ON PARTIAL LIKELIHOOD AND THE CONSTRUCTION OF FACTORISABLE TRANSFORMATIONS

H. S. BATTEY, D. R. COX[†], AND SU HYEONG LEE

SUMMARY. Models whose associated likelihood functions fruitfully factorise are an important minority allowing elimination of nuisance parameters via partial likelihood, an operation that is valuable in both Bayesian and frequentist inferences, particularly when the number of nuisance parameters is not small. After some general discussion of partial likelihood, we focus on marginal likelihood factorisations, which are particularly difficult to ascertain from elementary calculations. We suggest a systematic approach for deducing transformations of the data, if they exist, whose marginal likelihood functions are free of the nuisance parameters. This is based on the solution to an integro-differential equation constructed from aspects of the Laplace transform of the probability density function, for which candidate solutions solve a simpler first-order linear homogeneous differential equation. The approach is generalised to the situation in which such factorisable structure is not exactly present. Examples are used in illustration. Although motivated by inferential problems in statistics, the proposed construction is of independent interest and may find application elsewhere.

Some key words: inferential separation; marginal likelihood; matched comparisons; method of characteristics; partial differential equations; nuisance parameters.

1. INTRODUCTION

Parametric statistical inference in models with many parameters relative to the number of observations raises issues that are at least implicitly differential geometric. For inference on a scalar or vector interest parameter ψ, the profile log-likelihood function for ψ replaces nuisance parameters by their constrained maximum likelihood estimates. The approach may give highly miscalibrated inference for ψ when the number of nuisance parameters is large relative to the amount of information in the sample, and considerably suboptimal inference for even moderately many such parameters. Three broad approaches are to apply an interest-respecting orthogonal reparameterisation, to base inference on an adjusted version of the signed-likelihood root called r^*, or to construct a suitable partial likelihood.

Parameter orthogonalisation, proposed by Cox and Reid (1987) and based on the solution to a set of partial differential equations, leads to higher-order accuracy of nuisance maximum likelihood estimators in a moderate deviation range of ψ, and thereby removes the leading-order bias term in the maximum likelihood estimator of

Date: June 18, 2022.

Note on the posthumous version: A complete draft of the work was read and edited by all authors in November 2021. D. R. Cox died suddenly on 18 January 2022. This final version includes revisions and additions to that draft.

ψ. This can be observed from the expansions on p.150 of Barndorff-Nielsen and Cox (1994) on noting that parameter orthogonalisation, by definition, sets relevant blocks of the Fisher information matrix to zero. Two asymptotically equivalent versions of r^* are due to Barndorff-Nielsen (1990) and Fraser (1990), the latter using implicit conditioning on an approximately ancillary statistic. Partial likelihood, formalised to some extent by Cox (1975), uses only part of the full likelihood function for inference on ψ, possibly relinquishing some information. All three approaches, when available, often yield remarkably accurate inference when the amount of information is small, or nuisance parameters are numerous.

Cox (1975) stated five problems associated with partial likelihood, of which the first was: "to provide constructive procedures for finding useful partial likelihoods". The problem remains open, and the purpose of the present paper is to provide a modest step towards addressing it, focussing primarily on a class of matched comparison problems. Section 2 gives a formalisation of inferential separation based on partial likelihood, while §4 presents a version suitable for matched comparison settings.

2. Partial likelihood and inferential separation

Suppose the outcomes are realisations of random variables $Y_1, \ldots, Y_b$ whose joint distribution depends on parameters (ψ, λ). The notation b in place of the more conventional n is for consistency with the rest of the paper, in which n is the total sample size, Y_i is a vector of size m, and $b = n/m$ represents the number of blocks of size m. In the present section such block structure is not needed.

Consider inference on an interest parameter ψ from a parameter-based factorisation of the full likelihood function $L(\psi, \lambda; y)$ of the form

$$L(\psi, \lambda; y) = L_{\rm pa}(\psi; y) L_{\rm r}(\psi, \lambda; y). \tag{2.1}$$

The factor $L_{\rm pa}(\psi; y)$ is called the partial likelihood (Cox, 1975). Ideally, little or no information for inference on ψ is lost through relinquishment of the remainder likelihood $L_{\rm r}(\psi, \lambda; y)$.

A factorisation of the form (2.1) can be induced in various ways, as is apparent from the examples in Cox (1975). The important distinction between factorisations on the parameter space and those on the sample space is discussed in detail by Cox and Wermuth (1999) and Cox (2000). The present paper is concerned with factorisations on the parameter space induced through sample-space factorisations of the marginal and conditional types.

Let (S, R, A), where A is ancillary, be a jointly sufficient statistic for (ψ, λ) based on $Y = (Y_1, \ldots, Y_b)$, where any of Y_i, ψ or λ may be vectors. An ancillary statistic need not be available. When it is, (S, R, A) is treated as minimal sufficient.

The full likelihood function for (ψ, λ) is the density function f_Y of Y, viewed as a function of the parameters with y fixed at the observed value of Y:

$$f_Y(y; \psi, \lambda) \propto f_{S,R,A}(s, r, a; \psi, \lambda) = f_A(a) f_{S|A}(s \mid a; \psi, \lambda) f_{R|S,A}(r \mid s, a; \psi, \lambda).$$

That it may be beneficial, or indeed necessary, to use only part of the likelihood function for inference on ψ was noted by Bartlett (1936), who proposed conditional likelihood based on $f_{R|S,A}$ when this quantity is free of λ. Another special case

is $f_{S|A}(s \mid a; \psi, \lambda) = f_{S|A}(s \mid a; \psi)$ so that one choice of $L_{\mathrm{pa}}(\psi; y)$ is $f_{S|A}(s \mid a; \psi)$. Similarly, if

$$f_{S|A}(s \mid a; \psi, \lambda) = g_1(s, a; \psi) g_2(s, a; \psi, \lambda),$$

a λ-free partial likelihood for ψ can be constructed as $L_{\mathrm{pa}}(\psi; y) = g_1(s, a; \psi)$. In the absence of an ancillary statistic, $f_{S|A}(s \mid a; \psi, \lambda) = f_S(s; \psi, \lambda)$, and when this is free of λ, $L_{\mathrm{pa}}(\psi; y) = f_S(s; \psi)$ is called the marginal likelihood function (Fraser, 1968; Kalbfleish and Sprott, 1970).

3. Matched comparison problems

Matched comparison studies in blocks of size m represent a popular and effective method of experimentation for the assessment of m treatment effects, of which one typically represents a control or base treatment. Such designs entail matching b sets of m individuals based on their intrinsic features before randomising each treatment to one unit per block.

With $m = 2$ it is convenient to write $(Y_i)_{i=1}^b = (T_i, C_i)_{i=1}^b$ for outcomes on the treated and untreated individuals in each of the b pairs. The distributions of T_i and C_i are typically assumed equal up to the presence of a treatment parameter, ψ, and known modulo a pair-specific nuisance parameter, λ_i. Inclusion of a nuisance parameter per pair avoids having to specify in detail those aspects of the probabilistic model that are not of primary subject-matter importance. The nuisance parameters may, for example, encompass the effects of unmeasured covariates in the form $\lambda_i = h(x_i^{\mathrm{T}}\beta)$. It also makes the assumption of independence of T_i and C_i reasonable. There are limitations to this approach. Notably non-constancy of the treatment effect would not be directly detectable through consideration of all pairs simultaneously.

