Festschrift for Tarmo Pukkila on his 60th Birthday
Eds. E. P. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. P. H. Styan

Univ. of Tampere, 2006, ISBN 978-951-44-6620-5, pages 301–314

Calculating efficient semiparametric estimators for a broad class of missing-data problems

ALASTAIR J. SCOTT & CHRIS WILD

Abstract. We develop efficient methods for computing semiparametric estimates in a wide variety of situations involving missing data and response-selective sampling. The methods are based on the profile likelihood. Estimates of the covariance matrix and associated test statistics are obtained at the same time.

2000 MSC codes: 62G99, 62J99.

Key words and phrases: Missing data; Biased sampling; Profile likelihood.

1 Introduction

We consider a unified method for fitting essentially arbitrary regression models to a large class of missing data and/or response-selective sampling problems using semiparametric maximum likelihood. This paper gives the computational details underlying the profile likelihood methods used, for example, in Scott and Wild (1997), Lawless et al. (1999) and Neuhaus et al. (2002, 2006). However, it covers a much wider range of applications than the situations discussed in those papers.

Suppose that we have a finite population of N individuals under study. Let $\boldsymbol{v}$ represent a set of variables containing easily obtainable information that is available for all N individuals. For our development, $\boldsymbol{v}$ must have finite support, whereas all other variables may be either discrete or continuous. In addition to the information on $\boldsymbol{v}$, we assume that information on a (possibly multivariate) response variable $\boldsymbol{y}$, can be obtained for at least a subset of individuals in the study study, and that a more "expensive" set of explanatory variables $\boldsymbol{z}$ can also obtained for a (possibly different)subset. We may wish to use some of the variables in $\boldsymbol{v}$, say $\boldsymbol{v}_1$, as explanatory variables in our model; other variables in $\boldsymbol{v}$ can play the role of informative surrogates for expensive covariates in $\boldsymbol{z}$. The object is to estimate the parameters, $\boldsymbol{\theta}$, of the

regression model $f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{v}_1; \boldsymbol{\theta})$. Thus we have a parametric regression model for the conditional density of the response given explanatory variables. Our methods can handle situations in which data sources include observations on $\boldsymbol{z}$ and any or all of $(\boldsymbol{z}, \boldsymbol{v})$ $(\boldsymbol{z} \mid \boldsymbol{v})$, $(\boldsymbol{y}, \boldsymbol{v})$, $(\boldsymbol{y} \mid \boldsymbol{v})$, $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, $(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{v})$, $(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{v})$, and $(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{v})$, where "$\mid$" refers to information obtained from conditional sampling.

Let $g(\boldsymbol{z}, \boldsymbol{v})$ denote the density of the covariates and $g(\boldsymbol{z} \mid \boldsymbol{v})$ the conditional density of $\boldsymbol{z}$ given $\boldsymbol{v}$. With standard prospective sampling and no missing data, the likelihood factorises into a term involving $\boldsymbol{\theta}$ and a term involving $g()$ so that information on the distribution of the explanatory variables is orthogonal to information on $\boldsymbol{\theta}$. Consequently, we do not need to model the covariate distribution. This is very convenient in practice because we often have far too many covariates of different types for it to be feasible to model their joint distribution in any realistic fashion. Unfortunately, with response-selective sampling, as with most missing data mechanisms, information on $\boldsymbol{\theta}$ and $g()$ is no longer orthogonal and we are forced to use some sort of joint modelling. However, the practical need for methods which do not require parametric modelling of $g()$ is just as great. Thus, we consider semi-parametric methods in which the marginal distribution of $(\boldsymbol{z}, \boldsymbol{v})$ is left unspecified and estimated nonparametrically.

The class of likelihoods that we consider initially consists of all those of the form

$$\prod_{i=1}^{N} f(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \boldsymbol{v}_{1i}; \boldsymbol{\theta})^{\Delta_{1i}} g(\boldsymbol{z} \mid \boldsymbol{v}_i)^{\Delta_{2i}} f(\boldsymbol{y}_i \mid \boldsymbol{v}_i; \boldsymbol{\theta})^{\Delta_{3i}}, \tag{1}$$

where $f(\boldsymbol{y} \mid \boldsymbol{v}; \boldsymbol{\theta}) = \int f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{v}_1; \boldsymbol{\theta})\, dG(\boldsymbol{z} \mid \boldsymbol{v})$, Δ_{1i} and Δ_{2i} are binary indicators taking values 0 or 1, and Δ_{3i} can take values 0, 1 or -1. Examples where such forms arise in practice are given in the next section. Our estimator of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, is found by maximizing the *profile* likelihood, obtained by maximizing (1) over the (potentially infinite-dimensional) nuisance parameter $g(\boldsymbol{z} \mid \boldsymbol{v})$. Conditions under which this profile likelihood estimator has full semiparametric efficiency are given by Bickel et al. (1993) and Murphy and van der Vaart (2000) for the i.i.d. case. Simpler conditions that apply directly to our multi-sample likelihood (1) are given in Lee and Hirose (2006). In addition, a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be obtained from the inverse profile information matrix. This means that we can produce a single program to cater for all likelihoods in this class and, with minimal modification, a more extended class introduced later. This enables general software to be written whereby a new regression model can be catered for simply by coding a new function to calculate $f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{v}_1; \boldsymbol{\theta})$ and its derivatives.

Other semiparametric efficient approaches such as those of Robins et al. (1994, 1995) and Holcroft et al. (1997) require complicated modelling over and above finding a suitable model, $f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{v}_1; \boldsymbol{\theta})$, for the regression

of interest. Profile likelihood enables us to obtain standard errors from the inverse Hessian matrix of the likelihood profile without having to consider the intricacies of complex observational schemes that are necessary for the sandwich estimators of variance required by other estimating equation approaches. Analyses based upon the approach here require no modelling effort from the statistician on aspects relating to missingness patterns and thus result in analyses that are much more time-efficient for the statistician.

