On estimation of conditional density models with two-phase sampling

Zhiqiang Tan
Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA

ARTICLE INFO

Article history:
Received 25 April 2009
Received in revised form
21 January 2010
Accepted 26 January 2010
Available online 4 February 2010

Keywords:
Biased sampling
Conditional density model
Nonparametric likelihood
Profile likelihood
Response-selective sampling
Two-phase sampling

ABSTRACT

Suppose that the conditional density of a response variable given a vector of explanatory variables is parametrically modelled, and that data are collected by a two-phase sampling design. First, a simple random sample is drawn from the population. The stratum membership in a finite number of strata of the response and explanatory variables is recorded for each unit. Second, a subsample is drawn from the phase-one sample such that the selection probability is determined by the stratum membership. The response and explanatory variables are fully measured at this phase. We synthesize existing results on nonparametric likelihood estimation and present a streamlined approach for the computation and the large sample theory of profile likelihood in four different situations. The amount of information in terms of data and assumptions varies depending on whether the phase-one data are retained, the selection probabilities are known, and/or the stratum probabilities are known. We establish and illustrate numerically the order of efficiency among the maximum likelihood estimators, according to the amount of information utilized, in the four situations.

1. Introduction

For a study population, let $Y$ be a response variable and $X$ be a vector of explanatory variables. Consider a model for the conditional density of $Y$ given $X$,

$$
\frac{dP}{d\mu}(y|x) = f(y|x; \theta),
$$

(1)

where $\theta$ is a parameter of dimension $d$, and $\mu$ is a dominating measure. The distribution of $X$, denoted by $G$, is completely unspecified.

Suppose that data are collected in a two-phase sampling design as follows. Such a design is often used as a cost-effective way to allocate resources for data collection. Partition the range of $(Y, X)$ into $J \geq 2$ nonnegligible strata and let $S$ be the stratum variable taking values $1, \ldots, J$. First, an independent and identically distributed (iid) sample of size $N$ is drawn from the population and the stratum information $S$ is recorded. Next, a subsample of size $n$ is drawn from the phase-one sample and the variables $(Y, X)$ are measured. Several sampling schemes can be used at this phase, including Bernoulli sampling, multinomial sampling, and stratified sampling (see Imbens and Lancaster, 1996; Lawless et al., 1999). For simplicity, we focus on Bernoulli sampling in which units are independently selected such that

$$
P(R = 1|Y, X) = \pi_S,
$$

E-mail address: ztan@stat.rutgers.edu

0378-3758/2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.01.041
where $R$ is the selection indicator, and $\pi_j$ is the selection probability for units in stratum $j (j = 1, \ldots, J)$. The sampling is called response-selective or outcome-dependent if $\pi_j$ depends on $Y$. The objective is to estimate $\theta$ for the conditional density model (1).

The following notation will be used. For $j = 1, \ldots, J$, let

$$Q_i(x) = P(S = j | X = x), \quad Q_j = P(S = j) = \int Q_i(x) \, dG.$$

Write $Q(x) = (Q_1(x), \ldots, Q_J(x))^\top$, $Q = (Q_1, \ldots, Q_J)^\top$, and $\pi = (\pi_1, \ldots, \pi_J)^\top$. Label the subjects such that $R_i = 1$ for $i = 1, \ldots, n$ and $= 0$ otherwise. Denote by $N_j$ (or $n_j$) the number of subjects in stratum $j$ in the phase-one (or phase-two) sample. Note that $N_j$ and $n_j$ are random variables, but $\sum_j N_j = N$ is fixed.

There is an extensive literature in statistics and econometrics on conditional density models with two-phase sampling. In fact, a number of variations of this problem are of interest. We distinguish the following four situations.

(I) Both the phase-one and the phase-two data are available.

(II) Only the phase-two data are available and the selective probabilities $\pi_j$ are known.

(III) Only the phase-two data are available and the selective probabilities $\pi_j$ are unknown.

(IV) The phase-two data, together with or without the phase-one data, are available and the stratum probabilities $Q_j$ are known.

Methods for estimation have been proposed and studied at different levels of generality and completeness. Lawless et al. (1999) included a review of methods for situations I and II at the general level as described here. Cosslett (1993) provided a review of methods for situations II, III, and IV in the special case where sampling strata are defined in terms of only $Y$. Examples in this case for dichotomous or categorical $Y$ include Manski and McFadden (1981) for situation II, Anderson (1972) and Prentice and Pyke (1979) for situation III, and Cosslett (1981) for situations III and IV. No methods seem to have been investigated for situations III an IV when sampling strata are defined in terms of both $X$ and $Y$. See Cosslett (1993, Section 6.4) for a related discussion.