A key consideration is whether the nuisance parameters $\lambda_1, \ldots, \lambda_b$ are to be treated as realisations of random variables or as fixed arbitrary constants. Appendix A formalises the two formulations and argues in favour of the latter. While conceptually compelling, the analysis for fixed nuisance parameters is more challenging and context-dependent. When available, it tends to be based on a form of partial likelihood obtained from distributional factorisations on the sample space: either marginal or conditional likelihood. Motivated by this, the present paper seeks a function $s(Y_i)$ of the components of Y_i such that the density function of $S_i = s(Y_i)$ depends on ψ but not on λ_i. We call S_i a factorisable transformation of Y_i. A function s inducing such factorisable structure need not exist, raising the question of whether useful approximate versions may be found.

For larger values of m, typically $\psi = (\psi_1, \ldots, \psi_m)^{\mathrm{T}}$ represents a vector of treatment effects, and we let $Y_{i1}, \ldots, Y_{im}$ represent the independent outcome variables associated with each treatment in the ith block, $i = 1, \ldots, b$.

The formulation of matched comparison problems as inducing large numbers of nuisance parameters for inference on a scalar treatment effect was presented by Cox (1958b), and has been discussed from various perspectives by Cox and Hinkley (1978), Lindsay (1980,1985), Barndorff-Nielsen and Cox (1994), Sartori (2003), Kartsonaki and Cox (2016) and Battey and Cox (2020). The following results illustrate some of the difficulties.

Example 1. Bartlett (1936) noted that when two random variables, Y_{i1} and Y_{i2} say, are normally distributed of mean μ_i and variance σ^2 for $i = 1, \ldots, b$, the maximum likelihood estimator of σ^2 converges in probability to $\sigma^2/2$.

Example 2 (Cox, 1958b)**.** Let $(T_i, C_i)_{i=1}^b$ be independent pairs of independent binary variables, each valued in $\{0, 1\}$, and let

$$\mathrm{pr}(C_i = 1) = \frac{e^{\lambda_i}}{1 + e^{\lambda_i}}, \quad \mathrm{pr}(T_i = 1) = \frac{e^{\lambda_i + \psi}}{1 + e^{\lambda_i + \psi}}.$$

The interest parameter ψ is the logistic difference between the probabilities. Cox (1958b) indicated the appropriate conditional analysis, while Barndorff-Nielsen and Cox (1994, example 4.6) demonstrated difficulties with direct use of the likelihood function for estimation of ψ, showing that the estimator converges in probability to 2ψ.

Example 3. Suppose that T_i and C_i are independently exponentially distributed of rates $\lambda_i\psi$ and λ_i/ψ respectively for $i = 1, \ldots, b$. Thus T_i and C_i have a constant hazard ratio ψ^2. A slightly different parameterisation of this example was used by Lindsay (1980, 1985).

The likelihood equations for $\hat\psi$ and $\hat\lambda_1, \ldots, \hat\lambda_b$ are

$$\begin{aligned} 0 &= \nabla_{\lambda_i} \ell(\hat\psi, \hat\lambda_1, \ldots, \hat\lambda_b) = 2\hat\lambda_i^{-1} - (T_i\hat\psi + C_i/\hat\psi), \\ 0 &= \nabla_{\psi} \ell(\hat\psi, \hat\lambda_1, \ldots, \hat\lambda_b) = -\sum_{i=1}^b \hat\lambda_i T_i + \sum_{i=1}^b \hat\lambda_i C_i/\hat\psi^2. \end{aligned}$$

Thus, on substituting $\hat\lambda_i = 2(T_i\hat\psi + C_i/\hat\psi)^{-1}$ into the likelihood equation for $\hat\psi$ and simplifying, the maximum likelihood estimator $\hat\psi$ is the solution in ψ to

$$0 = \sum_{i=1}^b \frac{C_i/\psi - T_i\psi}{C_i/\psi + T_i\psi}. \tag{3.1}$$

Lindsay (1985) established that this estimator of the constant hazard ratio is consistent and asymptotically normally distributed as $n = 2b \to \infty$. This is in contradistinction to the seemingly similar Examples 1 and 2. However, the usual estimator of variance of $\hat\psi$ is miscalibrated. Lindsay (1985) recommended a treatment in which $\lambda_1, \ldots, \lambda_b$ are regarded as independent and identically distributed random variables from a parametric distribution of known form. Mispecification of the form can lead to major difficulties.

We favour a transformation to $S_i = T_i/C_i$, whose density function at $s > 0$ is $\psi^2/(1 + \psi^2 s)^2$ producing the marginal likelihood,

$$L_{\mathrm{pa}}(\psi; y) = \prod_{i=1}^b f_S(s_i; \psi),$$

as a fruitful choice of partial likelihood. This is to be viewed as a function of ψ. Since $(S_i)_{i=1}^b$ are independent and identically distributed, consistency of the marginal likelihood estimator $\hat\psi_{\mathrm{m}}$ is expected in view of standard maximum likelihood

theory, although verification of the usual regularity conditions is complicated by non-existence of moments of S_i. The proof of Proposition 3.1 establishes consistency directly.

Proposition 3.1. *Let $\hat{\psi}_{\mathrm{m}}$ be the marginal maximum likelihood estimator based on the transformed random variables $(S_i)_{i=1}^b$. Then $\hat{\psi}_{\mathrm{m}}$ is consistent for ψ as $b \to \infty$.*

Proof. The marginal likelihood equation for $\hat{\psi}_{\mathrm{m}}$ is

$$1 = \frac{2}{b} \sum_{i=1}^{b} \frac{\hat{\psi}_{\mathrm{m}}^2 S_i}{1 + \hat{\psi}_{\mathrm{m}}^2 S_i}. \tag{3.2}$$

A strong law of large numbers implies that, for any $\kappa > 0$,

$$\frac{1}{b} \sum_{i=1}^{b} \frac{\kappa^2 S_i}{1 + \kappa^2 S_i} \to_{a.s.} \frac{\kappa^2 \{(\kappa - \psi)(\kappa + \psi) + 2\psi^2(\log \psi - \log \kappa)\}}{(\kappa^2 - \psi^2)^2}.$$

Thus in the limit as $b \to \infty$, $\hat{\psi}_{\mathrm{m}}$ satisfies

$$1 = \frac{2\hat{\psi}_{\mathrm{m}}^2 \{(\hat{\psi}_{\mathrm{m}} - \psi)(\hat{\psi}_{\mathrm{m}} + \psi) + 2\psi^2(\log \psi - \log \hat{\psi}_{\mathrm{m}})\}}{(\hat{\psi}_{\mathrm{m}}^2 - \psi^2)^2}. \tag{3.3}$$

The right hand side of (3.3) is 1 only in the limit as $\hat{\psi}_{\mathrm{m}} \to \psi$. □

It follows directly from the consistency established in Proposition 3.1 that $(\hat{\psi}_{\mathrm{m}} - \psi)\{-\ell''_{\mathrm{pa}}(\hat{\psi})\}^{1/2}$ is asymptotically standard normally distributed, and similarly for the likelihood ratio statistic based on the marginal likelihood.