This paper is organized as follows. Section 2 gives examples showing how likelihoods of the form (1) arise and illustrates the generality of the form. Section 3 describes maximum likelihood estimation for the class described above and the results are extended to a wider class in Section 4. Section 5 describes a much more computationally-efficient specialization of the algorithm that can be used when $\boldsymbol{y}$ is discrete, or for continuous $\boldsymbol{y}$ where we have (or retain) only class-interval information on $(\boldsymbol{y}, \boldsymbol{v})$ when $\boldsymbol{z}$ is missing as in Lawless et al. (1999).

2 Examples

We begin with some examples to illustrate how likelihoods of the form (1) can arise in practice.

Example 1. Case-control designs and generalisations

For case-control studies, y is a binary response variable recording case status as $y = 1$ or control status as $y = 0$. As a typical example, the cases may be those individuals that have contracted a disease of interest and the controls are those that have not.

In a simple case-control study, a random sample of m_1 cases is taken and their covariate values $\boldsymbol{z}$ are subsequently ascertained. The same is done for a sample of m_0 controls. The resulting likelihood is

$$\prod_{\text{cases}} f(\boldsymbol{z} \mid y = 1) \prod_{\text{controls}} f(\boldsymbol{z} \mid y = 0) = \prod_{\text{sample}} f(y \mid \boldsymbol{z}; \boldsymbol{\theta}) g(\boldsymbol{z}) f(y; \boldsymbol{\theta})^{-1}, \quad (2)$$

from Bayes' theorem. This is a very simple example of (1) in which a notional $\boldsymbol{v}$ takes on a single value, $\Delta_1 = \Delta_2 = 1$ if $\boldsymbol{z}$ is observed and 0 otherwise, and $\Delta_3 = -1$ if $\boldsymbol{z}$ is observed and 0 otherwise.

In a population-based study or a two-stage study in which case-control status is ascertained for all population units at the first stage and $\boldsymbol{z}$ is measured for a sample of cases and a sample of controls at the second stage, the likelihood is (Scott and Wild (1991))

$$\Big\{ \prod_{\text{cases}} f(\boldsymbol{z} \mid y = 1) \prod_{\text{controls}} f(\boldsymbol{z} \mid y = 0) \Big\} f(y = 1)^{M_1} f(y = 0)^{M_0} = \prod_{i=1}^{N} \{ f(y \mid \boldsymbol{z}; \boldsymbol{\theta}) \}^{\Delta_1} g(\boldsymbol{z})^{\Delta_2} f(y; \boldsymbol{\theta})^{\Delta_3}, \quad (3)$$

where M_1 is the number of cases and M_0 the number of controls in the population or first stage, $\Delta_1 = \Delta_2 = 1$ if $\boldsymbol{z}$ is observed and 0 otherwise, and $\Delta_3 = 0$ if $\boldsymbol{z}$ is observed and $+1$ otherwise. We note that we get the same likelihood whether we take fixed sized samples of cases and controls at the second stage or use a random mechanism as in the missing-data example. However, sandwich estimators of the variance of $\hat{\boldsymbol{\theta}}$ will be different in general although, for the special case of logistic regression, only the variance of the intercept is affected.

With stratified case-control data, as described by Scott and Wild (1997), we have data on $(\boldsymbol{y}, \boldsymbol{v})$ at the first stage and then obtain $\boldsymbol{z}$ for subsamples of cases and controls drawn from strata, $S_1, \ldots, S_L$ say, defined by values of $\boldsymbol{v}$. The likelihood is of the form

$$\prod_{\ell=1}^{L} \prod_{\boldsymbol{v}_i \in S_\ell} \{f(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \boldsymbol{v}_{1i}; \boldsymbol{\theta})\}^{\Delta_1} g(\boldsymbol{z} \mid \boldsymbol{v}_i)^{\Delta_2} f(\boldsymbol{y}_i \mid \boldsymbol{v}_i; \boldsymbol{\theta})^{\Delta_3}, \tag{4}$$

where the Δ_js are as for (3). The same likelihood pertains whether the data is obtained purposefully in this way using subsamples of fixed or random size, or whether we have population $\boldsymbol{y}$ and $\boldsymbol{v}$ data available for the finite population from which the cases and controls were drawn in a simple case-control study. Lawless et al. (1999) discuss a wide variety of other examples and sampling schemes that produce likelihoods of the form (4).

Example 2. Two and three stage missing-at-random

Consider a three stage mechanism in which, at Stage 1, $\boldsymbol{v}$ is observed for a random sample of individuals. At the second stage $\boldsymbol{y}$ is observed for subset of individuals with the ith individual having probability $\pi_1(\boldsymbol{v}_i)$ of being included. At Stage 3, $\boldsymbol{z}$ is observed for a subset of the Stage 2 sample with the ith individual having probability $\pi_2(\boldsymbol{v}_i, \boldsymbol{y}_i)$ of being included in the third stage. Thus, we only have complete $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$ information for individuals sampled at Stage 3.

This is a special case of the missing data mechanism considered by Holcroft et al. (1997), which encompasses the mechanisms in Robins et al. (1994, 1995), for a designed study in which π_1 and π_2 would be known. More generally, Holcroft et al. (1997) also allow for missingness by happenstance in which the available data is of the form $\boldsymbol{v}$, $(\boldsymbol{y}, \boldsymbol{v})$ and $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$. An assumption that data is missing at random means, in general terms, that the probability that something is missing depends only upon data that is observed, e.g., the probability that $\boldsymbol{z}$ is missing from $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$ depends only $(\boldsymbol{y}, \boldsymbol{v})$. Thus, for a missing-at-random analysis, we will still have the same set up as above but with $\pi_1(\,)$ and $\pi_2(\,)$ being unknown. When data is missing at random these probabilities can, however, be estimated from the data, using a binary

regression model applied to an observed/missing response. This results in the use of missingness models $\pi_1(\ ;\alpha_1)$ and $\pi_2(\ ;\alpha_2)$ with their own sets of parameters.

