tilde{\omega})^2 \quad , \quad (20)$$

where

$$\tilde{\omega} = \frac{n^{\frac{1}{2}}\{\tilde{\phi} - \hat{\phi}\}}{\sigma_\phi} \quad .$$

Note that here, and throughout the paper, we are assuming that  $f(x, \theta)$  satisfies the regularity conditions necessary for the usual asymptotic approximations to apply.

Now suppose the data are generated by the model  $g_u$  in (4) for some given direction function  $u$ , with large  $n$  and  $\epsilon = O(n^{-\frac{1}{2}})$ . Then the likelihood under model  $g_u$  for  $(\epsilon, \theta)$  is

$$\begin{aligned} l_u(\epsilon, \theta) &= n\bar{E} \log g_u(x; \epsilon, \theta) \\ &= n\bar{E} \log f(x, \theta) + \epsilon n^{\frac{1}{2}} S_u - \frac{n}{2}\epsilon^2 + O(n^{-\frac{1}{2}}) \quad . \end{aligned}$$

From (16), the profile likelihood for  $(\epsilon, \phi)$  is therefore

$$\begin{aligned} L_u(\epsilon, \phi) &= \sup_{\{\theta | \phi(\theta) = \phi + \epsilon\sigma_\phi\lambda_u\}} l_u(\epsilon, \theta) + O(n^{-\frac{1}{2}}) \\ &= L_f\{\phi + \epsilon\sigma_\phi\lambda_u\} + \epsilon n^{\frac{1}{2}} S_u - \frac{n}{2}\epsilon^2 + O(n^{-\frac{1}{2}}) \quad . \end{aligned}$$

Using (19) this gives

$$L_u(\epsilon, \phi) = -\frac{1}{2}\omega^2 + n^{\frac{1}{2}}\epsilon(S_u - \lambda_u\omega) - \frac{n}{2}\epsilon^2(1 + \lambda_u^2) + O(n^{-\frac{1}{2}}) \quad .$$

Hence, omitting irrelevant constants, the asymptotic profile likelihood for  $\phi$  for any given  $u$  is

$$L_u(\phi) = \sup_{\epsilon} L_u(\epsilon, \phi) = -\frac{1}{2} \frac{(\omega + S_u\lambda_u)^2}{1 + \lambda_u^2} + O(n^{-\frac{1}{2}}) \quad . \quad (21)$$

This gives the MLE

$$\hat{\phi}_u = \hat{\phi} - n^{-\frac{1}{2}}\sigma_\phi\lambda_u S_u + O(n^{-1}) \quad . \quad (22)$$

The MLE is equivalent to estimating  $\epsilon$  by the consistent estimate  $n^{-\frac{1}{2}}S_u$  and substituting this for  $\epsilon$  in the formula for the bias in (16).

Note that if  $u = u_a$  in (17), which maximizes the directional component  $\lambda_u$  of the bias, then (22) is the same as the non-parametric estimate  $\tilde{\phi}$  in (13). This is because, from (11),

$$0 = n^{\frac{1}{2}}\bar{E}a(x, \tilde{\phi}) = n^{\frac{1}{2}}\bar{E}a(x, \hat{\phi}) + n^{\frac{1}{2}}(\tilde{\phi} - \hat{\phi}) + O(n^{-\frac{1}{2}}) \quad ,$$

and from (17),

$$n^{\frac{1}{2}}\bar{E}a(x, \hat{\phi}) = S_{u_a}\sigma_\phi\lambda_{u_a} \quad .$$

Further, from (18),  $1/(1 + \lambda_{u_a}^2) = \rho^2$ , so  $L_{u_a}(\phi) = L_{NP}(\phi)$ .

Now consider the family of likelihoods (21) for different direction functions  $u$ , say for  $u \in \mathcal{U}$ . For each  $u$ ,  $L_u$  is characterized by the upper and lower values of  $\omega$ , say  $\omega_{u+}(z)$  and  $\omega_{u-}(z)$  defined by

$$L_u\{\omega_{u+}(z)\} = L_u\{\omega_{u-}(z)\} = -\frac{1}{2}z^2 \quad .$$

Let the *envelope likelihood*  $L_{ENV}$  be similarly defined by the two limits

$$\omega_+(z) = \sup_{u \in \mathcal{U}} \omega_{u+}(z) \quad , \quad \omega_-(z) = \inf_{u \in \mathcal{U}} \omega_{u-}(z) \quad . \quad (23)$$

Equivalently,

$$L_{ENV}(\phi) = \sup_{u \in \mathcal{U}} L_u(\phi) \quad . \quad (24)$$

We look to the “worst case” for assessing the likelihood of each individual value of  $\phi$ , in the sense that for all  $\phi$  and all  $u \in \mathcal{U}$ ,

$$L_u(\phi) \leq L_{ENV}(\phi) \quad . \quad (25)$$

If  $z = z_\alpha$  then  $\{\omega_{u,-}(z_\alpha), \omega_{u+}(z_\alpha)\}$  are the upper and lower asymptotic confidence limits for  $\omega$  with coverage  $1 - \alpha$  under the model  $g_u$ . If  $\mathcal{U}$  is a *fixed* set of directions, the expanded interval  $\{\omega_-(z_\alpha), \omega_+(z_\alpha)\}$  is still a confidence interval in the wider sense (Shao, 2003, p.142) that (asymptotically) for all  $u \in \mathcal{U}$

$$P_u\{\omega_-(z_\alpha) \leq \omega \leq \omega_+(z_\alpha)\} \geq 1 - \alpha \quad . \quad (26)$$

In Section 2.1 we defined the set of statistically equivalent models  $\mathcal{G}$  in (7), those alternative models  $g_u$  within which  $f$  would be accepted as a reasonable explanation of the data (at significance level  $\alpha$ ). This is equivalent to taking  $\mathcal{U}$  to be the random set

$$\widehat{\mathcal{U}} = \{u : |S_u| \leq z_\alpha\} \quad . \quad (27)$$

Then for  $u \in \widehat{\mathcal{U}}$ ,

$$\omega_{u-}(z) = -S_u \lambda_u - z(1 + \lambda_u^2)^{\frac{1}{2}} \geq -z_\alpha \lambda_u - z(1 + \lambda_u^2)^{\frac{1}{2}} \quad (28)$$

$$\omega_{u+}(z) = -S_u \lambda_u + z(1 + \lambda_u^2)^{\frac{1}{2}} \leq z_\alpha \lambda_u + z(1 + \lambda_u^2)^{\frac{1}{2}} \quad . \quad (29)$$

Thus using the bounds for  $\lambda_u$  in (18),

$$\omega_-(z) = -z_\alpha \lambda_M - \frac{z}{\rho}, \quad \omega_+(z) = z_\alpha \lambda_M + \frac{z}{\rho} \quad . \quad (30)$$

(In these expressions we take  $\lambda_M$  as the positive square root of the right hand side of (18)). Equivalently,

$$L_{ENV}(\phi) = -\frac{1}{2}\rho^2 [\max\{|w| - z_\alpha \lambda_M, 0\}]^2 \quad . \quad (31)$$

We use the notation  $\widehat{\mathcal{U}}$  to emphasize that this set of directions depends on the data, and so although  $L_{ENV}$  still has its “worst case” interpretation (25) in terms of likelihood, the

interval  $\{\omega_-(z_\alpha), \omega_+(z_\alpha)\}$  no longer has the straightforward frequentist interpretation (26) of a confidence interval (but see Section 5, comment 4).

The envelope likelihood (31) assumes that the bounds for  $S_u$  and  $\lambda_u$  can be attained independently. If  $x$  is continuous this is the case since  $\lambda_u$  depends on all values of  $u(x, \hat{\theta})$  whereas  $S_u$  depends on  $u(x, \hat{\theta})$  only at a finite number of data points. In other cases, such as when  $x$  is discrete or when smoothness restrictions are placed on  $u$ ,  $\lambda_u$  and  $S_u$  cannot vary independently and the limits in (30) are merely bounds which can be tightened (see Section 4 for a discussion of the discrete case).

If  $\rho = \pm 1$  then  $\lambda_M = 0$  and so  $L_{ENV}$  in (31) is just  $-\frac{1}{2}\omega^2$ , the usual model likelihood  $L_f$ . Inference is then robust to first order. As  $|\rho|$  decreases,  $L_{ENV}$  moves further away from  $L_f$  and becomes progressively flatter. In the limit as  $\rho \rightarrow 0$ ,  $L_{ENV}$  becomes zero for all  $\phi$  and so is completely uninformative. Even if  $f$  gives a good fit to the data, it is a useless model if our aim is to estimate  $\phi$ .