Conditional likelihood is available and fruitful in the matched comparison context if (T_i, C_i) can be transformed bijectively to new variables, (S_i, R_i) say, such that the conditional density function of each S_i given $R_i = r_i$ depends on ψ but not on λ_i. When available, this situation is typically easy to detect from inspection of the log-likelihood function, as it only requires identifying $R_i = r(T_i, C_i)$ as a sufficient statistic for λ_i. The statistic S_i can then be chosen based on convenience of calculating the conditional density, provided that the transformation $(T_i, C_i) \to (S_i, R_i)$ is bijective.

Deducing a function $S_i = s(T_i, C_i)$ that gives a suitable marginal likelihood as in Example 3 is considerably more difficult and, to our knowledge, a systematic construction has not been attempted.

4. A systematic construction of marginal likelihood

For ease of exposition the case of $m = 2$ is discussed first, suppressing the pair index i on λ_i and (T_i, C_i) except when it is necessary to be explicit.

To establish a transformation $(T, C) \to (S, R)$, write the transformation equations as $s = s(t, c)$, and $r = r(t, c)$. Since the transformation is assumed bijective, the inverse equations are $t = t(s, r)$ and $c = c(s, r)$. A transformation satisfying $f_S(s; \psi, \lambda) = f_S(s; \psi)$ is sought, using only the joint probability density or mass function of T and C, denoted by $f_{T,C}(t, c; \psi, \lambda)$. The probability function of an arbitrary transformed random variable $S = s(T, C)$ is expressible in terms of this

either by specifying the Jacobian matrix of the transformation, or by using Laplace transforms. The latter is found to be more convenient. Thus consider

$$f_S(s;\psi,\lambda) = \frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau+i\infty} \exp\{zs(t,c)\}T_\lambda(s,z)dz, \tag{4.1}$$

where $s:\mathbb{R}^2\to\mathbb{R}$, τ is anywhere in the interval of convergence of the moment generating function of S and $T_\lambda(s,z)$ is the Laplace transform,

$$T_\lambda(s,z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-zs(x,y)\}f_{T,C}(x,y;\psi,\lambda)dxdy, \quad z\in\mathbb{C}.$$

Since (4.1) may only depend on λ through T_λ, Battey and Cox (2020) suggested choosing the function $s(t,c)$ to make T_λ independent of λ, identically in z, ψ and λ. It is in fact sufficient by Cauchy's theorem (e.g. Whittaker and Watson, 1927, §5) that independence be achieved only at points z of singularity, but this is more difficult and inconsequential unless the analytic continuation of the moment generating function of S has a singularity at zero, as will become clear from the discussion below.

A function s delivering independence of λ is sought by differentiating T_λ partially with respect to λ and solving the resulting integral equation for $s(t,c)$, identically in z, ψ, and λ. This equation is

$$\frac{\partial}{\partial\lambda}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-zs(t,c)\}f_{T,C}(t,c;\psi,\lambda)dtdc = 0. \tag{4.2}$$

Depending on the support of the distributions of T and C, the range of integration may be restricted.

The partial differential operator can be interchanged with the integral sign by Leibniz's theorem but it is more fruitful to first apply a change of variables from (t,c) to (w,v), say, chosen such that all dependence on λ is transferred from $f_{T,C}(t,c;\psi,\lambda)$ to $s(w,v;\psi,\lambda)$. The conditions required for this are that $f_{T,C}(t,c;\psi,\lambda)$ is expressible as

$$f_{T,C}(t,c;\psi,\lambda) = \kappa\frac{dw(t)}{dt}\frac{dv(c)}{dc}g\{w(t,\psi,\lambda),v(c,\psi,\lambda),\psi\}, \tag{4.3}$$

where κ is a constant and g depends on λ only through the bijective functions $w(t;\psi,\lambda)$ and $v(c;\psi,\lambda)$. Assuming this is satisfied, (4.2) is

$$\frac{\partial}{\partial\lambda}\int_{\mathcal{V}}\int_{\mathcal{W}} \exp[-zs\{t(w,\psi,\lambda),c(v,\psi,\lambda)\}]g(w,v,\psi)dwdv = 0. \tag{4.4}$$

By Leibniz's theorem, the chain rule and positivity of g, a sufficient and necessary condition for s to solve the nonlinear double integral equation (4.4) for any $z\neq 0$ is that it solves the first-order linear homogeneous differential equation:

$$\frac{\partial}{\partial\lambda}s\{t(w;\psi,\lambda),c(v;\psi,\lambda)\} = 0 \tag{4.5}$$

for all permissible w and v, i.e. all values for which the corresponding values of t and c are in the support of $f_{T,C}$ for values of ψ and λ in their respective parameter spaces. Thus, provided (4.3) is satisfied, any solution to (4.5) is a solution to (4.2) identically in ψ, λ and $z\neq 0$. If no parameter-free solution to (4.5) exists, there may still be a parameter-dependent solution, as discussed in §7.1.

In contrast to the original problem (4.2), established solution strategies are available for (4.5), either exact if an analytic solution exists, or approximate. The probabilistic model $f_{T,C}$ may supply additional information to aide solution.

Using the chain rule, the partial differential equation (4.5) can be written

$$\frac{\partial s\{t(w,\psi,\lambda),c(v,\psi,\lambda)\}}{\partial t}\frac{\partial t(w,\psi,\lambda)}{\partial \lambda}+\frac{\partial s\{t(w,\psi,\lambda),c(v,\psi,\lambda)\}}{\partial c}\frac{\partial c(v,\psi,\lambda)}{\partial \lambda}=0.$$

More compactly, with $a(t,c):=a(t,c;\lambda,\psi)=\partial t(w,\psi,\lambda)/\partial\lambda$ and $b(t,c):=b(t,c;\lambda,\psi)=\partial c(v,\psi,\lambda)/\partial\lambda$,

$$a(t,c)\frac{\partial s(t,c)}{\partial t}+b(t,c)\frac{\partial s(t,c)}{\partial c}=0. \tag{4.6}$$

This is a standard form of partial differential equation to which the method of characteristics applies. See e.g., Courant and Hilbert (1966, Chapter 2).

The approach extends naturally to $m>2$, although in that case bijectivity of the map implies that s has up to $(m-1)$ components and the analogue of (4.1) is

$$f_S(s;\psi,\lambda)=\frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau+i\infty}\cdots\int_{\gamma-i\infty}^{\gamma+i\infty}\exp\{z_1s_1+\cdots+z_{m-1}s_{m-1}\}T_\lambda(s,z)dz_1\cdots dz_{m-1}, \tag{4.7}$$

where, reintroducing subscripts for clarity and assuming the components $Y_{i1},\ldots,Y_{im}$ are independent, $T_\lambda(s,z)$ is

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\{-z_1s_1-\cdots-z_{m-1}s_{m-1}\}\prod_j f_{Y_{ij}}(y_{ij};\psi_j,\lambda_i)dy_{i1},\ldots,dy_{im}. \tag{4.8}$$

The generalisation of (4.3) is that $f_{Y_{ij}}(y_{ij};\psi_j,\lambda_i)$ can be written as

$$f_{Y_{ij}}(y_{ij};\psi_j,\lambda_i)=c_{ij}\frac{dv_{ij}(y_{ij})}{dy_{ij}}g_{ij}\{v_{ij}(y_{ij};\psi_j,\lambda_i),\psi_j\}, \tag{4.9}$$

where c_{ij} is a constant and g_{ij} only depends on λ_i through the bijective function $v_{ij}(y_{ij};\psi_j,\lambda_i)$.