Let us write $\tilde{\Delta}_j = 1$ when an observation is included in the $(j+1)$th stage of observation and 0 otherwise. The likelihood for the above scenario is:

$$\begin{aligned}
&\prod [g(\boldsymbol{v})\{1-\pi_1(\boldsymbol{v})\}]^{1-\tilde{\Delta}_1} [g(\boldsymbol{v})\pi_1(\boldsymbol{v}) f(\boldsymbol{y}\mid\boldsymbol{v})\{1-\pi_2(\boldsymbol{y},\boldsymbol{v})\}]^{\tilde{\Delta}_1(1-\tilde{\Delta}_2)} \\
&\qquad\qquad \cdot [g(\boldsymbol{v})\pi_1(\boldsymbol{v}) f(\boldsymbol{y}\mid\boldsymbol{v})\pi_2(\boldsymbol{y},\boldsymbol{v}) f(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{v})]^{\tilde{\Delta}_1\tilde{\Delta}_2} \\
&= \left\{\prod g(\boldsymbol{v})^{1-\tilde{\Delta}_1} f(\boldsymbol{y},\boldsymbol{v})^{\tilde{\Delta}_1(1-\tilde{\Delta}_2)} f(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})^{\tilde{\Delta}_1\tilde{\Delta}_2}\right\} \times \mathrm{Term}(\pi_1,\pi_2) \\
&= \left[\prod \{f(\boldsymbol{y}\mid\boldsymbol{z},\boldsymbol{v};\boldsymbol{\theta}) g(\boldsymbol{z}\mid\boldsymbol{v})\}^{\tilde{\Delta}_1\tilde{\Delta}_2} f(\boldsymbol{y}\mid\boldsymbol{v}_1;\boldsymbol{\theta})^{\tilde{\Delta}_1(1-\tilde{\Delta}_2)}\right] \\
&\qquad\qquad \times \left\{\prod g(\boldsymbol{v})\right\} \times \mathrm{Term}(\pi_1,\pi_2), \qquad (5)
\end{aligned}$$

where $f(\boldsymbol{y}\mid\boldsymbol{v};\boldsymbol{\theta}) = \int f(\boldsymbol{y}\mid\boldsymbol{z},\boldsymbol{v}_1;\boldsymbol{\theta})\, dG_1(\boldsymbol{z}\mid\boldsymbol{v})$. There are several things to note from the form of the likelihood (5). Most importantly, we see that $g(\boldsymbol{v})$ and any missingness model parameters are orthogonal to $\boldsymbol{\theta}$ and the conditional densities $g(\boldsymbol{z}\mid\boldsymbol{v})$ so likelihood-based inferences about $\boldsymbol{\theta}$ use only the first term of (5). This means that: (i) we do not have to construct missingness models and estimate their parameters; (ii) it is only the conditional densities $g(\boldsymbol{z}\mid\boldsymbol{v})$ that we need to estimate nonparametrically; (iii) all of the information on $\boldsymbol{\theta}$ is in the second and third stage data – there is no information on $\boldsymbol{\theta}$ in the data from those individuals for whom we observe $\boldsymbol{v}$ alone; (iv) the likelihood that we end up using is identical to that from a two-stage study in which we obtain data on $(\boldsymbol{y},\boldsymbol{v})$ for all individuals sampled at the first stage and then observe $\boldsymbol{z}$ for a second-stage subsample which individuals enter with probabilities $\pi_2(\boldsymbol{y},\boldsymbol{v})$; and (v) the likelihood we use is the same whether $\boldsymbol{v}$ is random, having been sampled at the first stage, or whether sampling is conditional upon $\boldsymbol{v}$.

We reiterate that only the first term of (1) is relevant for estimation of $\boldsymbol{\theta}$. When $\boldsymbol{v}$ is discrete, which we assume throughout, the first term of (5) is clearly equivalent to (1) with $\Delta_1 = \Delta_2 = \tilde{\Delta}_1\tilde{\Delta}_2$ and $\Delta_3 = \tilde{\Delta}_1(1-\tilde{\Delta}_2)$.

In addition to the examples above, the likelihood (1) allows us to include information that can be treated as independent sampled from any of the following distributions: $f(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})$, $f(\boldsymbol{y},\boldsymbol{z}\mid\boldsymbol{v})$, $f(\boldsymbol{y}\mid\boldsymbol{z},\boldsymbol{v})$, $f(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{v})$, $f(\boldsymbol{y},\boldsymbol{v})$, $f(\boldsymbol{y}\mid\boldsymbol{v})$, $f(\boldsymbol{z},\boldsymbol{v})$, and $f(\boldsymbol{z}\mid\boldsymbol{v})$. Examples of such schemes can be found in Scott and Wild (2001), Neuhaus et al. (2002, 2006) and Lee et al. (2006). This gives a great deal of scope for including supplementary information to cope with the lack of indentifiablity and the ill-conditioning problems that can occur with fitting prospective regressions to retrospectively sampled data.

3 Profile likelihood

3.1 Preliminaries

Recall that $\boldsymbol{v}$ has finite support. Denote the distinct values of $\boldsymbol{v}$ in the observed population by $\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_S$ and let S_s be the stratum containing all units with $\boldsymbol{v}_i = \tilde{\boldsymbol{v}}_s$. Then (1) can be written in the form

$$L[\boldsymbol{\theta}, \{g(\cdot \mid \boldsymbol{v}_s)\}] = \prod_{s=1}^{S} \prod_{i \in S_s} f(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \tilde{\boldsymbol{v}}_{1s}; \boldsymbol{\theta})^{\Delta_{1i}} g(\boldsymbol{z}_i \mid \tilde{\boldsymbol{v}}_s)^{\Delta_{2i}} f(\boldsymbol{y}_i \mid \tilde{\boldsymbol{v}}_s; \boldsymbol{\theta})^{\Delta_{3i}}. \tag{6}$$

We wish to maximize this as a function of $\boldsymbol{\theta}$ and the S conditional densities $g(\boldsymbol{z} \mid \tilde{\boldsymbol{v}}_s)$. As is standard in semiparametric maximum likelihood, we treat these densities as discrete with all of the mass being placed at the observed $\boldsymbol{z}$ values.

For simplicity, we write $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{v}_1)$ to include all covariates that are to be used in the model. Note that, when we consider only the sizes of probability atoms, $g(\boldsymbol{x} \mid \boldsymbol{v}) = g(\boldsymbol{z}, \boldsymbol{v}_1 \mid \boldsymbol{v}) = g(\boldsymbol{z} \mid \boldsymbol{v})$. The only remaining role of $\boldsymbol{v}$ in (6) is to divide the data set up into the S strata with separate conditional distributions of $\boldsymbol{x}$ to be estimated for each stratum.