The goal of this article is to synthesize existing results on nonparametric likelihood estimation and present a streamlined approach for the computation and the large sample theory of profile likelihood for situations I–IV. We discuss existing methods and results within the streamlined approach, and at the same time obtain interesting new results, which are summarized in the following paragraph. The profile likelihood approach itself is not new in general (e.g., Murphy and van der Vaart, 2000). However, the essence of this approach is yet to be fully appreciated for specific situations.

First, we provide a unified formula for computing the profile likelihood in situation I. This formula remains valid under certain conditions where the formula of Lawless et al. (1999) fails. Second, we provide a more elementary proof of the large sample theory of the profile likelihood as in the classical likelihood theory. Third, we establish the Bartlett identity for the profile likelihood as in the classical likelihood theory. Fourth, we propose use of the expectation-maximization (EM) algorithm (Dempster et al., 1977) in situation I, the iterative proportional scaling (IPS) algorithm (Deming and Stephan, 1940) in situation III, and the trust-region algorithm (e.g., Fletcher, 1987, Section 5.1) in situation IV, as alternative algorithms to the Newton–Raphson algorithm for computing profile likelihoods. Finally, we establish the order of efficiency among the maximum likelihood estimators, according to the amount of information utilized, in situations I–IV.

The rest of this article is organized as follows. Section 2 examines situation I where both the phase-one and the phase-two data are used. Section 3 examines situations II and III where only the phase-two data are used. Section 4 examines situation IV where the stratum probabilities are known. Section 5 compares the profile likelihoods and the maximum likelihood estimators based on different data and assumptions in Sections 2–4. Section 6 summarizes the general approach for the computation and the large sample theory of profile likelihoods and discusses other related topics.

2. Full likelihood

In this section, we study nonparametric likelihood estimation using both the phase-one and the phase-two data. The data consist of the stratum information $(S_1, \ldots, S_N)$ as well as the phase-two data $(Y_1, X_1), \ldots, (Y_n, X_n)$.

2.1. Computation of profile likelihood

Following Lawless et al. (1999) and Scott and Wild (1997, 2001, 2006), the log-likelihood is given by

$$L_n(\theta, G) = \sum_{i=1}^n \log f(Y_i | X_i ; \theta) G_i((X_i)) + \sum_{j=1}^J (N_j - n_j) \log Q_j,$$
where \( Q_j = \int Q(x; \theta) \, dG \) depending on \((\theta, G)\). The profile log-likelihood at \( \theta \) is defined as
\[
\ell_{pl}(\theta) = \max_G \ell_{n}(\theta, G),
\]
where \( G \) is supported on \( \{X_1, \ldots, X_n\} \). The maximum likelihood estimator (MLE) of \( \theta \) is \( \hat{\theta} = \arg\max \ell_{pl}(\theta) \).

By regarding \((\theta, G)\) as free arguments, define
\[
\kappa_n(\theta, Q) = \sum_{j=1}^{N_j} \exp \left( f(Y_j|x_i; \theta) \right) \frac{f(Y_j|x_i; \theta)}{N \sum_{k=1}^{J} \gamma_k Q_k(x_i; \theta)} \right) + \sum_{j=1}^{N_j} (N_j - n_j) \log Q_j,
\]
where \( \gamma_j = 1 - (N_j - n_j)/(N Q_j) \), \( j = 1, \ldots, J \). Lawless et al. (1999, Section 3.3) and Scott and Wild (1997, Appendix 1) derived the profile log-likelihood as
\[
\ell_{pl}(\theta) = \kappa_n(\theta, \hat{Q}(\theta)),
\]
where \( \hat{Q}(\theta) \) solves \( \partial \kappa_n / \partial Q = 0 \) with
\[
\frac{\partial \kappa_n}{\partial Q} = -\frac{N_j - n_j}{Q_j} \left[ \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{J} \gamma_k Q_k(x_i; \theta) - Q_j \right], \quad j = 1, \ldots, J.
\]
However, this formula is appropriate only if \( N_j - n_j > 0 \) for all but at most one stratum. Otherwise, the number of nondegenerate equations, \( \partial \kappa_n / \partial Q = 0 \), is strictly less than \( J - 1 \), the number of unknowns, \( Q_j \) taking into account the constraint \( \sum_{j=1}^{J} Q_j = 1 \). Particularly, this problem arises when two or more strata are empty in the phase-one sample \((N_j=0, \text{and hence } n_j=0)\). Lawless et al. (1999, Section 6) noticed that there are sometimes empty strata in the corners of a cross-classification of \((Y,V)\) with \( V \) a subvector of \( X \), and subsequently combined the corner strata in their simulations. Similarly, Scott and Wild (2006, Section 5) found that using (2) with the full set of \( Q_j \) leads to singular information matrices when \( N_j - n_j = 0 \) for two or more strata.