In principle, these ideas extend beyond matched comparison problems, by replacing $(T_i,C_i)_{i=1}^b$ from the previous discussion by pairs of observations $(Y_i,Y_1)_{i=2}^n$, say, where $Y_1,\ldots,Y_n$ are outcome variables whose distribution possibly depends on covariates $x_1,\ldots,x_n\in\mathbb{R}^p$. There are connections to maximal invariants used in the theory of invariant tests. Indeed, a referee has pointed out that Example 3 is a location model after a log transformation, from which it is clear that differences on the log-scale eliminate the nuisance parameter, and that this must be related to the existence of a maximal invariant, or maximal ancillary statistic for $(\lambda_1,\ldots,\lambda_b)$.

Several examples clarify these ideas.

5. Matched comparison examples

Example 4. Suppose that T_i and C_i are independently exponentially distributed of rates $\lambda_i\psi$ and λ_i/ψ respectively for $i=1,\ldots,b$. Example 3 shows that a factorisable transformation exists in the form $S_i=T_i/C_i$. The motivation for the present paper was that this should be recoverable from a seamless application of theory.

Equation (4.2) is, on treating each pair separately and suppressing the subscript on λ,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\lambda}\int_0^\infty\int_0^\infty \exp\{-zs(t,c)\}\left\{\lambda^2\exp(-\lambda\psi t)\exp(-\lambda c/\psi)\right\}dtdc \qquad (5.1)\\ &= \int_0^\infty\int_0^\infty \exp\{-zs(t,c)\}\left\{2\lambda-\lambda^2(\psi t+c/\psi)\right\}\exp(-\lambda\psi t)\exp(-\lambda c/\psi)dtdc. \end{aligned}$$

Integration shows that $s(t,c)=t/c$ verifies equation (4.2). The goal is to recover this transformation using a strategy that does not require $s(t,c)$ to be known a priori.

Following the recommendation above, change variables in equation (5.1) to $w=\lambda\psi t$ and $v=\lambda c/\psi$. The volume element transforms as $dtdc=(\psi\lambda)^{-1}(\lambda/\psi)^{-1}dwdv=\lambda^{-2}dwdv$ so that equation (5.1) is

$$0=\frac{\partial}{\partial\lambda}\int_0^\infty\int_0^\infty \exp[-zs\{w/(\lambda\psi),\psi v/\lambda\}]\exp\{-(w+v)\}dwdv. \qquad (5.2)$$

Dependence on λ has been transferred to the bivariate function s for which a solution to (5.2) is required. Interchanging the partial differential operator with the integrals shows that a solution to (5.2) is that of the first order linear partial differential equation

$$\frac{\partial}{\partial\lambda}s\{t(w,\psi,\lambda),c(v,\psi,\lambda)\}=0, \qquad (5.3)$$

where $t(w,\psi,\lambda)=w/(\lambda\psi)$ and $c(v,\psi,\lambda)=\psi v/\lambda$, and for the purpose of this argument, w and v should be treated as fixed. Equation (5.3) specifies that $s:\mathbb{R}^2\to\mathbb{R}$ must be constant identically in v, w and ψ as a function of λ. Thus, $\{s(t,c)=(t/c)^k=(\psi^{-2}w/v)^k : k\neq 0\}$ is an equivalence class of solutions to (5.3), suggesting $(T_i,C_i)\mapsto\{(T_i/C_i),A\}$ as a suitable transformation, where A represents any statistic that makes the transformation bijective, for instance $A=T_i$, $A=C_i$ or $A=T_iC_i$.

Example 5. Example 4 easily extends to triplets, quadruplets etc. With X_{i1}, X_{i2} and X_{i3} exponentially distributed of rates $\lambda_i\psi_1$, $\lambda_i\psi_2$ and λ_i respectively, the only change to the previous derivations is that s is a function of three variables, which are, on omitting triplet subscripts, $x_1=w_1/(\lambda\psi_1)$, $x_2=w_2/(\lambda\psi_2)$ and $x_3=w_3/\lambda$. The equation to be solved is thus (cf equation (4.6))

$$a(x_1,x_2,x_3)\frac{\partial s(x_1,x_2,x_3)}{\partial x_1}+b(x_1,x_2,x_3)\frac{\partial s(x_1,x_2,x_3)}{\partial x_2}+c(y_1,y_2,y_3)\frac{\partial s(y_1,y_2,y_3)}{\partial x_3}=0,$$

say, where $a(x_1,x_2,x_3)=\partial x_1(w_1,\psi_1,\lambda)/\partial\lambda=-x_1/\lambda$, $b(x_1,x_2,x_3)=\partial x_2(w_2,\psi_2,\lambda)/\partial\lambda=-x_2/\lambda$ and $c(x_1,x_2,x_3)=\partial x_3(w_3,\lambda)/\partial\lambda=-x_3/\lambda$. There are multiple solutions. Among those yielding a bijective map from (x_1,x_2,x_3) to (s_1,s_2,r) are $s_1(x_1,x_2,x_3)=x_1/x_3$, $s_2(x_1,x_2,x_3)=x_2/x_3$ and $r=x_3$.

Example 6. An extension of Example 4 to which the appropriate transformation is not already known has T_i and C_i Weibull distributed of shape α and rate parameters $\lambda_i\psi$ and λ_i/ψ respectively. Thus there is one nuisance parameter per pair and another shared over all pairs. The appropriate change of variables is to $w=\lambda\psi t^\alpha$ and

$v = (\lambda/\psi)c^\alpha$ so that the analogue of equation (5.2) is

$$0 = \frac{\partial}{\partial\lambda}\int_0^\infty\int_0^\infty \exp(-zs[\{w/(\lambda\psi)\}^{1/\alpha}, \{\psi v/\lambda\}^{1/\alpha}])\exp\{-(w+v)\}dwdv.$$

A transformation to eliminate both α and λ would need to solve the pair of simultaneous partial differential equations

$$\begin{aligned} a_\lambda(t,c)\frac{\partial s(t,c)}{\partial t} + b_\lambda(t,c)\frac{\partial s(t,c)}{\partial c} &= 0, \\ a_\alpha(t,c)\frac{\partial s(t,c)}{\partial t} + b_\alpha(t,c)\frac{\partial s(t,c)}{\partial c} &= 0, \end{aligned}$$

where $a_\lambda(t,c) = -t/\alpha\lambda$, $b_\lambda(t,c) = -c/\alpha\lambda$, $a_\alpha(t,c) = -t\log(t)/\alpha$, $b_\alpha(t,c) = -c\log(c)/\alpha$. There is clearly no such solution. The transformation $s(t,c) = t/c$ eliminates λ, which is to be favoured over elimination of α because each pair of observations introduces a pair-specific λ_i. The transformed random variables $(S_i)_{i=1}^b = (T_i/C_i)_{i=1}^b$ are independent and identically distributed with probability density function

$$f_S(s;\psi,\alpha) = \frac{\alpha\psi^2 s^{\alpha-1}}{(1+\psi^2 s^\alpha)^2}, \quad s > 0. \tag{5.4}$$

To assess whether it is possible to also eliminate α, consider a pair of these transformed random variables, S_i and S_j, say. The density function of a transformation $u(s_i, s_j)$ satisfies

$$f_U(u;\psi,\alpha) = \frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau+i\infty}\exp\{zu(s_i,s_j)\}T_\alpha(u,z)dz,$$

where

$$T_\alpha(u,z) = \int_0^\infty\int_0^\infty \exp\{-zu(x,y)\}f_{S_i,S_j}(x,y;\psi,\alpha)dxdy,$$

and analogously to before we seek a solution to $(\partial/\partial\alpha)T_\alpha(u,z) = 0$. It does not seem possible to specify a change of variables from (s_i, s_j) to (w, v) such that the dependence on α is transferred from

$$f_{S_i,S_j}(s_i,s_j;\psi,\alpha) = \frac{\alpha\psi^2 s_i^{\alpha-1}}{(1+\psi^2 s_i^\alpha)^2}\frac{\alpha\psi^2 s_j^{\alpha-1}}{(1+\psi^2 s_j^\alpha)^2}, \quad s_i > 0, s_j > 0,$$

to $u\{s_i(w,\psi,\alpha), s_j(v,\psi,\alpha)\}$. In other words, the analogue of equation (4.9) appears to be violated for this relatively complicated form of $f_S(s;\psi,\alpha)$ given in equation (5.4).