Our problem is now to maximize

$$\ell(\boldsymbol{\theta}, g_1, \ldots, g_S) = \prod_{s=1}^{S} \prod_{i \in S_s} f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})^{\Delta_{1i}} g(\boldsymbol{x}_i \mid S_s)^{\Delta_{2i}} f(\boldsymbol{y}_i \mid S_s; \boldsymbol{\theta})^{\Delta_{3i}},$$

with $f(\boldsymbol{y} \mid S_s; \boldsymbol{\theta}) = \int f(\boldsymbol{y} \mid \boldsymbol{z}, \tilde{\boldsymbol{v}}_{s1}; \boldsymbol{\theta})\, dG(\boldsymbol{z} \mid \tilde{\boldsymbol{v}}_s)$. The corresponding log-likelihood is of the form $\ell(\boldsymbol{\theta}, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_S) = \sum_s \ell_s(\boldsymbol{\theta}, \boldsymbol{g}_s)$, where $\boldsymbol{g}_s = \boldsymbol{g}(\boldsymbol{x} \mid S_s)$ Thus, the profile log-likelihood, in which we maximise out the $\boldsymbol{g}_s$'s for fixed $\boldsymbol{\theta}$, is of the form

$$\ell_P(\boldsymbol{\theta}) = \ell\{\boldsymbol{\theta}, \hat{\boldsymbol{g}}_1(\cdot\,; \boldsymbol{\theta}), \ldots, \hat{\boldsymbol{g}}_S(\cdot\,; \boldsymbol{\theta})\} = \sum_s \ell_s\{\boldsymbol{\theta}, \hat{\boldsymbol{g}}_s(\cdot\,; \boldsymbol{\theta})\} = \sum_s \ell_{Ps}(\boldsymbol{\theta}).$$

This means that we need only to be able to solve the problem of obtaining the profile for $\boldsymbol{\theta}$ (and its derivatives) for a single stratum, i.e. to maximize

$$L(\boldsymbol{\theta}, \boldsymbol{g}) = \prod_i f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})^{\Delta_{1i}} g(\boldsymbol{x}_i)^{\Delta_{2i}} f(\boldsymbol{y}_i; \boldsymbol{\theta})^{\Delta_{3i}}. \tag{7}$$

In terms of implementation in software, we need to write a function to find $\ell_P(\boldsymbol{\theta}) = \sup_{\boldsymbol{g}}\{\log L(\boldsymbol{\theta}, \boldsymbol{g})\}$ where $L(\boldsymbol{\theta}, \boldsymbol{g})$ is given by (7). When we have multiple strata, we can send the data from each stratum in turn to that function and accumulate the results.

3.2 Basic profile likelihood algorithm

In essence, we will let $\delta_i = g(\boldsymbol{x}_i)$ and work with $\ell(\boldsymbol{\theta}, \boldsymbol{\delta})$ with $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots)$ as if it were an ordinary parametric log-likelihood. One of the problems with

this, however, is that it often results in working with very large arrays. It is particularly important to keep the dimension of $\boldsymbol{\delta}$ as small as possible. Trapping replicate data points is one way of reducing the size of these arrays. Thus, we will write in terms of replicated data. Whether or not we do this makes no difference at all to the profile that we obtain for $\boldsymbol{\theta}$, but it can substantially reduce storage.

Let $A = \{\, i : \Delta_{1i} = 1 \,\}$, $B = \{\, i : \Delta_{2i} = 1 \,\}$, and $C = \{\, i : \Delta_{3i} \neq 0 \,\}$. Let $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_J$ be the distinct values of $\boldsymbol{x}$ in B, m_j be the multiplicity of $\tilde{\boldsymbol{x}}_j$ in B and $\delta_j = g(\tilde{\boldsymbol{x}}_j)$. Let $\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_K$ be the distinct values of $\boldsymbol{y}$ in C and let $r_k = \sum_{\{i:\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_k\}} \Delta_{3i}$. Note that r_k can be positive, negative or zero. On noting that $f(\mathcal{Y}) = \sum f(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_j)\delta_j$, the log-likelihood from (7) can be written in the form

$$\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_A \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) + \sum_j m_j \log \delta_j + \sum_k r_k \log\Big\{ \sum_{j=1}^{J} f(\tilde{\boldsymbol{y}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})\delta_j \Big\}. \tag{8}$$

Since the δ_j parameters have to satisfy the constraints $0 < \delta_j < 1$ and $\sum \delta_j = 1$, we reparameterize in terms of $\rho_j = \log(\delta_j/\delta_J)$ and work with $\ell(\boldsymbol{\theta}, \boldsymbol{\rho})$. With this parametrization $\delta_j = \exp(\rho_j)/\sum \exp(\rho_\ell)$ (with $\rho_J \equiv 0$) and the constraints are satisfied automatically. The profile log-likelihood is then $\ell_P(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\rho}}(\boldsymbol{\theta}))$, where $\hat{\boldsymbol{\rho}}(\boldsymbol{\theta})$ satisfies $\partial \ell(\boldsymbol{\theta}, \boldsymbol{\rho})/\partial \boldsymbol{\rho} = \mathbf{0}$, and ℓ_P has profile score vector

$$\boldsymbol{U}_P(\boldsymbol{\theta}) = \frac{\partial \ell_P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\rho})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\rho} = \hat{\boldsymbol{\rho}}(\boldsymbol{\theta})}$$

(see Seber and Wild (1989), equation (2.69)) and observed profile information matrix

$$\boldsymbol{\mathcal{I}}_P = \left(-\frac{\partial^2 \ell_P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathsf{T}}} \right) = \{ \boldsymbol{\mathcal{J}}_{\theta\theta} - \boldsymbol{\mathcal{J}}_{\theta\rho} \boldsymbol{\mathcal{J}}_{\rho\rho}^{-1} \boldsymbol{\mathcal{J}}_{\theta\rho}^{\mathsf{T}} \}\Big|_{\boldsymbol{\rho} = \hat{\boldsymbol{\rho}}(\boldsymbol{\theta})} \tag{9}$$

written in terms of the blocks of $\boldsymbol{\mathcal{J}}$, the observed information matrix obtained from $\ell(\boldsymbol{\theta}, \boldsymbol{\rho})$ (see Seber and Wild (1989), just prior to equation (2.72)).