## 2.4 Example 1 : estimating the mean of a log-normal distribution

Consider the problem of estimating the mean of a highly skewed distribution, such as the distribution of weekly wages from an income survey. Income data usually have a very long right tail, and a small number of outlying observations can have a strong influence on the sample mean. A more stable approach is to fit a parametric model and deduce the mean as a function of the fitted parameters. For income data the log-normal distribution is the natural choice.

Thus let  $x = \log(\text{wages})$  and  $f = N(\mu, \sigma^2)$ , so that  $\theta = (\mu, \sigma^2)$ . The MLE under  $f$  is  $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ , the sample mean and variance respectively of the observations  $x_1, x_2, \dots, x_n$ . It is sensible to estimate the mean wage on the log scale, so  $\phi = \log E\{\exp(x)\}$  for which

$$a(x, \phi) = 1 - \exp(x - \phi) \quad .$$

For this example we have

$$\hat{\phi} = \hat{\mu} + \frac{1}{2}\hat{\sigma}^2 \quad , \quad \tilde{\phi} = \log \bar{E}\{\exp(x)\} \quad , \quad \sigma_\phi^2 = \sigma^2 + \frac{1}{2}\sigma^4 \quad , \quad \sigma_a^2 = \exp(\sigma^2) - 1 \quad .$$

Hence

$$\rho^2 = \frac{\sigma^2 + \frac{1}{2}\sigma^4}{\exp(\sigma^2) - 1} \leq 1 \quad .$$

Note that  $\rho$  depends only on  $\sigma^2$ . As expected,  $\rho \rightarrow 1$  if  $\sigma^2 \rightarrow 0$ , as the distribution of  $x$  becomes more symmetrical, and  $\rho \rightarrow 0$  if  $\sigma^2 \rightarrow \infty$ , as the distribution becomes more highly skewed.

Fig. 1 shows likelihoods for an example with  $n = 100$ . This sample is simply illustrative of income data, simulated to reflect some key percentiles of the actual income distribution in the UK in 2002. According to the Shapiro-Wilk test for normality these data give an acceptable fit ( $W = 0.978$ , P-value = 0.092). The likelihoods are plotted against  $\exp(\phi)$ , so the horizontal axis is mean income in pounds per week. On the graph,  $L_f$  is the solid line,  $L_{ENV}$  the stippled line and  $L_{NP}$  the dotted line. The horizontal line at  $L = -2$  indicates the corresponding asymptotic 95% confidence intervals for average wage. The interval for

$L_{ENV}$  is considerably wider than that for  $L_f$ , but does not extend to such high values as the nonparametric interval for  $L_{NP}$ .

Although for these data the omnibus test for  $f$  indicates an acceptable fit,  $|S_{u_a}| = 3.29 > 1.96$ , so there is a significant lack of fit in the *specific* direction  $u = u_a$ . This explains why  $L_{ENV}$  does not cover  $L_{NP}$ . In these data there is one rather large and influential observation which does not upset the omnibus test but largely accounts for the significantly large value of  $S_{u_a}$ . Removing this observation has a much more marked effect on  $L_{NP}$  than on  $L_{ENV}$ .

## 2.5 Geometrical insights

We follow Copas and Eguchi (2005) in giving an intuitive outline of the geometrical ideas behind much of the material we have been discussing. In his paper on the related topic of testing non-nested models, Kent (1986) similarly added a section giving a geometrical overview. Our set-up is also quite similar to that of Small and Wang (2003, section 2.5). See Amari (1985) for an authoritative account of the structure of the underlying mathematics.

We focus on the three principal functions in our discussion,  $a = a(x, \phi(\theta))$ ,  $u = u(x, \theta)$  and  $s = (s_1, s_2, \dots, s_p)$ , the components of the score function  $s(x, \theta)$ . All these functions belong to the linear space

$$\mathcal{L}_\theta = \{v : E_\theta v = 0, E_\theta v^2 < \infty\} .$$

We define inner product and norm within this space to be

$$\langle v_1, v_2 \rangle_\theta = E_\theta(v_1 v_2) , \quad \|v\|_\theta = \langle v, v \rangle_\theta^{\frac{1}{2}} = (E_\theta v^2)^{\frac{1}{2}} .$$

An important subspace of  $\mathcal{L}_\theta$  is the set of linear combinations of the score vector,

$$\mathcal{L}_s = \{\alpha^T s : \alpha \in \mathcal{R}^p\} .$$

Geometrically,  $\mathcal{L}_s$  is the tangent plane to the model  $f(x, \theta)$ : from (5) we see that an increment in the direction  $\alpha^T s$  induces an increment in  $\theta$ . Similarly, we can view  $\mathcal{L}_\theta$  as the tangent space of a non-parametric family of distributions at  $f = f(x, \theta)$ .

A key geometrical concept is *projection*, denoted here by the symbol  $\Pi$  (using the notation of Henmi and Eguchi, 2004). If  $w \in \mathcal{L}_\theta$  and  $v$  is a vector  $(v_1, v_2, \dots, v_p)$  of functions in  $\mathcal{L}_\theta$ , then the projection of  $w$  onto linear combinations of  $v$  (the linear hull of  $v$ ) is  $\Pi(w|v)$  defined by

$$\|w - \Pi(w|v)\|_\theta^2 = \min_{\alpha \in \mathcal{R}^p} \|w - \alpha^T v\|_\theta^2 .$$

This is like least squares, regressing  $w$  onto the elements of  $v$ , and so

$$\Pi(w|v) = E_\theta(w v^T) \{E_\theta(v v^T)\}^{-1} v .$$

We can give similar statistical meanings to the projection properties

$$\begin{aligned} \Pi\{w|\Pi(w|v)\} &= \Pi(w|v) \\ \langle w, \Pi(w|v) \rangle_\theta &= \|\Pi(w|v)\|_\theta^2 \\ \|w\|_\theta^2 &= \|w - \Pi(w|v)\|_\theta^2 + \|\Pi(w|v)\|_\theta^2 . \end{aligned} \tag{32}$$

Most of the results discussed in Section 2.2 are simply properties of various projections defined within  $\mathcal{L}_\theta$ . These are illustrated in Figure 2. For example, the asymptotic variance of the MLE  $\hat{\phi}$  in (12) is  $n^{-1}\sigma_\phi^2$  with

$$\sigma_\phi^2 = \|\Pi(a|s)\|_\theta^2 .$$

Similarly, the quantities  $\rho^2$  in (14) and  $\lambda_u^2$  in (15) are

$$\rho^2 = \frac{\|\Pi(a|s)\|_\theta^2}{\|a\|_\theta^2} , \quad \lambda_u^2 = \frac{\|\Pi(a|u)\|_\theta^2}{\|\Pi(a|s)\|_\theta^2} .$$

Now the orthogonality of  $u$  and  $s$  implies that

$$\Pi(a - \Pi(a|s)|u) = \Pi(a|u) , \quad (33)$$

and so, using the triangle equality (32), we get

$$\begin{aligned} & \|a\|_\theta^2 - \|\Pi(a|s)\|_\theta^2 - \|\Pi(u|a)\|_\theta^2 \\ &= \|a - \Pi(a|s)\|_\theta^2 - \|\Pi(a - \Pi(a|s)|u)\|_\theta^2 \\ &= \|a - \Pi(a|s) - \Pi(a - \Pi(a|s)|u)\|_\theta^2 \\ &\geq 0 . \end{aligned} \quad (34)$$

Hence

$$\lambda_u^2 = \frac{\|\Pi(a|u)\|_\theta^2}{\|\Pi(a|s)\|_\theta^2} \leq \frac{\|a\|_\theta^2 - \|\Pi(a|s)\|_\theta^2}{\|\Pi(a|s)\|_\theta^2} = \frac{1 - \rho^2}{\rho^2} ,$$

giving a geometrical proof of (18). From (33) and (34), this inequality is attained when

$$\Pi(a|u) = a - \Pi(a|s) .$$

But  $\Pi(a|u)$  is just a constant multiple of  $u$ , so  $u$  must be

$$u_a = \frac{a - \Pi(a|s)}{\|a - \Pi(a|s)\|_\theta} ,$$

as in (17).

The notation also extends to the discussion in Section 2.3. For example, equation (22) for the difference between the MLEs of  $\phi$  under models  $f$  and  $g_u$  is

$$\hat{\phi} - \hat{\phi}_u = n^{-\frac{1}{2}} \bar{\mathbf{E}}\{\Pi(a|u)\} .$$

## 3 Extension to the basic theory

### 3.1 Extended definition of parameter of interest

In Section 2.2 we defined  $\phi$ , the parameter of interest, in terms of the scalar estimating function  $a(x, \phi)$  in (8). This covers simple parameters such as moments or percentiles,

but does not, for example, include non-linear functions of moments such as variances or correlation coefficients. In many problems we need a more general definition.