The following example, although artificial and having an obvious solution, serves as a reassuring illustration that the distributions involved need not belong to the exponential family.

Example 7. Suppose that T_i and C_i are independently Cauchy distributed of location λ_i and shape ψ, the interest parameter. In the notation of §4,

$$\begin{aligned} T_\lambda(s,z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-zs(t,c)\}\frac{1}{\pi\psi}\left\{1+\left(\frac{t-\lambda}{\psi}\right)^2\right\}^{-1}\left\{1+\left(\frac{c-\lambda}{\psi}\right)^2\right\}^{-1} dtdc \\ &= \frac{1}{\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-zs(\psi w+\lambda, \psi v+\lambda)\}\frac{dwdv}{(1+w^2)(1+v^2)}, \end{aligned}$$

where we have used a change of variables to $w = (t-\lambda)/\psi$ and $v = (c-\lambda)/\psi$. The function $s : \mathbb{R}^2 \to \mathbb{R}$ must be constant identically in v, w and ψ as a function of λ showing that any λ-free function of $(t-c) = \psi(w-v)$ is a solution to (4.5).

6. Further examples

The strategy extends beyond the block structure of the previous examples, or equivalently to settings with b blocks of size one or a single block of size n.

Example 8. . Let $Y_1, \ldots, Y_n$ be exponentially distributed with means $\mathbb{E}(Y_i) = \lambda \exp(x_i^{\mathrm{T}}\psi)$ for covariates $x_i \in \mathbb{R}^p$. A simplified version of this example was studied by Cox and Reid (1987) from a different perspective. For an arbitrary pair (Y_i, Y_1), say:

$$T_\lambda(s,z) = \int_{-0}^{\infty}\int_{-0}^{\infty} \exp\{-zs(\lambda\exp(x_i^{\mathrm{T}}\psi)w_i, \lambda\exp(x_1^{\mathrm{T}}\psi)w_1)\}\exp\{-(w+v)\}dw_i dw_1,$$

where we have used a change of variable from y_i to $w_i = \lambda^{-1}\exp(-x_i^{\mathrm{T}}\psi)y_i$. The differential equation to be solved in $s(y_j, y_1)$ is

$$\begin{aligned} 0 &= \frac{\partial s\{y_i(w_i,\lambda,\psi), y_1(w_1,\lambda,\psi)\}}{\partial\lambda} \\ &= w_i(y_i,\psi,\lambda)\exp(x_i^{\mathrm{T}}\psi)\frac{\partial s(y_i,y_1)}{\partial y_i} + w_1(y_1,\psi,\lambda)\exp(x_1^{\mathrm{T}}\psi)\frac{\partial s(y_i,y_1)}{\partial y_1}, \end{aligned}$$

for which solutions are of the form $\{s(y_i, y_1) = (y_i/y_1)^k : k \neq 0\}$. Thus, inference based on $(Y_i/Y_1)_{i=1}^n$ is free of the nuisance parameter λ.

Example 9. Cox (1972, §11) discussed accelerated life models as convincing alternatives to proportional hazards representations in some contexts. One convenient parametric form, particularly when censoring is present, is the log-logistic model (e.g. Davison, 2003, p.190), which has a potentially non-monotonic hazard function. Let Y_i be a log-logistic random variable with shape parameter $\beta > 0$, determining monotonicity or otherwise, and scale parameter $\alpha_i = \lambda\exp(x_i^{\mathrm{T}}\psi)$ for $\lambda > 0$. The density function of Y_i is given by

$$\frac{(\beta/\alpha_i)(y/\alpha_i)^{\beta-1}}{(1+(y/\alpha_i)^\beta)^2}, \quad y > 0.$$

Solutions to the differential equation associated with an arbitrary pair (Y_i, Y_1) are of the form $\{s(y_i, y_1) = (y_i/y_1)^k : k \neq 0\}$, again showing that inference based on $(Y_i/Y_1)_{i=1}^n$ is free of the nuisance parameter λ.

7. Extending applicability

The previous analysis was limited by two aspects: that the marginal likelihood factorisation is not universally valid and indeed may only hold in rather special cases, such as models built on transformation groups; secondly, the requirement (4.3), which arises from the solution strategy for (4.2). It seems likely that applicability of the ideas of §4 can be extended, and we outline some possible routes to this.

7.1. **Perturbed factorisations.** Existence of a marginal likelihood factorisation can be relaxed by supposing that such a factorisation is approximately valid, with the factor purported to depend only on ψ in fact depending weakly on λ. If this more general condition is satisfied, λ can be replaced by an arbitrary value without materially affecting inference for ψ.

The relaxation can be incorporated by specifying that the right hand side of (4.5) is not exactly zero but rather a slowly varying function, h say, of λ. In the context of §4, the resulting equation is

$$\frac{\partial}{\partial\lambda}s\{t(w,\psi,\lambda),c(v,\psi,\lambda)\}=h(\lambda), \tag{7.1}$$

and since $h(\lambda)$ is unknown, one approach is to expand it locally around a base point, λ_0, leading to the approximation

$$\begin{aligned}\frac{\partial}{\partial\lambda}s\{t(w,\psi,\lambda),c(v,\psi,\lambda)\} &= a(t,c)\frac{\partial s(t,c)}{\partial t}+b(t,c)\frac{\partial s(t,c)}{\partial c}\\ &= \kappa+\varepsilon_1(\lambda-\lambda_0)+\varepsilon_2(\lambda-\lambda_0)^2,\end{aligned}$$

where κ is a constant, ε_1 and ε_2 are small and, as explained in §4, $a(t,c)=\partial t(w;\psi,\lambda)/\partial\lambda$ and $b(t,c)=\partial c(v;\psi,\lambda)/\partial\lambda$. This is now a first-order linear inhomogenous differential equation that can, in principle, be solved by the method of characteristics when κ, λ_0, ε_1 and ε_2 are treated as known. Provided that the limit of the resulting solution as $\varepsilon_1,\varepsilon_2\to 0$ is operational for small $(\lambda-\lambda_0)$, this specifies a transformation that is approximately factorisable in the earlier sense.

An alternative, simpler, route is to seek a solution to the original homogeneous equation $(\partial/\partial\lambda)s\{t(w,\psi,\lambda),c(v,\psi,\lambda)\}=0$, acknowledging that the transformation $s(t,c)$ will in general depend on the unknown parameters. If such a solution depends on ψ only, the distribution of $s(T,C;\psi)$ is free of λ at the true value of ψ. This leads to a procedure for constructing confidence sets for ψ analogous that proposed by Bartlett (1936) in the context of conditional likelihood. Specifically, any ψ_0 for which the sample of suitably standardised statistics $s(t,c;\psi_0)$ is consistent with its theoretical distribution, assuming $\psi_0=\psi$, constitutes a confidence set for ψ.