We apply a Newton–Raphson based algorithm, with updating steps $\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \boldsymbol{\mathcal{I}}_P^{-1} \boldsymbol{U}_P \,|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{(a)}}$ for $a = 1, 2, \ldots$ to maximize the profile log-likelihood $\ell_P(\boldsymbol{\theta})$. At each iteration we solve for the accompanying $\hat{\boldsymbol{\rho}}(\boldsymbol{\theta}^{(a+1)})$ also by using Newton–Raphson to maximise $\ell(\boldsymbol{\theta}^{(a+1)}, \boldsymbol{\rho})$ with respect to $\boldsymbol{\rho}$. The derivatives required for all of this are given in Appendix A. Our "Newton–Raphson based algorithm" employs simple versions of the hill-climbing techniques discussed in Section 13.3.1 of Seber and Wild (1989) to improve robustness. We have been surprised by how stable and reliable the method has turned out to be in practice, despite the often very large dimensions of the parameter vectors involved.

A clear disadvantage with the algorithm discussed in this section is that the dimension of the parameter vectors can be very large indeed (equal to the sample size n in the worst case), requiring substantial storage. For example, using Newton–Raphson for the maximization requires that we store $\boldsymbol{\mathcal{J}}_{\rho\rho}$. This is also necessary if we wish to calculate the profile information matrix to obtain variance estimates for $\hat{\boldsymbol{\theta}}$. We will find in the next section that we can obtain a substantial reduction in dimensionality, and consequent increase in speed, when $\boldsymbol{y}$ is discrete. In particular, for binary regression models we can work with a nuisance parameter of dimension 1 rather than $\boldsymbol{\rho}$, which has dimension $J-1$ where J is the number of distinct values observed for $\boldsymbol{x}$. The same sort of reduction can be made for continuous $\boldsymbol{y}$ in situations where only class interval information on $\boldsymbol{y}$ is available for those data points for which $\boldsymbol{x}$ is not fully observed.

Before going on, however, we do note that when we have several strata, the method discussed in the current subsection never needs to store a $\boldsymbol{\mathcal{J}}_{\rho\rho}$ matrix for more than one stratum. These matrices are used to find the contribution of the current stratum to the overall information matrix $\boldsymbol{\mathcal{I}}_P$ and then discarded when we move on to process the data from the next stratum; see the discussion surrounding (7). This makes it feasible to handle reasonably large data sets. For example, Jiang (2004) was able to run simulations fitting linear models to two-phase samples with strata of size $N = 5000$ and sub-sample sizes of $n = 1000$ using these methods.

4 An extension to the class of likelihoods

There are important extensions of the likelihood class (1) which do not affect the essential nature of the maximization problem. (We have postponed consideration of these in the interests of intelligibility.) For example, Lawless et al. (1999) discuss failure time data in which whether of not a data point is fully observed depends on membership of strata defined in terms of both $\boldsymbol{y}$ and $\boldsymbol{x}$. They also used strata involving both $\boldsymbol{y}$ and $\boldsymbol{x}$ to avoid the problem of empty or near-empty strata. Neuhaus et al. (2002) deals with retrospectively sampled family data. Here, $\boldsymbol{y}$ records the set of binary responses for each member of a cluster (or family) and sampling is conditional upon observation of some specified pattern in the responses from a cluster. This data can also be supplemented in various ways, for example by knowledge of stratum sizes in the finite population from which the individuals were sampled.

We can expand (1) to cater for such examples as follows. Let $\boldsymbol{h}_{\boldsymbol{v}}(\boldsymbol{y}, \boldsymbol{x})$ be a known function of the data, where we may use different functions for different $\boldsymbol{v}$. Imagine that the probability of being observed depends upon the value of $\boldsymbol{h}$. We may, for example, sample conditionally upon $\boldsymbol{h}$ values and then observe $(\boldsymbol{y}, \boldsymbol{x})$. We may also use a random mechanism by which $\boldsymbol{h}$ values are obtained according to some probability, $\pi_4(\boldsymbol{h})$ say, and then $(\boldsymbol{y}, \boldsymbol{x})$ are sampled obtained from the conditional distribution of $(\boldsymbol{y}, \boldsymbol{x})$ given $\boldsymbol{h}$.

We may supplement such data with finite population data, or data from a random sample of $\boldsymbol{h}$ values. Alternatively, we may may have data produced by a random process, observe the $\boldsymbol{h}$ values as they arise and then further observe $(\boldsymbol{y}, \boldsymbol{x})$ with probability $\pi_5(\boldsymbol{h})$. In all of these situations, the likelihood is of the form,

$$\prod_{s=1}^{S} \prod_{i:\boldsymbol{v}_i=\boldsymbol{v}_s} f(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \boldsymbol{v}_{1s}; \boldsymbol{\theta})^{\Delta_{1i}} g(\boldsymbol{z}_i \mid \boldsymbol{v}_s)^{\Delta_{2i}} f\{\boldsymbol{h}_s^{[i]} \mid \boldsymbol{v}_s; \boldsymbol{\theta}\}^{\Delta_{3i}}, \tag{10}$$

where $\boldsymbol{h}_s^{[i]} = \boldsymbol{h}_{\boldsymbol{v}_s}(\boldsymbol{y}_i, \boldsymbol{x}_i)$, for suitably chosen Δ_{ji}s. This is a simple generalization of (6) and almost all the work in Section 3.2 applies directly. Again we need only consider the profiling problem for a single stratum, and (7) becomes

$$L(\boldsymbol{\theta}, g) = \prod f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})^{\Delta_{1i}} g(\boldsymbol{x}_i)^{\Delta_{2i}} f\{\boldsymbol{h}(\boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{h}^{[i]}; \boldsymbol{\theta}\}^{\Delta_{3i}}, \tag{11}$$

where $f\{\boldsymbol{h}; \boldsymbol{\theta}\} = \int_{\boldsymbol{h}(\boldsymbol{y},\boldsymbol{x})=\boldsymbol{h}} dF(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})\, dG(\boldsymbol{x})$. Suppose that $\tilde{\boldsymbol{h}}_k$ occurs with multiplicity r_k. Then (8), the form we use for computation, becomes