Suppose now that scalar  $\phi$  is a function of a vector  $\eta$  of intermediate parameters, say  $\phi = h(\eta)$  for some given function  $h(\eta)$ . Each component of  $\eta$  is defined analogously to (8), so that for any distribution  $g$ , vector  $\eta$  is given by

$$E_g a(x, \eta) = 0 \quad ,$$

where  $a(x, \eta)$  is now a vector function of the same dimension as  $\eta$ . Under  $f$ ,  $\eta = \eta(\theta)$ , defined in the same way as  $\phi(\theta)$  in (10). The scaling constraint (11) now assumes that  $E_\theta \partial a\{x, \eta(\theta)\} / \partial \eta$  is the identity matrix. This means that, as in (16),

$$\eta = \eta(\theta) - \epsilon E_\theta [u(x, \theta) a\{x, \eta(\theta)\}] + O(\epsilon^2) \quad .$$

Now define

$$\phi(\theta) = h\{\eta(\theta)\} \quad , \quad \hat{\phi} = \phi(\hat{\theta}) = h\{\eta(\hat{\theta})\} \quad .$$

Then

$$\phi = h(\eta) = \phi(\theta) - \epsilon E_\theta [u(x, \theta) a^*\{x, \eta(\theta)\}] + O(\epsilon^2) \quad ,$$

where

$$a^*\{x, \eta(\theta)\} = \left\{ \frac{\partial h\{\eta(\theta)\}}{\partial \eta} \right\}^T a\{x, \eta(\theta)\} \quad . \quad (35)$$

This means that as far as first order bias is concerned the theory is exactly the same as before, simply replacing the previous scalar estimating function  $a(x, \phi)$  by the scalar composite estimating function  $a^*\{x, \eta(\theta)\}$  in (35). The variance  $\sigma_a^2$  becomes

$$\left\{ \frac{\partial h\{\eta(\theta)\}}{\partial \eta} \right\}^T \text{Var}_\theta a\{x, \eta(\theta)\} \left\{ \frac{\partial h\{\eta(\theta)\}}{\partial \eta} \right\} \quad ,$$

and the previous quantities  $\sigma_\phi^2$ ,  $\rho$  and  $\lambda_u$  then follow from (12), (14) and (15). For first order asymptotic inference, all we need are the values of these quantities evaluated at the model MLE  $\theta = \hat{\theta}$ .

For a simple example let  $\phi = \text{Var}_g(x) = \eta_2 - \eta_1^2$ , where  $\eta = (\eta_1, \eta_2)$  are the first two moments of  $x$  under  $g$ . Then the two components of  $a(x, \eta)$  are  $(\eta_1 - x)$  and  $(\eta_2 - x^2)$ . For this,  $E_\theta \partial a / \partial \eta$  is the identity matrix as required and

$$a^* = -2\eta_1(\eta_1 - x) + \eta_2 - x^2 = \phi - (x - \eta_1)^2 \quad .$$

Note that the choice of  $a$  and  $\eta$  is not unique. For example we could equally well take the two components of  $a(x, \eta)$  to be  $(\eta_1 - x)$  and  $\{\eta_2 - (x - \eta_1)^2\}$  so that now  $\phi = \eta_2$ . However this gives the exactly the same function  $a^*$ . As noted before, it is not possible to define the variance directly in terms of a single estimating function  $a(x, \phi)$ .

## 3.2 Grouped data

An important special case of Section 3.1 is when the data take the form of samples in different groups or strata. Each of the  $J$  strata corresponds to a component of the vector  $\eta$ , and  $\phi$  is a global parameter defined across all of these strata. Suppose that within the  $j$ th strata the observed  $x$ 's have model distribution  $f_j(x, \theta_j)$  with model parameter  $\theta_j$ , score function  $s_j(x, \theta_j)$  and information matrix  $I_j$ ,  $j = 1, 2, \dots, J$ . Each strata has its own parameter of interest  $\eta_j$  defined by the estimating function  $a_j(x, \eta_j)$ . Following the notation and assumptions of Section 2 for each strata separately, we have

$$\mathbb{E}_{\theta_j} \frac{\partial a_j}{\partial \eta_j} = 1 \quad , \quad \sigma_{a_j}^2 = \text{Var}_{\theta_j}(a_j) \quad , \quad \rho_j^2 = \frac{\{\mathbb{E}_{\theta_j}(a_j s_j)\}^T I_j^{-1} \mathbb{E}_{\theta_j}(a_j s_j)}{\sigma_{a_j}^2} \quad .$$

Within the  $j$ th strata we observe the random sample  $x_{ij}$ ,  $i = 1, 2, \dots, n_j$ . To bring this within the basic set-up we think of these data as a single random sample of size  $n = \sum_j n_j$  of the joint random variable  $(x, j)$ .

Under the model the joint distribution of  $(x, j)$  is given by

$$P(j) = p_j \quad , \quad x|j \sim f_j(x, \theta_j) \quad . \quad (36)$$

The model MLEs are  $\hat{\theta}_j$ , the ordinary MLEs from each strata separately, and  $\hat{p}_j = n_j/n$ ,  $j = 1, 2, \dots, J$ . In terms of the joint observation  $(x, j)$  the estimating function defining  $\eta_k$  is

$$\frac{1}{p_k} \begin{cases} 0 & \text{if } j \neq k \\ a_k(x, \eta_k) & \text{if } j = k \end{cases} \quad ,$$

and so if we are interested in the overall scalar parameter  $\phi = h(\eta_1, \eta_2, \dots, \eta_J)$  then the composite estimating function is

$$a^*\{(x, j), \eta\} = \frac{1}{p_j} \frac{\partial h}{\partial \eta_j} a_j(x, \eta_j) \quad .$$

If  $\hat{\phi}$  is the model MLE of  $\phi$  and  $\rho$  is the corresponding correlation parameter for the experiment as a whole, we get

$$n \text{Var}_{\theta}(\hat{\phi}) = \sum_j w_j \rho_j^2 \quad , \quad \rho^2 = \frac{\sum w_j \rho_j^2}{\sum w_j} \quad ,$$

where the weights  $w_j$  are given by

$$w_j = \frac{\sigma_{a_j}^2}{p_j} \left( \frac{\partial h}{\partial \eta_j} \right)^2 \quad .$$

## 3.3 Example 2 : exponential survival with censoring

Suppose that we model failure time  $y$  as an exponential random variable with rate parameter  $\theta$ , so that

$$f(y, \theta) = \theta e^{-\theta y} \quad y \geq 0 \quad .$$

Let  $t_1, t_2, \dots, t_K$  be potential censoring times with sequential censoring probabilities  $p_1, p_2, \dots, p_K$ , so that the model distribution of (potential) censoring time  $t$  is

$$P(t = t_i) = \nu_{i-1} p_i \quad , \quad \nu_0 = 1 \quad , \quad \nu_i = \prod_{j=1}^i (1 - p_j) \quad ; \quad i = 1, 2, \dots, K \quad .$$

The censoring means that we can only observe

$$x = \min\{y, t\} \quad .$$

Assuming uninformative censoring the model distribution  $P_f(x)$  of  $x$  is the mixed discrete and continuous distribution given by probability mass function/density function

$$\begin{aligned} e^{-\theta t_i} \nu_{i-1} p_i & \quad \text{if } x = t_i \\ \nu_{i-1} \theta e^{-\theta x} & \quad \text{if } x \in (t_{i-1}, t_i) \quad , \end{aligned}$$

for  $i = 1, 2, \dots, K+1$  with  $t_0 = p_{K+1} = 0$  and  $t_{K+1} = \infty$ . It follows that  $-\partial^2 \log P_f(x) / \partial \theta^2$  is  $\theta^{-2}$  for uncensored observations and zero for censored observations. Thus the model information is

$$I = -E_\theta \left( \frac{\partial^2}{\partial \theta^2} \log P_f(x) \right) = \frac{1}{\theta^2} \left( 1 - \sum_{i=1}^K e^{-\theta t_i} \nu_{i-1} p_i \right) \quad .$$

Suppose that for any survival distribution  $g$  we are interested in the log survival probability  $\phi = \log P_g(y > \tau)$  for some fixed time  $\tau$ . Fixing  $\tau$  fixes  $k$  such that  $t_{k-1} \leq \tau < t_k$ . Then we can write  $\phi = \sum_{i=1}^k \log \eta_i$  where the subsidiary parameters  $\eta_i$  are

$$\eta_i = P_g(y > t_i | y > t_{i-1}) \quad (i = 1, 2, \dots, k-1) \quad , \quad \eta_k = P_g(y > \tau | y > t_{k-1}) \quad .$$

These are defined by the estimating functions, for  $i = 1, 2, \dots, k-1$ ,

$$a_i(x, \eta_i) = \frac{1}{\nu_{i-1} \exp(-\theta t_{i-1})} \begin{cases} 0 & \text{if } x \leq t_{i-1} \\ \eta_i & \text{if } t_{i-1} < x < t_i \\ \eta_i - 1 & \text{if } x \geq t_i \end{cases} \quad . \quad (37)$$