In reducing or eliminating dependence on the nuisance parameter, it is possible that dependence on the interest parameter is also weakened to the extent that marginal likelihood is ineffective. This is a limitation of the procedure which needs to be checked for.

Example 10. Let T_i and C_i for $i=1,\ldots,b$ be independently exponentially distributed of means $\mu_i-\psi$ and $\mu_i+\psi$ respectively, or equivalently of rates $\lambda_i/(1-\lambda_i\psi)$

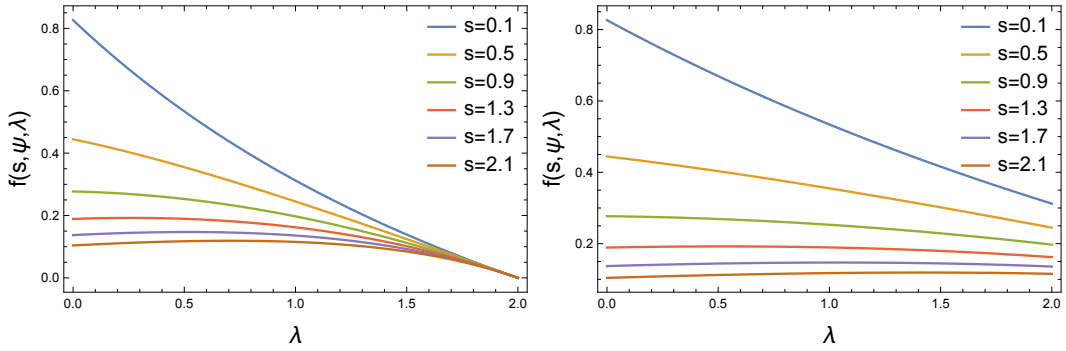

FIGURE 1. Plot of equation (7.2) as a function of λ for $\psi = 0.5$ (left), $\psi = 0.25$ (right), and for the specified values of s.

and $\lambda_i/(1+\lambda_i\psi)$, where $-1 < \lambda_i\psi < 1$. There is no obvious choice of partial likelihood function.

Dropping subscripts, the appropriate change of variables is to $w = t\lambda/(1-\lambda\psi)$ and $v = c\lambda/(1+\lambda\psi)$, therefore

$$\begin{aligned} a(t,c) &:= \frac{\partial t(w;\psi,\lambda)}{\partial\lambda} = -\frac{w}{\lambda^2} = \frac{t}{\lambda(1-\lambda\psi)}, \\ b(t,c) &:= \frac{\partial c(v;\psi,\lambda)}{\partial\lambda} = -\frac{v}{\lambda^2} = \frac{c}{\lambda(1+\lambda\psi)}. \end{aligned}$$

A set of solutions to the resulting homogeneous equation,

$$\frac{t}{\lambda(1-\lambda\psi)}\frac{\partial s(t,c)}{\partial t} + \frac{c}{\lambda(1+\lambda\psi)}\frac{\partial s(t,c)}{\partial c} = 0,$$

is $\{s(t,c) = (c/t^\beta)^k : k \neq 0\}$, where $\beta = (1-\lambda\psi)/(1+\lambda\psi)$, and it is convenient to take $k = 1$. The transformation depends on the unknown parameters λ and ψ, and T/C provides a crude estimate of β so that $C/T^{T/C}$ is one candidate transformation.

Instead consider a leading order Taylor series expansion around $\psi\lambda = 0$. This gives $s(t,c) \approx \tilde{s}(t,c) = c/t$. Thus $\tilde{s}(t,c) = c/t$ is the parameter-free transformation whose marginal density function depends only weakly on λ provided that $\psi\lambda$ is small. The latter condition is not unreasonable because the formulation requires $-1 < \psi\lambda < 1$.

However, direct calculation shows that the density function of $\tilde{S} = C/T$ is determined by the product $\psi\lambda$ as

$$f_{\tilde{S}}(s;\psi,\lambda) = \frac{1-(\psi\lambda)^2}{\{1+s+\psi\lambda(1-s)\}^2}, \quad s > 0. \tag{7.2}$$

Thus, although the transformation has been effective in approximately eliminating dependence on λ for any given ψ, as illustrated in Figure 1, a partial likelihood function based on (7.2) caries information only about the product $\psi\lambda$. We have found instead an approximately ancillary statistic for (ψ,λ). Further discussion of this example is in §8.

7.2. **Other solution strategies for the integro-differential equation** (4.2)**.** A further limitation of the proposal in §4 is the requirement (4.3). This arises from the solution strategy for (4.2) in terms of a partial differential equation. In developing the ideas from §4, we considered several other solution strategies. While we were unable to fully operationalise them, it seems valuable to report them here as possible routes for further exploration. We illustrate these ideas in the context of Example 4.

7.2.1. *Change of variables after differentiation under the integral sign.* A re-expression of the second equation in (5.1), in which the differentiation has been performed under the integral sign, is

$$0 = \int_0^\infty \exp(-\lambda c/\psi) \left[\int_0^\infty \exp\{-zs(t,c)\}\{2\lambda - \lambda^2(\psi t + c/\psi)\} \exp(-\lambda\psi t) dt \right] dc. \tag{7.3}$$

Thus any $s(t,c)$ that makes the function in square brackets orthogonal to $\exp(-\lambda c/\psi)$ identically in λ, ψ and z, also solves (5.1).

By changing variables to $x = \sqrt{c}$, (7.3) is

$$0 = \int_{-\infty}^\infty \frac{\exp(-\lambda x^2/\psi)}{2x} \left[\int_0^\infty \exp\{-zs(t,x^2)\}\{2\lambda - \lambda^2(\psi t + x^2/\psi)\} \exp(-\lambda\psi t) dt \right] dx.$$

and any $s(t,x^2)$ that makes the term in square brackets an even function, identically in z, ψ and λ satisfies the equation, because $\exp(-\lambda x^2/\psi)$ is an even function and $1/2x$ is odd.

7.2.2. *Hilbert-Schmidt orthogonalisation after differentiation under the integral sign.* From any linearly independent functions $f_0, \ldots, f_K$, orthogonal functions $\phi_0, \ldots, \phi_K$ can be constructed as described by Szëgo (1967, Chapter II) or Whittaker and Watson (1965, §11.6). Thus, in view of (7.3), set

$$\phi_0(c) = f_0(c) = \exp(-\lambda c/\psi).$$

For any $\kappa \neq 1$, $f_1(c) = \exp(-\lambda\kappa c/\psi) := f_1^{(\kappa)}(c)$ is linearly independent of f_0, and a family of functions $\phi_1^{(\kappa)}$, orthogonal to ϕ_0 can be constructed as (Szëgo 1967, Chapter II)

$$\phi_1^{(\kappa)}(c) = (D_0 D_1^{(\kappa)})^{-1/2} D_1^{(\kappa)}(c), \tag{7.4}$$

where $D_0 := \langle f_0, f_0 \rangle$, and

$$\begin{aligned} D_1^{(\kappa)}(c) &:= \langle f_0, f_0 \rangle f_1^{(\kappa)}(c) - \langle f_0, f_1^{(\kappa)} \rangle f_0(c), \\ D_1^{(\kappa)} &:= \langle f_0, f_0 \rangle \langle f_1^{(\kappa)}, f_1^{(\kappa)} \rangle - \langle f_0, f_1^{(\kappa)} \rangle^2, \end{aligned}$$

and for any pair of functions f and g, each in $\mathbb{L}_2(a,b)$, $\langle f, g \rangle = \int_a^b f(x)g(x)dx$.