$$\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_A \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) + \sum_j m_j \log \delta_j + \sum_k r_k \log\Big\{\sum_j f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}) \delta_j\Big\}, \tag{12}$$

and $f\{\tilde{\boldsymbol{h}}_k \mid \boldsymbol{x}; \boldsymbol{\theta}\} = \int_{S_k(\boldsymbol{x})} dF(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ with $S_k(\boldsymbol{x}) = \{\, \boldsymbol{y} : \boldsymbol{h}(\boldsymbol{y}, \boldsymbol{x}) = \tilde{\boldsymbol{h}}_k \,\}$. The profile likelihood algorithm for the larger class of likelihoods in this subsection differs from that in Section 3.2 only in that $f\{\tilde{\boldsymbol{h}}_k \mid \boldsymbol{x}; \boldsymbol{\theta}\}$ replaces $f(\tilde{\boldsymbol{y}}_k \mid \boldsymbol{x}; \boldsymbol{\theta})$ in the last term of the log-likelihood. In computational terms, this means that to accommodate a new model $f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ when $\boldsymbol{h}(\boldsymbol{y}, \boldsymbol{x})$ is more complicated than simply $\boldsymbol{h}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{y}$, an additional function must be written to evaluate $f\{\boldsymbol{h} \mid \boldsymbol{x}; \boldsymbol{\theta}\}$ and its derivatives with respect to $\boldsymbol{\theta}$.

5 Exploiting discreteness

As we noted in Section 3.2, a big disadvantage of the brute-force approach outlined there is the dimension of the maximization problem. A substantial reduction of this dimension can be achieved in some important special cases, specifically whenever the $\boldsymbol{h}(\boldsymbol{y}, \boldsymbol{x})$ can take only a finite set of values, $(\tilde{\boldsymbol{h}}_1, \tilde{\boldsymbol{h}}_2, \ldots, \tilde{\boldsymbol{h}}_K)$ say.

If we maximise (12) with respect to $\boldsymbol{\delta}$ for fixed $\boldsymbol{\theta}$, using a Lagrange multiplier to cater for the constraint $\sum_j \delta_j = 1$, cf. Scott and Wild (1997, circa equation (A1.2)) we find that $\hat{\boldsymbol{\delta}}(\boldsymbol{\theta})$ satisfies the set of equations

$$\delta_j = m_j \Big/ \Big\{(m_+ + r_+) - \sum_k \frac{r_k f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}{\sum_{j'} f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_{j'}; \boldsymbol{\theta}) \delta_{j'}}\Big\}, \qquad j = 1, \ldots, J, \tag{13}$$

where $m_+ = \sum_j m_j$ is the total number of observations in B and $r_+ = \sum r_k$. Suppose that all possible values have been observed at least once so that $\sum_{k=1}^{K} f(\tilde{\boldsymbol{h}}_k \mid \boldsymbol{x}; \boldsymbol{\theta}) = 1$. If $r_k \neq 0$, set

$$Q_k = \sum_j f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}) \delta_j$$

and write

$$\tilde{p}_k = \tilde{p}_k(Q_k) = (m_+ + r_+) - \frac{r_k}{Q_k}. \tag{14}$$

Then the system (13) can be written in the form

$$\delta_j = \frac{m_j}{(m_+ + r_+) - \sum_k r_k \frac{f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}{Q_k}} = \frac{m_j}{\sum_k \tilde{p}_k f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}, \tag{15}$$

where the set of Q_k's corresponding to nonzero r_k satisfy the system

$$Q_k = \sum_j \frac{m_j f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}{(m_+ + r_+) - \sum_k \frac{m_k}{Q_k} f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})} = \sum_j \frac{m_j f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}{\sum_k \tilde{p}_k f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}. \tag{16}$$

The system of equations (16) is equivalent to the set of 'score' equations, $\partial \ell^* / \partial \boldsymbol{Q} = \boldsymbol{0}$ for fixed $\boldsymbol{\theta}$, where

$$\ell^*(\boldsymbol{\theta}, \boldsymbol{Q}) = \sum_A \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \sum_j m_j \log\Big\{\sum_k \tilde{p}_k(Q_k) f(\tilde{\boldsymbol{h}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})\Big\} + \sum r_k \log Q_k. \tag{17}$$

Thus we can either obtain the profile log-likelihood using $\ell_P(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\theta}))$, where $\boldsymbol{\delta}$ has dimension $(J-1)$, or using $\ell_P(\boldsymbol{\theta}) = \ell^*(\boldsymbol{\theta}, \hat{\boldsymbol{Q}}(\boldsymbol{\theta}))$, where $\boldsymbol{Q}$ has dimension $(K-1)$.

Further, and perhaps more importantly, it follows from Theorem 2.2 in Seber and Wild (1989) that

$$\frac{\partial^2 \ell_P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}} = \frac{\partial^2 \ell^*(\boldsymbol{\theta}, \boldsymbol{Q})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}} - \frac{\partial^2 \ell^*(\boldsymbol{\theta}, \boldsymbol{Q})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{Q}^\mathsf{T}} \left(\frac{\partial^2 \ell^*(\boldsymbol{\theta}, \boldsymbol{Q})}{\partial \boldsymbol{Q}\, \partial \boldsymbol{Q}^\mathsf{T}} \right)^{-1} \frac{\partial^2 \ell^*(\boldsymbol{\theta}, \boldsymbol{Q})}{\partial \boldsymbol{Q}\, \partial \boldsymbol{\theta}^\mathsf{T}}.$$

Recalling the standard form for the inverse of a partitioned matrix, it follows that the inverse profile information, $-(\partial^2 \ell_P(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T})^{-1}$ is equal to the leading $p \times p$ submatrix of $\boldsymbol{\mathcal{I}}^{*-1}$, where $\boldsymbol{\mathcal{I}}^* = \partial^2 \ell^*(\boldsymbol{\phi}) / \partial \boldsymbol{\phi}\, \partial \boldsymbol{\phi}^\mathsf{T}$, with $\boldsymbol{\phi} = \binom{\boldsymbol{\theta}}{\boldsymbol{Q}}$. This means that, for making inferences about $\boldsymbol{\theta}$, we can proceed as if $\ell^*(\boldsymbol{\phi})$ was the log-likelihood. We can obtain the semiparametric maximum likelihood estimator of $\boldsymbol{\theta}$ by setting $\partial \ell^*(\boldsymbol{\theta}, \boldsymbol{Q}) / \partial \boldsymbol{\phi} = 0$, we can estimate its covariance matrix with the appropriate block of $\boldsymbol{\mathcal{I}}^{*-1}$ and we can test hypotheses about $\boldsymbol{\theta}$ using appropriate differences in $-2\ell^*$. (This last statement needs some justification but the general case follows in exactly the same way as the special case treated in Scott and Wild (1989).)