Estimating function  $a_k(x, \eta_k)$  is the value of (37) for  $i = k$  but with  $t_k$  replaced by  $\tau$ . (The factor before the bracket in (37) is needed to ensure that the  $a_i$ 's satisfy the required scaling constraints). Under the model the  $a_i$ 's are uncorrelated with

$$\text{Var}_\theta(a_i) = \frac{\eta_i(1 - \eta_i)}{\nu_{i-1} e^{-\theta t_{i-1}}} \quad .$$

Evaluating the model variance of the composite estimating function  $a^*$  in (35), we find

$$\text{Var}_\theta(a^*) = \sum_{i=1}^k \frac{1}{\eta_i^2} \text{Var}_\theta(a_i) = \frac{e^{\theta \tau}}{\nu_{k-1}} - \sum_{i=1}^{k-1} \frac{e^{\theta t_i} p_i}{\nu_i} - 1 \quad .$$

Now suppose we have a large sample of size  $n$ , and let  $\tilde{\eta}_i$  be the nonparametric estimates of  $\eta_i$  given by  $\bar{E} a_i(x, \tilde{\eta}_i) = 0$ . This gives  $\tilde{\phi} = \sum_1^k \log \tilde{\eta}_i$ , the usual Kaplan-Meier estimate of log survival at  $\tau$ . If  $\hat{\theta}$  is the model MLE of  $\theta$  then the model MLE of  $\phi$  is  $\hat{\phi} = -\hat{\theta} \tau$ . Thus

$$\rho^2 = \frac{\text{Var}_\theta(\hat{\phi})}{\text{Var}_\theta(\tilde{\phi})} = \frac{\tau^2}{\text{Var}_\theta(a^*) I} = \frac{\theta^2 \tau^2}{(1 - \sum_{i=1}^K e^{-\theta t_i} \nu_{i-1} p_i) \left( \frac{e^{\theta \tau}}{\nu_{k-1}} - \sum_{i=1}^{k-1} \frac{e^{\theta t_i} p_i}{\nu_i} - 1 \right)} \quad .$$

To estimate  $\rho$  we use the MLE  $\hat{\theta}$  given by

$$\hat{\theta} = \frac{n - \sum_{i=1}^K \text{Frequency}(x = t_i)}{\sum_{j=1}^n x_j}$$

and the MLE of  $p_i$  given by

$$\hat{p}_i = \frac{\text{Frequency}(x = t_i)}{\text{Frequency}(x \leq t_i)} .$$

An alternative version is to take the parameter of interest to be the complementary log-log survival function (log cumulative hazard) rather than the cumulative hazard itself. This means we now define  $\phi = \log(-\sum_1^k \log \eta_i)$ . The only change to the above development is that the  $\sigma_\phi^2$  now becomes  $1/(I\theta^2)$ . The value of  $\rho$  remains exactly the same (as both  $\hat{\phi}$  and  $\tilde{\phi}$  transform in the same way).

As an illustration of these calculations, Armitage *et al.* (1969) discuss the results of a clinical trial comparing two surgical treatments (A and B) for advanced breast cancer. Amongst other measures, we have the (censored) survival times in months from treatment to death for a total of 187 patients. On treatment A there were  $n = 101$  patients, including 12 censored observations giving the values of  $t_i$  with  $K = 12$  (although other potential censoring times are clearly possible it is only the observed times that matter in the sense of having  $\hat{p}_i > 0$ ). We take a grid of values for  $\tau$  and use the complementary log-log version of the above model.

Figure 3 shows asymptotic confidence intervals for the survival function  $S(\tau) = P(y > \tau)$  for treatment A, plotted against  $\tau$ . The solid lines show the fitted exponential survival function and its point-wise 95% confidence limits. The dotted lines show the Kaplan-Meier survival curve with confidence limits. The dashed lines are the worst-case limits from (31) allowing for all possible mis-specification functions  $u$  for which  $|S_u| < 1.96$ . The envelope likelihood limits cover the Kaplan-Meier limits for almost all values of  $\tau$ , confirming the excellent fit of the exponential model. The corresponding survival curves for treatment B tend to zero at a slower rate than for treatment A, but otherwise are rather similar to Figure 3. Under treatment B there were 71 observed survival times and 15 censored cases.

As an illustration of Section 3.2, the two treatments in this trial can be modeled together using (36), defining  $j = 1, 2$  as treatments A, B respectively. Thus  $J = 2$  with  $n_1 = 101$  and  $n_2 = 86$ . The relative effectiveness of the treatments in terms of the probability of surviving to time  $\tau$  can be measured by

$$\phi = \log\{P(y > \tau|A)/P(y > \tau|B)\} .$$

Figure 4 shows estimates and point-wise confidence limits for  $\phi$  plotted against  $\tau$ . If  $\hat{\theta}_A$  and  $\hat{\theta}_B$  are the estimated failure rate parameters for the two treatments, the usual analysis gives

$$\frac{\hat{\theta}_A - \hat{\theta}_B}{\sqrt{\text{Var}(\hat{\theta}_A - \hat{\theta}_B)}} = 2.47 , \quad (38)$$

suggesting reasonably strong evidence that B is better than A. However when model uncertainty is taken into account, the result is much less clear. Although the solid lines in Figure 4 are all below zero, the upper parts of both the Kaplan-Meier and worst case

confidence intervals are positive for all  $\tau$  (meaning that  $B$  may actually be worse than  $A$ ). Evidently the significant difference suggested by (38) depends strongly on the parametric assumption of exponential survival, although there is no reason to doubt that this is a perfectly reasonable model for these particular data.

## 4 Tighter likelihood bounds for finite discrete models

We now consider the case when the sample space of  $x$  is finite, for example a finite discrete distribution or a contingency table. The set-up follows directly as a special case of Sections 2.1 and 2.2, but tighter bounds for the envelope likelihood  $L_{ENV}$  in (31) are possible, as we now discuss. A contingency table example is used as an illustration.

### 4.1 Likelihoods for the finite case

Suppose that  $x$  is a discrete random variable taking one of  $m$  distinct values  $d_j$ ,  $j = 1, 2, \dots, m$ . The working model  $f$  is then the discrete distribution

$$f(d_j, \theta) = P_\theta(x = d_j) , \quad j = 1, 2, \dots, m .$$

The functions  $s(x, \theta)$ ,  $u(x, \theta)$  and  $a(x, \phi)$  are defined as in the general case, as are the model MLEs  $\hat{\theta}$  and  $\hat{\phi}$  and the non-parametric estimate  $\tilde{\phi}$ . Let

$$p_j = f(d_j, \hat{\theta}), \quad a_j = a(d_j, \hat{\phi}), \quad u_j = u(d_j, \hat{\theta}), \quad s_j = s(d_j, \hat{\theta}) . \quad (39)$$

These quantities satisfy

$$\sum p_j u_j = \sum p_j s_j u_j = \sum p_j a_j = \sum p_j s_j = 0 , \quad \sum p_j u_j^2 = 1 . \quad (40)$$

Given data  $x_1, x_2, \dots, x_n$  let  $q_j$  be the observed relative frequency of  $x = d_j$ , and  $\chi^2$  be the usual goodness-of-fit statistic. Then

$$\sum q_j s_j = 0 , \quad \chi^2 = n \sum_{j=0}^m \frac{(q_j - p_j)^2}{p_j} .$$

Other quantities defined earlier become

$$S_u = n^{\frac{1}{2}} \sum q_j u_j , \quad \sigma_a^2 = \sum p_j a_j^2 , \quad \lambda_u = \sigma_\phi^{-1} \sum p_j a_j u_j , \quad E(as) = \sum p_j a_j s_j ,$$

where  $\sigma_\phi$  (and hence  $\rho$ ) are defined as in (14).