For any $\kappa > 0$, with the f_0 and $f_1^{(\kappa)}$ defined above,

$$\begin{aligned} \langle f_0, f_1^{(\kappa)} \rangle &= \int_0^\infty \exp(-\lambda c/\psi) \exp(-\lambda\kappa c/\psi) dc = \psi/(\lambda + \kappa\lambda), \\ \langle f_1^{(\kappa)}, f_1^{(\kappa)} \rangle &= \int_0^\infty \exp(-\lambda\kappa c/\psi) \exp(-\lambda\kappa c/\psi) dc = \psi/(2\kappa\lambda), \end{aligned}$$

and $(f_0, f_0) = \psi/(2\lambda)$. Thus

$$(D_0 D_1^{(\kappa)})^{-1/2} = \frac{2\sqrt{2}\kappa^{1/2}(\kappa+1)\lambda^{3/2}}{(\kappa-1)\psi^{3/2}},$$

and using equation (7.4),

$$\phi_1^{(\kappa)}(c) = \frac{\sqrt{2\kappa\lambda}(\kappa+1)}{(\kappa-1)\psi^{1/2}}\exp(-\lambda\kappa c/\psi) - \frac{2\sqrt{2\kappa\lambda}}{(\kappa-1)\psi^{1/2}}\exp(-\lambda c/\psi).$$

It can be checked by integration that $\int_0^\infty \phi_0(c)\phi_1^{(\kappa)}(c)dc = 0$. It follows that a family of solutions to (7.3) are the functions $s(t,c)$ that solve

$$\phi_1^{(\kappa)}(c) = \int_0^\infty \exp\{-zs(t,c)\}\{2\lambda - \lambda^2(\psi t + c/\psi)\}\exp(-\lambda\psi t)dt.$$

7.2.3. *Stein operators.* Consider operators $\mathcal{A}_Q$, which characterize a distribution Q in the sense that

$$\mathbb{E}_Q(\mathcal{A}_Q f)(X) = 0 \quad \forall f \in \mathcal{F} \iff X \sim Q, \tag{7.5}$$

where $\mathcal{F}$ is the space of smooth and bounded functions. Stein (1972) showed that for Q the standard normal distribution, the operator A_Q is $(\mathcal{A}_Q f)(x) = f'(x) - xf(x)$. For Q an exponential distribution of rate ρ, the corresponding $\mathcal{A}_Q$ is (Luk, 1994),

$$(\mathcal{A}_Q f)(x) = (1-\rho x)f'(x) + xf''(x). \tag{7.6}$$

Let $w(c;\psi,\lambda)$ denote the function in square brackets in (7.3). Since (7.5) is a complete characterisation of Q, any function w satisfying (7.3) also satisfies

$$(\psi/\lambda)w(c;\psi,\lambda) = (1-\lambda c/\psi)f'(c) + cf''(c)$$

for some smooth and bounded function f, and so the equation to be solved for $s(t,c)$, identically in z, λ and ψ, is

$$\begin{aligned}&(1-\lambda c/\psi)f'(c;z,\lambda,\psi) + cf''(c;z,\lambda,\psi)\\ =\ &\lambda\psi^{-1}\int_0^\infty \exp\{-zs(t,c)\}\{2\lambda - \lambda^2(\psi t + c/\psi)\}\exp(-\lambda\psi t)dt\end{aligned}$$

for any convenient choice of $f \in \mathcal{F}$.

8. Reducing the role of nuisance parameters through other routes

While complete elimination of nuisance parameters is typically highly effective when available, other routes to reducing their role may sometimes be more fruitful.

Rather than seeking a transformation whose distribution is free of the nuisance parameters, we may instead seek one whose expectation is free of them, a more modest goal, leading to empirical averages as point estimators. In principle this should be easier to operationalise in a systematic way than complete elimination from the distribution of the transformation, although we have not obtained a unifying formulation. The variance of this point estimator in general depends on the nuisance parameters, and the challenge is then to find an accurate estimate of the composite, thereby evading separate estimation of each one.

Example 11. Suppose that T_i and C_i are as in Example 10. Then

$$\hat{\psi} := \frac{1}{2b}\sum_{i=1}^{b}(C_i - T_i) \to_p \frac{1}{2b}\sum_{i=1}^{b}\mathbb{E}(C_i - T_i) = \psi$$

is a consistent point estimator as $n \to \infty$. The variance of $\hat{\psi}$, unsurprisingly, depends on all b nuisance parameters. In particular

$$\mathrm{var}(\hat{\psi}) = \frac{\psi^2}{2b} + \frac{1}{2b^2}\sum_{i=1}^{b}\frac{1}{\lambda_i^2}.$$

However, since

$$\frac{1}{b}\sum_{i=1}^{b}T_iC_i \to_p \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}(T_iC_i) = \frac{1}{b}\sum_{i=1}^{b}\frac{1}{\lambda_i^2} - \psi^2,$$

a consistent estimator of the variance of $\hat{\psi}$ is

$$\hat{\sigma}^2 := \frac{1}{2b^2}\sum_{i=1}^{b}T_iC_i + \frac{\hat{\psi}^2}{b} \to_p \frac{\psi^2}{2b} + \frac{1}{2b^2}\sum_{i=1}^{b}\frac{1}{\lambda_i^2}.$$

Although there are compelling reasons for preferring likelihood-ratio inference in low dimensions, Example 10 is a case for which direct use of the likelihood is not recommended, and for which partial likelihood may be infeasible, while a pivotal quantity $(\hat{\psi} - \psi)/\hat{\sigma}$ can be derived from simple algebraic operations. Following Wald (1950), confidence sets can in principle be constructed from this quantity, although its approximate normality requires investigation, as the usual regularity conditions do not hold.

In Example 11 the original nuisance parameters are neither estimated nor eliminated. Instead, an implicit reparametrisation is performed, producing a scalar composite nuisance parameter. Accumulation of estimation error from multitudinous nuisance parameters is thereby avoided. This points to a more general strategy, in principle applying even when $p > n$, in which transformations are sought to make the problem depend on the interest parameter and, at most, a small set of one-dimensional summaries of the original nuisance parameters.

9. Closing remarks

In an earlier related paper (Battey and Cox, 2020) we closed with some open problems having a differential geometrical bearing. The first of these questioned whether a connection could be established between data-based transformations for the elimination of nuisance parameters via marginal or conditional likelihood, and the interest-respecting reparameterisations of Cox and Reid (1987). The following remarks are informal.

Fraser (1964) made a connection between the sample and parameter spaces through the notion of a local location model. The parameterised distribution function $F_Y(y;\theta)$

of Y, say, is viewed a function of both y and the parameter θ. Let ε be a quantile defined by $F(y_\varepsilon(\theta);\theta) = \varepsilon$. Then

$$0 = \frac{\partial F(y_\varepsilon(\theta);\theta)}{\partial y_\varepsilon}\frac{\partial y_\varepsilon(\theta)}{\partial \theta} + \frac{\partial F(y_\varepsilon;\theta)}{\partial \theta},$$

and since ε is arbitrary

$$\frac{\partial y(\theta)}{\partial \theta} = -\frac{\partial F(y;\theta)/\partial\theta}{f(y;\theta)}.$$

The exposition in Fraser (1964) is different.