Since K is almost always very much less than J, using $\ell^*(\boldsymbol{\theta}, \boldsymbol{Q})$ in place of $\ell(\boldsymbol{\theta}, \boldsymbol{\delta})$ can result in large reductions in storage requirements and in computing time. For example, in the important special case of fitting a binary regression model to case-control data, $K = 2$ and we have essentially reduced the problem from one involving a vector of n nuisance parameters to one involving a single scalar nuisance parameter.

We note that, although we can treat $\ell^*(\boldsymbol{\phi})$ as if it were a log-likelihood for many purposes, it is not in fact a true log-likelihood. In particular, ℓ^* typically has a minimum rather than a maximum in $\boldsymbol{Q}$ when the nonzero r_ks are all positive as is the case with missing data problems.

We have experimented with several reparameterizations of $\boldsymbol{Q}$ to take care of positivity constraints including $Q_k = \exp(\rho_k)/\{1 + \sum \exp(\rho_\ell)\}$ and

$$Q_k = \exp(\xi_k)/\{1 + \exp(\xi_k)\} \quad \text{or} \quad \xi_k = \text{logit}(Q_k).$$

The former, which also takes care of the summation constraint, leads to singular information matrices when two or more $r_k = 0$, whereas both parameterizations lead to the identical sets of derivatives when at most one $r_k = 0$. Thus, we routinely use $\ell^*(\boldsymbol{\theta}, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ contains only those ξ_k for which $r_k \neq 0$, for computation. Now,

$$\ell_P(\boldsymbol{\theta}) = \ell^*(\boldsymbol{\theta}, \hat{\boldsymbol{\xi}}(\boldsymbol{\theta})),$$

where $\hat{\boldsymbol{\xi}}(\boldsymbol{\theta})$ satisfies $\partial \ell(\boldsymbol{\theta}, \boldsymbol{\xi})^*/\partial \boldsymbol{\xi} = \mathbf{0}$. We then calculate the profile score vector and profile information matrix of $\ell_P(\boldsymbol{\theta})$ using

$$\boldsymbol{U}_P(\boldsymbol{\theta}) = \frac{\partial \ell^*(\boldsymbol{\theta}, \boldsymbol{\xi})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}(\boldsymbol{\theta})} \quad \text{and} \quad \boldsymbol{\mathcal{I}}_P = \{\boldsymbol{\mathcal{J}}^*_{\theta\theta} - \boldsymbol{\mathcal{J}}^*_{\theta\xi}\boldsymbol{\mathcal{J}}^{*-1}_{\xi\xi}\boldsymbol{\mathcal{J}}^{*\mathsf{T}}_{\theta\xi}\}\Big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}(\boldsymbol{\theta})}.$$

Expressions for the required derivatives are given in Appendix B.

6 Conclusion

The computational procedures outlined in this paper enable us to make efficient inferences about parameters of interest in a large number of situations involving missing data and response-biased sampling. In many of these situations it has simply not been feasible to implement efficient procedures before this and people have proposed a variety of *ad hoc* alternatives. Most of these alternatives should become obsolete when software to implement the methods described in this paper is available. Such software is currently under development and a prototype package is available from Chris Wild (c.wild@auckland.ac.nz) on request. A more polished version should soon be available on his web site at http://www.stat.auckland.ac.nz.

Appendix A. Derivatives of $\ell(\boldsymbol{\theta}, \boldsymbol{\rho})$

We give results for the general case described in Section 4. Equation (8) is the special case of equation (12) in which $\boldsymbol{h}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{y}$, and $f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}) = f(\tilde{\boldsymbol{y}}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})$ so that its derivatives with respect to $\boldsymbol{\theta}$ are simply the derivatives of the regression model $f()$. The user needs to specify

$$D_{kj}^{(\theta)} = \frac{\partial f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \text{and} \quad D_{kj}^{\theta\theta} = \frac{\partial D_{kj}^{(\theta)}}{\partial \boldsymbol{\theta}^{\mathsf{T}}}.$$

We repeat (12) more compactly as

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\delta}) &= \ell(\boldsymbol{\theta}, \boldsymbol{\rho}) \\ &= \sum_A \log f_i + \sum_j m_j \log \delta_j + \sum_k r_k \log\Big\{\sum_j f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}) \delta_j\Big\}, \end{aligned} \tag{18}$$

where, $f_i = f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})$ and $\delta_j = \exp(\rho_j) / \sum_\ell \exp(\rho_\ell)$ with $\rho_K \equiv 0$. After some manipulation, we find that

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\theta}} &= \sum_A \frac{\partial \log f_i}{\partial \boldsymbol{\theta}} + \sum_k r_k \frac{\boldsymbol{D}_k^{(\theta)}}{D_k}, \\ \frac{\partial \ell}{\partial \rho} &= \check{\boldsymbol{m}} + \sum_k r_k \check{\boldsymbol{R}}_k - (m_+ + r_+)\check{\boldsymbol{\delta}} \\ \boldsymbol{\mathcal{J}}^{(\theta\theta)} &= \sum_A \frac{\partial^2 \log f_i}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathsf{T}}} - \sum_k r_k \left\{ \frac{\boldsymbol{D}_k^{(\theta\theta)}}{D_k} - \left(\frac{\boldsymbol{D}_k^{(\theta)}}{D_k}\right)\left(\frac{\boldsymbol{D}_k^{(\theta)}}{D_k}\right)^{\mathsf{T}} \right\} \\ \boldsymbol{\mathcal{J}}^{(\theta\rho_j)} &= \sum_k r_k \left\{ \frac{Q_{kj}^* \delta_j}{D_k} \frac{\boldsymbol{D}_k^{(\theta)}}{D_k} - \frac{Q_{kj}^{*(\theta)} \delta_j}{D_k} \right\}, \qquad j = 1, \ldots, J-1 \\ \boldsymbol{\mathcal{J}}^{(\rho\rho)} &= (m_+ + r_+)\{\operatorname{diag}(\check{\boldsymbol{\delta}}) - \check{\boldsymbol{\delta}}\check{\boldsymbol{\delta}}^{\mathsf{T}}\} - \sum_k r_k \{\operatorname{diag}(\check{\boldsymbol{R}}_k) - \check{\boldsymbol{R}}_k \check{\boldsymbol{R}}_k^{\mathsf{T}}\} \end{aligned}$$