Lawless et al. (1999) considered the multinomial logistic reparametrization
\[
Q_j = \frac{\exp(\rho_j)}{\sum_{k=1}^{J} \exp(\rho_k)}
\]
to take account of the range constraints \( 0 < Q_j < 1 \) and the constraint \( \sum_{j=1}^{J} Q_j = 1 \). Alternatively, Scott and Wild (2006) considered the reparametrization
\[
Q_j = \frac{\exp(\zeta_j)}{1 + \exp(\zeta_j)}
\]
by ignoring the constraint \( \sum_{j=1}^{J} Q_j = 1 \), and suggested using \( \kappa_n(\theta, \zeta) \) with \( \zeta \) containing only those \( \zeta_j \)'s for which \( N_j - n_j > 0 \). In terms of \( Q_j \), this suggestion is equivalent to using \( \kappa_n(\theta, \hat{Q}) \) with \( \hat{Q} \) the vector of those \( Q_j \)'s for which \( N_j - n_j > 0 \). The dimension of \( \hat{Q} \) is data-dependent because both \( N_j \) and \( n_j \) are data-dependent.

We provide a unified formula for computing the profile log-likelihood (Proposition 1). Compared with \( \partial \kappa_n / \partial Q = 0 \), Eq. (6) below can be solved for the full set of \( Q_j \) regardless of the values of \( N_j - n_j \). Note that the values of \( Q_j \)'s with \( N_j - n_j > 0 \) form a self-contained subsystem of (6). The values of \( Q_j \)'s with \( N_j - n_j = 0 \) are determined by those of \( Q_j \)'s with \( N_j - n_j > 0 \), and so is the profile log-likelihood.

**Proposition 1.**

(i) For fixed \( \theta \), the log-likelihood \( \ell_{n}(\theta, G) \) achieves a maximum over distributions supported on \( \{X_1, \ldots, X_n\} \) at \( \hat{G}(\theta) \) with
\[
\hat{G}(X_i; \theta) = \frac{N^{-1} \sum_{j=1}^{J} \hat{\gamma}_j Q(x_i; \theta)}{\sum_{j=1}^{J} \hat{\gamma}_j Q(x_i; \theta)}, \quad i = 1, \ldots, n,
\]
where \( \hat{\gamma}_j = 1 - (N_j - n_j)/(N Q_j) \), and \( \hat{Q} = \hat{Q}(\theta) \) satisfies
\[
Q_j = \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{J} \gamma_k Q_k(x_i; \theta), \quad j = 1, \ldots, J.
\]

The profile log-likelihood at \( \theta \) is given by (2).

(ii) Furthermore, for any \( G \) supported on \( \{X_1, \ldots, X_n\} \) and any \( Q \) (free of \( G \)) satisfying \( Q_j > 0 \) for \( j = 1, \ldots, J \) and \( \sum_{j=1}^{J} \gamma_j Q(x_i; \theta) > 0 \) for \( i = 1, \ldots, n \),
\[
\max_G \ell_{n}(\theta, G) \leq \min_{Q} \kappa_n(\theta, Q),
\]
where the equality holds at \( \hat{G}(\theta) \) and \( \hat{Q}(\theta) \).

The result (i) accommodates and serves as a justification for the suggestion of Scott and Wild (2006) described after (4). Multiplying (6) by \( \gamma_j \) yields
\[
Q_j = \frac{N_j - n_j}{N} + \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{J} \gamma_k Q_k(x_i; \theta).
\]
Summing over \( j \) implies that \( \sum_{j=1}^{n} Q_j = 1 \) is automatically satisfied. Therefore, either reparametrization (3) or (4) is valid, but the latter seems more convenient.

Eq. (6) can be solved by using the Newton–Raphson algorithm. Alternatively, the log-likelihood for fixed \( \theta \) can be maximized by using the expectation-maximization (EM) algorithm (Dempster et al., 1977) or, more specifically in the context of nonparametric likelihood, the self-consistency algorithm (Efron, 1967; Turnbull, 1976). Here, the self-consistency iteration is given by

\[
G^{(t+1)}(X_i) = N^{-1} + \sum_{j=1}^{n} \frac{N-j}{N} Q(X_i | \theta) G^{(t)}(X_i), \quad i = 1, \ldots, n,
\]

where \( Q^{(t)} = \sum_{j=1}^{n} Q_i(X_i | \theta) G^{(t)}(X_i) \), \( t = 1, 2, \ldots \). The self-consistency equation obtained as \( t \to \infty \) is equivalent to (5). In our simulations, the EM algorithm converges at a slower rate, but is more stable, than the Newton–Raphson algorithm. A hybrid scheme taking advantage of both algorithms is to use the EM algorithm for initial iterations and then switch to the Newton–Raphson algorithm.