These ideas have been fruitfully employed in the development of the tangent exponential model, starting with Fraser (1988) and Fraser and Reid (1988). Davison and Reid (2022) provide summaries from a different perspective with more detailed accounts of the historical development.

We close with an acknowledgement of the limitations of this work. The strategy we sought to follow was to start from some examples for which we already knew the answer and to find a general theory that recovers those answers. Where the paper falls short is in its extension of this theory to other structures, raising two possibilities: either that we have not taken full advantage of our ideas, or that they are less general than we would hope. Which of these possibilities is true can only be ascertained after further attempts at development, possibly along the lines of §7.

Appendix A. Treating nuisance parameters as fixed or random

The next two paragraphs follow an exposition due to Lindsay (1980). Let $(\mathcal{Y}, \mathcal{A})$ denote the common Borel space for the independent random variables $Y_1, \ldots, Y_b$. For each i, Y_i has a parametric density function $f(y;\psi,\lambda_i)$ with respect to σ-finite measure on $(\mathcal{Y}, \mathcal{A})$, known up to ψ and λ_i, where (ψ, λ_i) is assumed to belong to the Cartesian product parameter space $\Theta = \Psi \times \Lambda$ so that ψ and λ_i are variation independent. Let $\prod_{i=1}^b (\psi, \lambda_i)$ denote the measure on $(\mathcal{Y}^b, \mathcal{A}^b)$ induced by $(Y_1, \ldots, Y_b)$.

A key consideration is whether the nuisance parameters $\lambda_1, \ldots, \lambda_b$ are treated as realisations of random variables or as fixed arbitrary constants. Lindsay (1980) formalised the distinction by introducing the mixture model

$$f(y;\psi,Q) = \int f(y;\psi,\lambda)dQ(\lambda),$$

say, whose induced measure on $(\mathcal{Y}^b, \mathcal{A}^b)$ is denoted by $(\psi, Q)^b$. Let $\delta(\lambda)$ be the probability measure assigning mass 1 to $\{\lambda\}$. Lindsay (1980) noted that the parameter space Θ can be embedded in

$$\Theta^* := \{(\psi, Q) : \psi \in \Psi, Q \in \mathcal{Q}\}$$

provided that the space of probability measures $\mathcal{Q}$ is sufficiently rich to include each of $\delta(\lambda_1), \ldots, \delta(\lambda_b)$. In practice, $\mathcal{Q}$ is typically taken as a parametric family, although Kiefer and Wolfowitz (1956) allowed an infinite-dimensional $\mathcal{Q}$ under some regularity conditions.

This raises important conceptual issues that go to the foundations of modelling. While, in principle, $\Theta \subseteq \Theta^*$ when $\mathcal{Q}$ is unrestricted, to treat $\lambda_1, \ldots, \lambda_b$ as being drawn from a non-degenerate distribution rather than as fixed unknown constants is an extra assumption whose implications may be minor or considerable depending on context. A more extreme version of essentially the same problem is the weighing machine example of Cox (1958a).

The random effects formulation, if it is to be believed, has the appealing feature that ψ and Q together capture stable aspects of the system and are therefore relevant for predicting future observations. For scientific understanding via ψ, lack of stability of the model with respect to the nuisance parameters $\lambda_1, \ldots, \lambda_b$ is immaterial except insofar as it might affect inference on ψ. For the matched comparison studies that motivate the present paper, whereby ψ represents a treatment effect and $\lambda_1, \ldots, \lambda_b$ capture block-specific effects, the conceptual motivation for treating these as fixed unknown constants seems strong and leads, at least in principle, to an answer free of assumptions about the inter-block variation.

Acknowledgements. It is a pleasure to thank the referees for their insights and suggestions. HSB was supported by a UK Engineering and Physical Sciences Research Council Research Fellowship (EP/T01864X/1). SHL's work on the paper was done while an undergraduate student at Imperial College London

Conflict of interest statement. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability statement. There are no data associated with this work.

References

1. Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.
2. Barndorff-Nielsen, O. E. (1990). Approximate interval probabilities. *J. Roy. Statist. Soc. Ser. B*, 52, 485–496.
3. Bartlett, M. S. (1936). Statistical information and properties of sufficiency. *Proc. Roy. Soc. A*, 154, 124–137.
4. Battey, H. S. and Cox, D. R. (2020). High dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry*, 3, 119–148.
5. Courant, R. and Hilbert, D. (1966). *Methods of Mathematical Physics, Volume II*. Interscience Publishers, New York.
6. Cox, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–372.
7. Cox, D. R. (1958b). Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.
8. Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc. B*, 49, 1–39.
9. Cox, D. R. and Wermuth, N. (1999). Likelihood factorizations for mixed discrete and continuous variables. *Scand. J. Statist.*, 25, 209–220.

10. Cox, D. R. (2000). Some remarks on lieklihood factorization. *IMS Lecture Note Series*, 36, 168–172.
11. Cox, D. R. and Hinkley, D.(1978). *Theoretical Statistics*. Chapman and Hall, London.
12. Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.
13. Davison, A. C. and Reid, N. (2022). The tangent exponential model. *arXiv: 2106.10496*.
14. Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Roy. Statist. Soc. Ser. B*, 35, 189–233.
15. Fraser, D. A. S. (1964). Local conditional sufficiency. *J. Roy. Statist. Soc. Ser. B*, 26, 52–62.
16. Fraser, D. A. S. (1968). *The Structure of Inference*. Wiley, New York.
17. Fraser, D. A. S. (1988). Normed likelihood as saddlepoint approximation. *J. Multivar. Anal.*, 27, 181–193.
18. Fraser, D. A. S. and Reid, N. (1988). On conditional inference for a real parameter: a differential approach on the sample space. *Biometrika*, 75, 251–264.
19. Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, 77, 65–76.
20. Kalbfleish, J. D. and Sprott, D. A. (1970). Applications of likelihood methods to problems involving a large number of nuisance parameters (with discussion). *J. Roy. Statist. Soc. Ser. B*, 32, 175–208.
21. Kartsonaki, C. and Cox, D. R. (2016). Some matched comparisons of two distributions of survival time. *Biometrika*, 103, 219–224.
22. Lindsay, B. G. (1980). Nuisance parameters, mixture models and the efficiency of partial likelihood estimators. *Phil. Trans. Roy. Soc. London*, 196, 639–665.
23. Lindsay, B. G. (1985). Using empirical partially Bayes inference for increased efficiency. *Ann. Statist.*, 13, 914–931.
24. Szëgo, G. (1967). *Orthogonal Polynomials*. American Mathematical Society, Providence.
25. Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.
26. Whittaker, E. T. and Watson, G. N. (1927). *A Course of Modern Analysis* (1965 reprint of the fourth edition). Cambridge University Press, London.

H. S. Battey (to whom correspondence should be addressed):
Department of Mathematics, Imperial College London, UK
Email address: h.battey@imperial.ac.uk

D. R. Cox († Nuffield College, University of Oxford, UK)

Su Hyeong Lee: Mathematical Institute, University of Oxford, UK
Email address: su.lee@spc.ox.ac.uk