Let

$$b = \frac{n^{\frac{1}{2}} \sum q_j a_j}{\chi \sigma_a (1 - \rho^2)^{\frac{1}{2}}} . \quad (41)$$

By using Lagrange multipliers to find the max/min over  $a_j$  of  $\sum q_j a_j$  for fixed  $E(as)$  and  $\sigma_a^2$ , and given  $\sum p_j a_j = 0$ , it is easy to show that  $-1 \leq b \leq +1$ . A further exercise

with Lagrange multipliers, now finding the max/min of  $\sum q_j u_j = n^{-\frac{1}{2}} S_u$  over  $u_j$  with the constraints in (40) and for given  $\sum p_j a_j u_j = \sigma_\phi \lambda_u$ , shows that

$$\frac{\chi}{\lambda_M} [b\lambda_u - \{(1-b^2)(\lambda_M^2 - \lambda_u^2)\}^{\frac{1}{2}}] \leq S_u \leq \frac{\chi}{\lambda_M} [b\lambda_u + \{(1-b^2)(\lambda_M^2 - \lambda_u^2)\}^{\frac{1}{2}}] .$$

With  $|S_u| < z_\alpha$  as before, the bounds in (28) and (29) can therefore be tightened to

$$\begin{aligned} \omega_{u-}(z) &= -S_u \lambda_u - z(1 + \lambda_u^2)^{\frac{1}{2}} \\ &\geq \max\{-|\lambda_u| z_\alpha, -\frac{\chi}{\lambda_M} [b\lambda_u^2 + |\lambda_u| \{(1-b^2)(\lambda_M^2 - \lambda_u^2)\}^{\frac{1}{2}}]\} - z(1 + \lambda_u^2)^{\frac{1}{2}} \end{aligned} \quad (42)$$

$$\begin{aligned} \omega_{u+}(z) &= -S_u \lambda_u + k(1 + \lambda_u^2)^{\frac{1}{2}} \\ &\leq \min\{+|\lambda_u| z_\alpha, -\frac{\chi}{\lambda_M} [b\lambda_u^2 - |\lambda_u| \{(1-b^2)(\lambda_M^2 - \lambda_u^2)\}^{\frac{1}{2}}]\} + z(1 + \lambda_u^2)^{\frac{1}{2}} \end{aligned} \quad (43)$$

These bounds are exact in the sense that there exist mis-specification directions  $u$  for which (42) and (43) are attained, and are generally closer together than the bounds given in (28) and (29). The envelope likelihood defined in (24) is then given by

$$\begin{aligned} \omega_-(z) &= \\ &\min_{0 \leq t \leq 1} \left\{ \max[-\lambda_M z_\alpha t^{\frac{1}{2}}, -\lambda_M b\chi t - \lambda_M \chi \{t(1-t)(1-b^2)\}^{\frac{1}{2}}] - z(1 + \lambda_M^2 t)^{\frac{1}{2}} \right\}, \end{aligned} \quad (44)$$

$$\begin{aligned} \omega_+(z) &= \\ &\max_{0 \leq t \leq 1} \left\{ \min[\lambda_M z_\alpha t^{\frac{1}{2}}, -\lambda_M b\chi t + \lambda_M \chi \{t(1-t)(1-b^2)\}^{\frac{1}{2}}] + z(1 + \lambda_M^2 t)^{\frac{1}{2}} \right\} . \end{aligned} \quad (45)$$

Here,  $t = \lambda_u^2 / \lambda_M^2$  so the limits on  $t$  follow from (18). For given  $z$ , (44) and (45) are easy to calculate numerically by finding the max/min over a suitably fine grid of values of  $t$  in  $[0, 1]$ . The plot of  $L_{ENV}(\phi)$  then follows by taking a grid of values of  $z \geq 0$  and plotting  $-z^2/2$  against  $\hat{\phi} + n^{-\frac{1}{2}} \sigma_\phi \omega_\pm(z)$ . This plots the two wings of the envelope separately, filled in by taking  $L_{ENV}(\phi) = 0$  for all values of  $\phi$  for which

$$\omega_-(0) < \omega < \omega_+(0) . \quad (46)$$

The interval (46) is usually non-empty, meaning that  $L_{ENV}(\phi) = 0$  for a positive range of values of  $\phi$ . Exceptions are when the data fit the model exactly ( $\chi^2 = 0$ ) or when  $\lambda_M = 0$  ( $\rho = \pm 1$ ), in which cases  $L_{ENV}(\phi) = L_{NP}(\phi) \geq L_f(\phi)$ . We now have

$$L_{NP}(\omega) = -\frac{\rho^2}{2} (\omega + \lambda_M b\chi)^2.$$

## 4.2 Example 3: Estimating an average causal effect

Consider a study with a total of  $n$  subjects, each of which is assigned to one of two treatments  $t$ , gives a binary response  $y$ , and is characterized by a vector  $v$  of categorical covariates. The data can be summarized by a contingency table with cell frequencies  $n_{ytv}$ . Our aim is to estimate the treatment effect adjusting for the covariates  $v$ .

Coding the two binary variables  $y$  and  $t$  as 0/1, the conventional model here is the logistic regression

$$P(y = 1|t, v) = p_{tv} = \frac{\exp(\beta_0 + \beta_1 t + \beta_2^T v)}{1 + \exp(\beta_0 + \beta_1 t + \beta_2^T v)} . \quad (47)$$

This specifies the conditional distribution of  $y$  given  $t$  and  $v$ . If  $(t, v)$  are thought of as random as well, the data take the form of a random sample from the discrete distribution over all possible values of  $x = (y, t, v)$ , and so is a special case of the above theory. As the model makes no assumptions about the marginal probabilities of  $(t, v)$ , these will simply end up as estimated from the observed marginal frequencies. Equivalently, as far as first order asymptotic approximations are concerned, we can assume that the marginal totals

$$n_{tv} = n_{0tv} + n_{1tv} \quad , \quad n_v = n_{0v} + n_{1v} \quad , \quad n = \sum_v n_v$$

are fixed. The only explicit parameters of the model are therefore  $\theta = (\beta_0, \beta_1, \beta_2)$ .

Under this model,  $\beta_1$  is the log-odds ratio for the treatment effect conditional on  $v$ , assumed to be constant over different values of  $v$ . It is also assumed that the dependence of  $y$  on  $t$  and  $v$  is captured by the additive form of the logit in (47). However, the meaning of  $\beta_1$  is unclear if this model is misspecified. An alternative and more primitive measure of treatment effect is

$$\phi = E_v(p_{1v} - p_{0v}) \quad ,$$

where  $E_v$  denotes expectation over the discrete distribution of  $v$  with probabilities  $n_v/n$ . This has a simple interpretation as the *average causal effect*, the difference between the overall proportion of positive responses had all subjects in the study been treated with  $t = 1$  and the overall proportion had all been treated with  $t = 0$ . Equivalently,  $\phi$  is defined by the estimating function  $a(x, \phi)$  given by

$$a\{(y, t, v); \phi\} = \phi - \frac{y(2t - 1)n_v}{n_{tv}} .$$

If  $\phi(\theta)$  is the value of  $\phi$  at the model, then from (47) we have

$$\frac{\partial \phi(\theta)}{\partial \theta} = \{E_v(\gamma_{1v}), E_v(\gamma_{1v} - \gamma_{0v}), E_v v^T (\gamma_{1v} - \gamma_{0v})\}^T ,$$

where  $\gamma_{tv} = p_{tv}(1 - p_{tv})$ . This gives

$$\sigma_\phi^2 = n \text{Var}_\theta(\hat{\phi}) = n \left( \frac{\partial \phi(\theta)}{\partial \theta} \right)^T V \left( \frac{\partial \phi(\theta)}{\partial \theta} \right)$$

where  $V$  is the usual asymptotic covariance matrix of the logistic regression estimates  $\hat{\theta}$ .

The non-parametric estimate of  $\phi$  is

$$\tilde{\phi} = E_v(\tilde{p}_{1v} - \tilde{p}_{0v}) \quad ,$$

where

$$\tilde{p}_{tv} = \frac{n_{1tv}}{n_{tv}} .$$

Hence

$$\sigma_a^2 = n \text{Var}_\theta(\tilde{\phi}) = E_v \left\{ n_v \left( \frac{\gamma_{1v}}{n_{1v}} + \frac{\gamma_{0v}}{n_{0v}} \right) \right\} . \quad (48)$$

Other quantities required for the likelihood calculations are

$$S_a = n^{\frac{1}{2}} \bar{E}a\{(y, t, v), \hat{\phi}\} = n^{\frac{1}{2}}(\hat{\phi} - \tilde{\phi}) \quad , \quad \rho^2 = \frac{\sigma_\phi^2}{\sigma_a^2} \quad , \quad \chi^2 = n E_{tv} \left\{ \frac{(\tilde{p}_{tv} - \hat{p}_{tv})^2}{\hat{\gamma}_{tv}} \right\} \quad ,$$

where  $\hat{\gamma}_{tv} = \hat{p}_{tv}(1 - \hat{p}_{tv})$ .

In considering mis-specified models in this example we are allowing for  $p_{1v}$  and  $p_{0v}$  to be arbitrary functions of  $v$ . But the fact that we are continuing to describe the distribution of outcome  $y$  by these two probabilities means that we are still making a tacit ignorability assumption about the allocation of patients to treatments, in the sense of Rosenbaum and Rubin (1983).