Here, $D_k = \sum_j f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}) \delta_j$, $\check{\boldsymbol{R}}_k$ is the vector with elements $D_{kj}^{(\theta)} \delta_j / D_k$, $j = 1, \ldots, J-1$, $\check{\boldsymbol{m}} = (m_1, \ldots, m_{J-1})$, $\check{\boldsymbol{\delta}} = (\delta_1, \ldots, \delta_{J-1})$, $\boldsymbol{D}_k^{(\theta)} = \sum_j D_{kj}^{(\theta)} \delta_j$, $\boldsymbol{D}_k^{(\theta\theta)} = \sum_j D_{kj}^{(\theta\theta)} \delta_j$, and $\sum_j$ denotes sums from 1 to J.

Appendix B. Derivatives of $\ell^*(\boldsymbol{\theta}, \xi)$

$$\begin{aligned} \ell^*(\boldsymbol{\theta}, \boldsymbol{\xi}) &= \ell^*(\boldsymbol{\theta}, \boldsymbol{Q}) \\ &= \sum_A \log f_i - \sum_j m_j \log\Big\{\sum_k \tilde{p}_k f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})\Big\} + \sum r_k \log Q_k, \end{aligned} \tag{19}$$

where $\tilde{p}_k = \tilde{p}_k(Q_k) = (m_+ + r_+) - r_k/Q_k$, and $\xi_k = \text{logit}(Q_k)$. We find that

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \sum_A \frac{\partial \log f_i}{\partial \boldsymbol{\theta}} - \sum_j m_j \frac{\partial}{\partial \boldsymbol{\theta}} \log\Big(\sum_k \tilde{p}_k f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})\Big),$$

$$\frac{\partial \ell^*}{\partial \xi_k} = r_k e^{-\xi_k}\left(Q_k - \sum_j \frac{m_j f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}{\sum_{k'} \tilde{p}_{k'} f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}\right), \qquad k = 1, \ldots, K,$$

$$\boldsymbol{\mathcal{J}}^{(\theta\theta)} = -\sum_A n_i \frac{\partial^2 \log f_i}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}} + \sum_j m_j \left[\frac{E_j^{*(\theta\theta)}}{E_j} - \left\{\frac{E_j^{*(\theta)}}{E_j}\right\}\left\{\frac{E_j^{*(\theta)}}{E_j}\right\}^\mathsf{T}\right],$$

$$\boldsymbol{\mathcal{J}}^{(\theta\xi_k)} = r_k e^{-\xi_k} \sum_j m_j \left[\frac{D_{kj}^{*(\theta)}}{E_j} - \left(\frac{f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})}{E_j}\right)\left\{\frac{E_j^{(\theta)}}{E_j}\right\}\right], \qquad k = 1, \ldots, K,$$

$$\boldsymbol{\mathcal{J}}^{(\xi\xi)} = \text{diag}()\{\text{diag}(\boldsymbol{Q}) - \boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\} - \sum_j m_j \{\text{diag}(\boldsymbol{F}_j) + \boldsymbol{F}_j \boldsymbol{F}_j^\mathsf{T}\},$$

where $E_j = \sum_{k'} \tilde{p}_{k'} f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}), E_j^{(\theta)} = \sum_{k'} \tilde{p}_{k'} D_{k'j}^{*(\theta)}$ and $\boldsymbol{F}_j$ is the vector with kth element $r_k e^{-\xi_k} f(\tilde{h}_k \mid \tilde{\boldsymbol{x}}_j; \boldsymbol{\theta})/E_j$.

References

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Holcroft, C. A., Rotnitzky, A., and Robins, J. M. (1997). Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference*, 65, 349–374.

Jiang, Y. (2004). *Semiparametric maximum likelihood for multi-phase response-selective sampling and missing data problems*. PhD Thesis, University of Auckland.

Lawless, J. F., Wild, C. J., and Kalbfleish (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B*, 61, 413–38.

Lee, A. J. and Hirose, Y. (2006). Semiparametric efficiency bounds for regression models under generalized case-control sampling. Under review for *Annals of Statistics*.

Lee, A. J., Scott, A. J., and Wild, C. J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 97, 23–37.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood (with discussion). *Journal of the American Statistical Association*, 95, 449–485.

Neuhaus, J., Scott, A. J., and Wild, C. J. (2002). The analysis of retrospective family studies. *Biometrika*, 89, 23–37.

Neuhaus, J., Scott, A. J., and Wild, C. J. (2006). Family-specific approaches to the analysis of case-control family data. *Biometrics*, 62, to appear.

Robins, J. M., Rotnitzky, A., Zhao, L. P., and Lipsitz, S. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society: Series B*, 57, 409–424.

Scott, A. J. and Wild, C. J. (1989). Hypothesis testing in case-control studies. *Biometrika*, 76, 806–808.

Scott, A. J. and Wild, C. J. (1991). Fitting logistic models in stratified case-control studies. *Biometrics*, 47, 497–510.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57–71.

Scott, A. J. and Wild, C. J. (2001). Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96, 3–27.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. New York: Wiley.

ALASTAIR SCOTT
Department of Statistics
University of Auckland
Auckland, New Zealand
a.scott@auckland.ac.nz

CHRIS WILD
Department of Statistics
University of Auckland
Auckland, New Zealand
c.wild@auckland.ac.nz