Burton and Wells (1989) discuss an observational study designed to compare rates of hospitalization for two different treatments for kidney dialysis patients. The test treatment ( $t = 1$ ) is the use of ambulatory peritoneal dialysis, the control treatment ( $t = 0$ ) is standard haemodialysis. The background to the study was a concern that although the test treatment had clinical and practical advantages over the standard procedure it may lead to the patients needing more or longer periods in hospital. Thus the response of interest is the rate of hospitalization: we take  $y = 1$  to mean an observed rate of 20 or more days in hospital per year,  $y = 0$  for less than 20 days per year (later arrivals who were on treatment for less than 3 months were omitted). This was not a randomized trial: the proportion of patients on the new treatment seemed to vary with the patient's age and appeared to change during the period of the trial. We thus use age and date of entry into the trial as covariates, both measured on a five-point scale. The data therefore take the form of a  $2 \times 2 \times 5 \times 5$  contingency table. The total sample size is  $n = 244$ . Some of the cell frequencies in the contingency table are therefore quite small, meaning that the non-parametric probabilities  $\tilde{p}_{tv}$  suffer much more sampling uncertainty than the corresponding fitted probabilities  $\hat{p}_{tv}$  calculated from the model.

Model (47) gives  $\hat{\beta}_1 = 1.899 \pm 0.299$ , very strong evidence that the new treatment is associated with more days in hospital. Under the logistic model the average of  $\hat{p}_{tv}$  over the marginal distribution of  $v$  is 0.252 for  $t = 0$  and 0.683 for  $t = 1$ , giving  $\hat{\phi} = 0.431 \pm 0.060$ . The average causal effect is to increase the chance of needing more than 20 days in hospital by a factor of almost three. If all these patients had been treated with  $t = 1$  the model estimates that there would have been three times as many patients with  $y = 1$  than if they had all been given  $t = 0$ .

The non-parametric calculation gives  $\tilde{\phi} = 0.459 \pm 0.065$  and  $\chi^2 = 65.9$ , and under the model we get  $\rho = 0.922$  and  $\sigma_\phi = 0.877$ . Figure 5 shows the likelihoods  $L_f(\phi)$  (solid line),  $L_{NP}(\phi)$  (dotted line) and  $L_{ENV}(\phi)$  (dashed line). We see that  $L_{ENV}$  comfortably covers  $L_{NP}$ , reflecting the fact that  $S_{u_a}$  in (17) is 1.00, much less than  $z_\alpha = 1.96$  defining the boundary of  $\hat{U}$  in (27). The fact that  $L_f$  is only a modest shift from  $L_{NP}$  indicates that the conventional analysis is not particularly sensitive to the logistic assumptions, in particular to the assumption that the effects of age and entry date are additive and linear. However, had other but equally plausible assumptions been made about how  $y$  depends on  $t$  and  $v$ ,

the resulting likelihood for  $\phi$  could have been noticeably different; Figure 5 suggests that for a measure of overall uncertainty it might be safer to increase the nominal standard error of  $\hat{\phi}$  by about 50%.

## 5 Comments

1. Our discussion has focussed on assessing the fit of the chosen model  $f$  within the wider model  $g_u$ ; there is no discussion of the goodness of fit of  $f$  *per se*, or of any adjustment for multiple testing with different  $u$ 's on the same data. Suppose we are interested in some particular value  $\phi = \phi_0$ , and that at this value  $L_f(\phi_0)$  is large and negative but for some  $u \in \hat{\mathcal{U}}$  in (27) we find that  $L_u(\phi_0)$  is much closer to zero. Then if we claim from  $L_f(\phi_0)$  that  $\phi_0$  is untenable in the light of the data, we are immediately open to the challenge that our analysis is entirely due to the choice of model: had we chosen an equally plausible model  $g_u$  instead we would come to the opposite conclusion. By replacing  $L_f(\phi)$  by  $L_{ENV}(\phi)$  we aim for a more cautious inference which is not open to such challenges.

2. We have assumed throughout that the data take the form of an independent and identically distributed sample of a single random variable  $x$ . We have brought more realistic settings into this general framework by assuming that data structures are also random: in Section 3.2 we have assumed that the strata of each observation is chosen at random, in Section 3.3 the censoring times are random and in Section 4.2 the treatments and covariate levels are jointly random. First-order asymptotic results are the same as those of the more natural approach of separate samples within *stata*/covariates or of conditional inference given the pattern of non-informative censoring. For example in the case of two independent samples we can check directly that if we have two separate mis-specification directions  $u_1$  and  $u_2$  then the profile likelihood for  $\phi = \phi(\eta_1, \eta_2)$  depends to first order on a locally linear function  $S = S(S_{u_1}, S_{u_2})$ . Assuming that  $|S|$  is less than the corresponding percentage point of its null distribution when  $\epsilon_1 = \epsilon_2 = 0$  then gives exactly the same envelope likelihood as before. For finite samples however the situation is much more subtle: how we condition on such data structures is then an important aspect of inference. By confining our discussion to first-order asymptotics we have avoided all such subtleties.

3. In Section 2.3 we described  $L_{NP}(\phi)$  in (20) as the non-parametric (psuedo-) likelihood for  $\phi$ . However  $L_{NP}$  is not non-parametric in the usual sense. Although it is centred on the fully non-parametric estimate  $\tilde{\phi}$ , we have used the variance  $n^{-1}\sigma_a^2$  evaluated at the parametric model  $f$  rather than non-parametrically. In Example 3, for instance, we have evaluated  $\sigma_a^2$  in (48) at the fitted probabilities  $\hat{p}_{tv}$  rather than at the observed relative frequencies  $\tilde{p}_{tv}$ . This gives more stable variance estimates, but may be misleading if the model is substantially mis-specified. Similarly, in Example 2, the confidence limits we have shown for the Kaplan-Meier survival curves are based on variance calculations under the exponential model.

4. Also in Section 2.3, we looked at the interval  $\mathcal{I} = \{\omega_-(z_\alpha), \omega_+(z_\alpha)\}$  and pointed out that although the confidence interval property  $P(\omega \in \mathcal{I}) \geq 1 - \alpha$  holds when sup / inf in (23) is taken over  $u$ 's in a fixed set  $\mathcal{U}$ , it does not hold for the interval (30) derived from the random set  $\hat{\mathcal{U}}$ . Suppose that the data are generated from  $g_u$  for some fixed  $\epsilon$  and  $u$ . Then it is easy to show that the asymptotic coverage of  $\mathcal{I}$  in (30) is an increasing function of  $\lambda_u$ ,

with lower bound  $P(\omega \in \mathcal{I}) \geq H_1(n^{\frac{1}{2}}\epsilon)$  where

$$H_1(n^{\frac{1}{2}}\epsilon) = 1 - \left\{ \Phi \left( -\frac{1 + \sqrt{1 - \rho^2}}{\rho} z_\alpha + n^{\frac{1}{2}}\epsilon\lambda_M \right) + \Phi \left( -\frac{1 + \sqrt{1 - \rho^2}}{\rho} z_\alpha - n^{\frac{1}{2}}\epsilon\lambda_M \right) \right\} ,$$

with equality attained when  $\lambda_u = \lambda_M$  in (18). This cannot be greater than or equal to  $1 - \alpha$  for all  $\epsilon$ . However, we might hope that  $H_1(n^{\frac{1}{2}}\epsilon)$  is large for those values of  $\epsilon$  for which  $u$  is likely to be included in  $\widehat{\mathcal{U}}$ . A similar asymptotic calculation shows that

$$H_2(n^{\frac{1}{2}}\epsilon) = P(u \in \widehat{\mathcal{U}}) = P(|S_u| \leq z_\alpha) = 1 - \{ \Phi(-z_\alpha + n^{\frac{1}{2}}\epsilon) + \Phi(-z_\alpha - n^{\frac{1}{2}}\epsilon) \} ,$$

and that the events  $\{\omega \in \mathcal{I}\}$  and  $\{u \in \widehat{\mathcal{U}}\}$  are independent. Figure 6 plots  $H_1$  against  $H_2$  for  $\alpha = 0.05$  and for  $\rho = 0.1, 0.2, \dots, 0.9, 0.95, 0.975, 0.99$ . Each curve traces the point  $(H_1, H_2)$  from  $n^{\frac{1}{2}}\epsilon = 0$  (right hand end) to  $n^{\frac{1}{2}}\epsilon = \infty$  (left hand end). Although there is no guarantee that  $H_1 \geq 0.95$ , we see from Figure 6 that the coverage can only fall short of this threshold when  $P(u \in \widehat{\mathcal{U}})$  is quite small, *i.e.* when  $g_u$  is unlikely to be judged statistically equivalent to  $f$ .

5. The finite case discussed in Section 4 can be generalized to a much wider parametric setting. The values of  $u_j = u(d_j, \hat{\theta})$  in (39) can be written as a linear combination of a finite number of basis functions  $w_i(x)$ ,

$$u_j = \sum_i \beta_i w_i(d_j) .$$

If, for all  $i$ ,  $w_{ij} = w_i(d_j)$  satisfy the orthonormal constraints

$$\sum_j p_j w_{ij} = \sum_j p_j s_j w_{ij} = 0 , \quad \sum_j p_j w_{ij}^2 = 1 , \quad \sum_j p_j w_{ij} w_{lj} = 0 \quad (l \neq i) \quad (49)$$

and  $\sum_i \beta_i^2 = 1$ , then the  $u_j$ 's automatically satisfy the constraints set out in (40). Writing the sums in (49) as expectations over the fitted model we see that this formulation applies just as well to a more general setting in which  $u(x) = u(x, \hat{\theta})$  is restricted to the parametric form

$$u(x) = \sum_{i=1}^k \beta_i w_i(x) , \quad (50)$$

where  $w_1(x), w_2(x), \dots, w_k(x)$  are given orthonormal basis functions with respect to the distribution  $f(x, \hat{\theta})$ . For example if  $f$  is normal we could take  $k = 2$  with  $w_1(x)$  and  $w_2(x)$  as Hermite polynomials of degrees 3 and 4. This would allow for small departures from normality in the direction of skewness and kurtosis. With  $u(x)$  in (50) we get

$$\lambda_u = \frac{1}{\sigma_\phi} \sum_i \beta_i E_{\hat{\theta}}\{w_i(x)a(x, \hat{\phi})\} , \quad S_u = n^{\frac{1}{2}} \sum_i \beta_i \bar{E}w_i(x) , \quad \chi^2 = n \sum_i \{\bar{E}w_i(x)\}^2 . \quad (51)$$

As  $g_u$  is now fully parametric we could find the MLEs of the  $\beta_i$ 's in the usual way: the fitted distribution then gives the analogue of the non-parametric estimate  $\tilde{\phi}$ , and  $\chi^2$  in (51) is the natural  $\chi^2$  statistic for testing the significance of these  $\beta_i$ 's. The quantity  $b$  in (41) follows

by writing  $\sum q_i a_i$  as  $\bar{E}a(x, \hat{\phi})$ , which leads to likelihood functions for  $\phi$  using exactly the same development as set out in Section 4.1.

6. When  $\rho = \pm 1$ , likelihood is first-order robust in the sense that  $L_{ENV}(\phi) = L_f(\phi)$ . This happens if  $\theta$  itself is the parameter of interest, so  $\phi = \theta$  and  $a(x, \theta) \propto s(x, \theta)$  (for simplicity assume that  $\theta$  is scalar). The likelihood difference  $L_u(\theta) - L_f(\theta)$  is then of order  $O(n^{-\frac{1}{2}})$ , a second-order approximation compared to the first-order approximations discussed in this paper. Approximating this difference is much more delicate than the first-order asymptotics presented in Section 2.3, since the second-order terms involve the dependence of  $u(x, \theta)$  on  $\theta$ , not just its value at  $\theta = \hat{\theta}$ . Suppose that  $u^* = u^*(x)$  is a function of  $x$  satisfying the normalization constraints (6) at  $\theta = \hat{\theta}$ , and define

$$u(x, \theta) = \frac{u^* - E_{\theta}u^* - I^{-1}\{E_{\theta}(u^*s)\}s}{\sqrt{\{E_{\theta}u^{*2} - (E_{\theta}u^*)^2 - I^{-1}(E_{\theta}u^*s)^2\}}} .$$

This means that the direction  $u(x, \theta)$  is the same for different  $\theta$  save for the weak dependence necessary for  $u(x, \theta)$  to satisfy (6) for all  $\theta$ . In this case we find

$$L_u(\theta) = L_f(\theta) + \frac{1}{2}\omega^2 I^{-1}E_{\hat{\theta}}(u^*t)\bar{E}(u^*) + O(n^{-1}) , \quad (52)$$

where  $t = t(x) = \partial s(x, \hat{\theta})/\partial \theta + s^2(x, \hat{\theta})$  (by the Bartlett identity  $t$  has zero expectation under the fitted model). Note that the correction term in (52) is symmetrical about  $\theta = \hat{\theta}$ , so the effect is to adjust the scale but leave the MLE unchanged. Bounds to the correction term when  $|S_u| \leq z_{\alpha}$  follow as in Section 2.3.

This is closely related to the robust likelihood proposed by Royal and Tsou (2003). Their adjusted likelihood, and its approximation in our setting, is

$$L_{RT}(\theta) = \frac{-\bar{E}(\partial s/\partial \theta)}{\bar{E}(s^2)}L_f(\theta) = L_f(\theta) + \frac{1}{2}\omega^2 I^{-1}\bar{E}(t) + O(n^{-1}) . \quad (53)$$

This is the same as (52) for the particular direction  $u^* = \sigma_{t^*}^{-1}t^*$  where  $t^* = t - E_{\hat{\theta}}(st)I^{-1}s$  and  $\sigma_{t^*}^2 = \text{Var}_{\hat{\theta}}(t^*)$ . Royal and Tsou make no such assumption about the direction of misspecification. In the discrete case (but not generally) we can also cover all possible  $u^*$ 's by finding the unrestricted bounds of the correction term in (52). In the notation of Section 4.1 we get the envelope likelihood

$$L_{ENV}(\theta) = L_f(\theta) + \frac{1}{4}\omega^2 I^{-1}\{\bar{E}(t) + n^{-\frac{1}{2}}\sigma_{t^*} \chi\} + O(n^{-1}) . \quad (54)$$

We can show that the size of the correction in (54) is always *greater* than that in (53). Likelihood  $L_{ENV}$  always indicates an increase in uncertainty whereas  $L_{RT}$  may give more or less uncertainty depending on the sign of  $\bar{E}(t)$ .

## References

- Amari, S. (1985) *Differential-Geometric Methods in Statistics*. Lecture Notes in Statistics, vol. **28**. New York: Springer.
- Armitage, P., McPherson, C. K. and Copas, J. B. (1969) Statistical studies of prognosis in advanced breast cancer. *J. Chronic Dis.*, **22**, 343-360.
- Burton, P. R. and Wells, J. (1989) A selection adjusted comparison of hospitalization on continuous ambulatory peritoneal dialysis and haemodialysis. *J. Clin. Epidemiol.*, **42**, 531-539.
- Copas, J. B. and Eguchi, S. (2005) Local model uncertainty and incomplete data bias (with discussion). *J. Roy. Statist. Soc. B*, **67**, 459-513.
- Cox, D. R. (2006) *Principles of Statistical Inference*. Cambridge: Cambridge University Press.
- Desmond, A. F. and Godambe, V. P. (1998) Estimating functions. In *Encyclopedia of Biostatistics* (eds. P. Armitage and T. Colton), **2**, pp. 1375-1386. Chichester: Wiley.
- Gustafson, P. (2001) On measuring sensitivity to parametric model misspecification. *J. Roy. Statist. Soc. B*, **63**, 81-94.
- Henmi, M. and Eguchi, S. (2004) A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, **91**, 929-941.
- Huber, P. J. (1981) *Robust Statistics*. London: Chapman and Hall.
- Kent, J. T. (1986) The underlying structure of non-nested hypothesis tests. *Biometrika*, **73**, 333-343.
- Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.
- Royall, R. and Tsou, T. (2003) Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *J. Roy. Statist. Soc. B*, **65**, 391-404.
- Severini, T. A. (2000) *Likelihood methods in Statistics*. Oxford: Clarendon Press.
- Shao, J. (2003) *Mathematical Statistics*, 2nd edition. New York: Springer-Verlag.
- Small, C. G. and Wang, J. (2003) *Numerical Methods for Non-linear Estimating Equations*. Oxford: Clarendon Press.

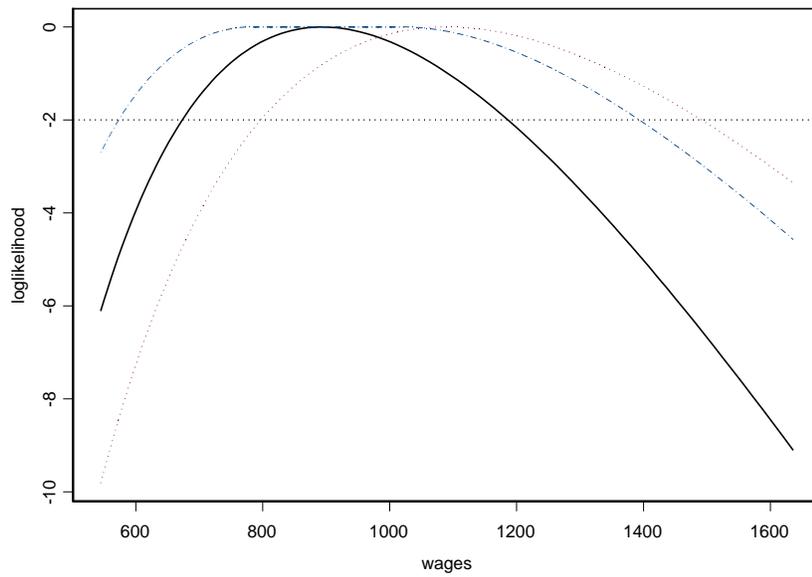


Figure 1: Likelihoods for income data

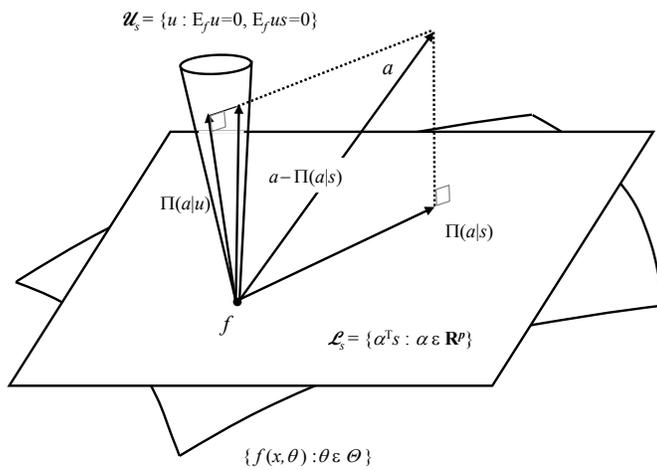


Figure 2: Tangent and orthogonal spaces

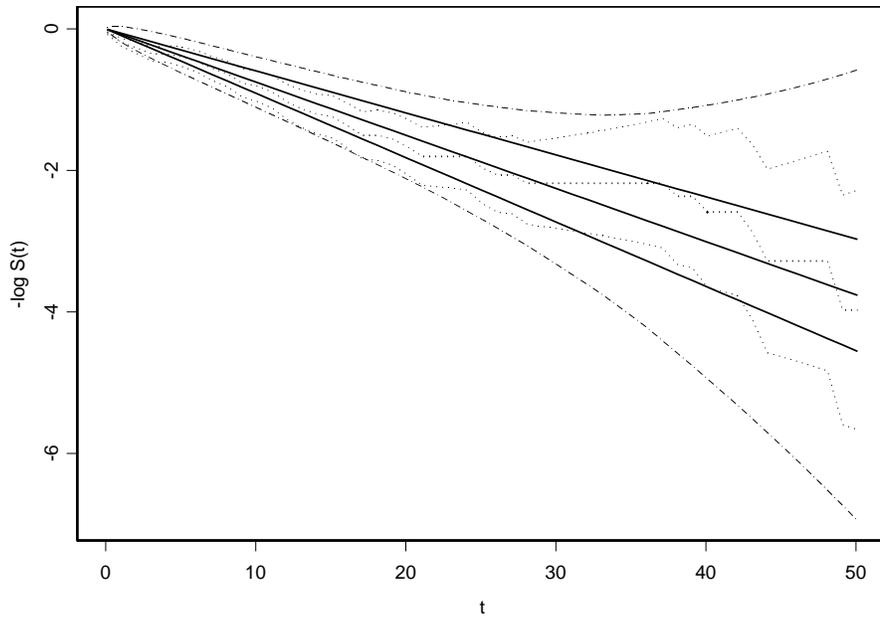


Figure 3: Confidence intervals for Treatment A

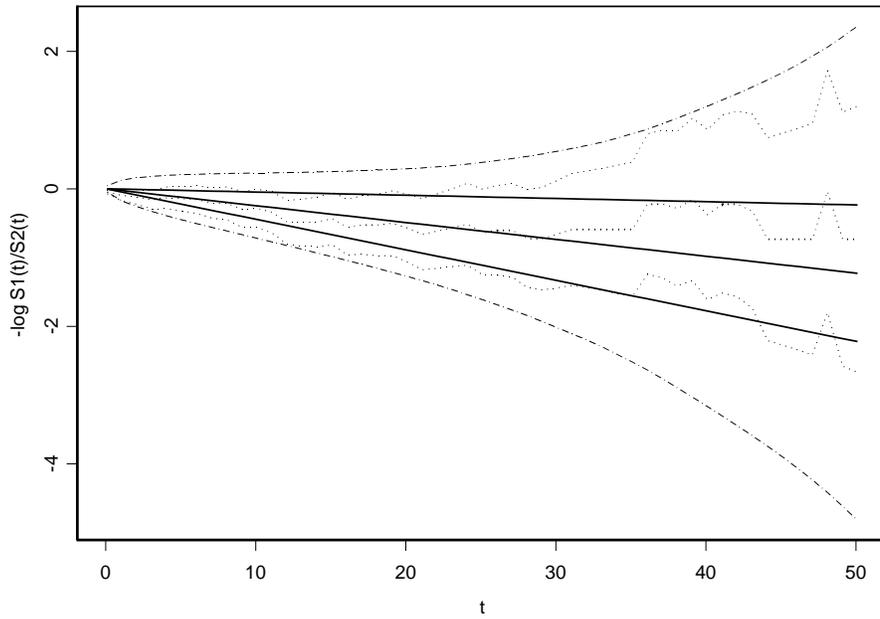


Figure 4: Confidence intervals for comparing Treatments A and B

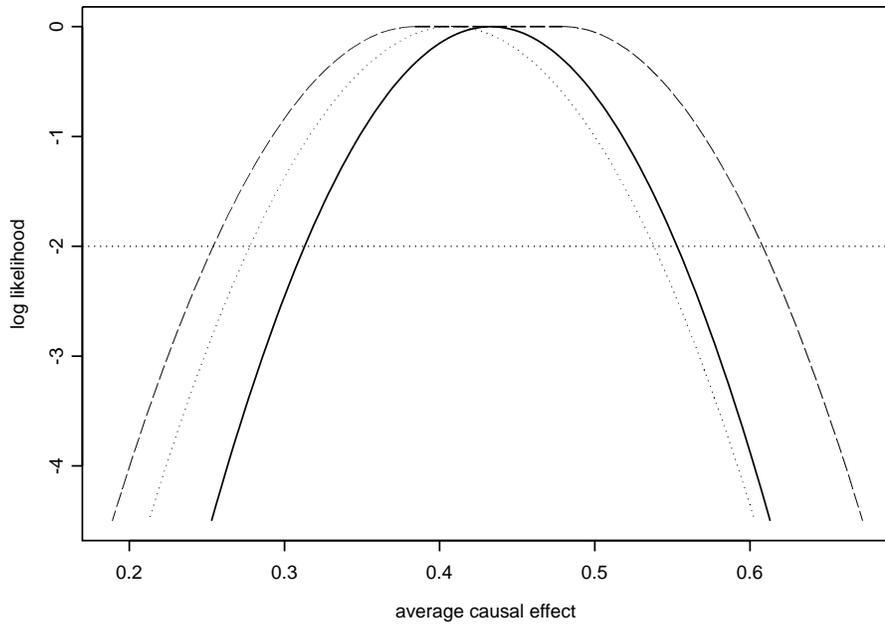


Figure 5: Likelihoods for dialysis data

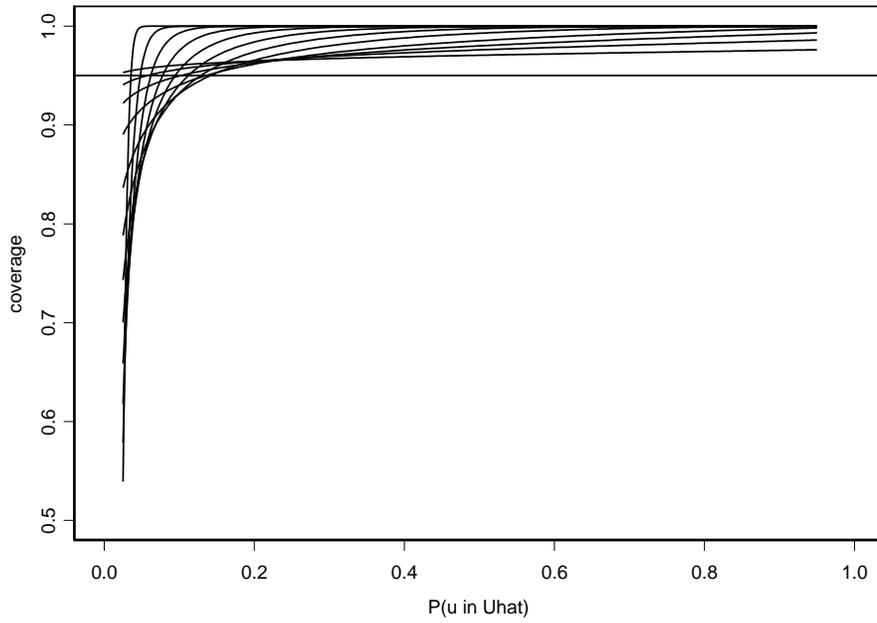


Figure 6: Coverage of the envelope likelihood confidence interval