The MLE \( \hat{\theta} \) can be computed by further using Newton–Raphson to maximize \( p_{\hat{\theta}}(\theta) \). For this purpose, the gradient and Hessian of \( p_{\hat{\theta}}(\theta) \) are given by

\[
\frac{\partial p_{\hat{\theta}}}{\partial \theta} = \frac{\partial k_n}{\partial \theta} \bigg|_{Q = Q_{\theta}},
\]

\[
\frac{\partial^2 p_{\hat{\theta}}}{\partial \theta^2} = \left[ \frac{\partial^2 k_n}{\partial \theta^2} - \frac{\partial^2 k_n}{\partial \theta \partial Q} \left( \frac{\partial^2 k_n}{\partial Q^2} \right) - \frac{\partial^2 k_n}{\partial Q \partial \theta^T} \right] \bigg|_{Q = Q_{\theta}},
\]

where \( O^- \) denotes the Moore–Penrose generalized inverse for a matrix \( O \). Our implementation consists of two levels. The first is to write a function for \( p_{\hat{\theta}}(\theta) \) together with the gradient and Hessian. The second is to maximize \( p_{\hat{\theta}}(\theta) \) and find \( \hat{\theta} \).

Eq. (7) may lead to another possible algorithm for solving Eq. (6). Consider the following iteration, \( t = 1, 2, \ldots \),

\[
Q^{(t+1)}_j = \frac{N-j}{N} + \frac{1}{N} \sum_{i=1}^{n} \frac{Q_{ji}^{(t)} Q_i(X_i | \theta)}{\sum_{j=1}^{n} Q_{ji}^{(t)} Q_i(X_i | \theta)},
\]

\[
i^{(t+1)}_j = 1 - \frac{N-j}{N Q^{(t+1)}_j},
\]

where \( Q^{(t)}_j = Q_{Nt}/N \) and \( i^{(t)}_j = Q_{Nt}/N \) if \( Nt > 0 \) and \( =1 \) otherwise. Scott and Wild (1997, p. 60) suggested a related algorithm for finding the MLE \( \hat{\theta} \), i.e., solving (6) and \( 0 = \partial k_n / \partial \theta \) jointly. An additional step is to compute \( \theta^{(t)} \) as a solution to \( 0 = \partial k_n / \partial \theta \big|_{Q = Q_{\theta}} \). Then \( Q^{(t+1)}_j \) and \( i^{(t+1)}_j \) are computed with \( \theta \) evaluated at \( \theta^{(t)} \). However, as observed in our simulations for solving (6), the iteration fails to converge to the desired solution in the case where \( \hat{Q}_j < (N_j - n_j)/N \), i.e., \( \hat{Q}_j < 0 \) for some \( j \). In fact, the iteration maintains \( Q^{(t)}_j \geq (N_j - n_j)/N \) and \( i^{(t)}_j \geq 0 \) for each \( t \geq 1 \).

2.2. Theory of profile likelihood

Breslow et al. (2003) established a large sample theory of the profile log-likelihood \( p_{\hat{\theta}}(\theta) \) and the MLEs \( (\hat{\theta}, \hat{C}) \). Under regularity conditions, it is shown that \( p_{\hat{\theta}}(\theta) \) admits an asymptotic expansion in terms of the efficient score and information similar to that of a parametric log-likelihood, and hence satisfies the following properties.

**Prototypical properties of profile likelihood:**

(i) The inverse curvature of \( p_{\hat{\theta}}(\theta) \) at the maximum is a consistent estimator of the asymptotic variance of the MLE \( \hat{\theta} \).

(ii) The log-likelihood ratio statistic \( -2 \ln(p_{\hat{\theta}}(\theta) - p_{\hat{\theta}}(\hat{\theta})) \) is asymptotically chi-squared with the degrees of freedom equal to \( d \).

The proof of Breslow et al. (2003) relies on general results for the profile likelihood in semiparametric models by Murphy and van der Vaart (2000).

We provide a more elementary proof of results (i) and (ii) by treating \( p_{\hat{\theta}}(\theta) \) as a criterion function of its own and \( \hat{\theta} \) as a maximizer to \( p_{\hat{\theta}}(\theta) \). The key lemma is Proposition 2, corresponding to the Bartlett identity in the classical likelihood theory. The remaining proof is standard, invoking Taylor expansions as in the classical large sample theory of MLEs and log-likelihood ratio statistics.

**Proposition 2.** Suppose that as \( N \to \infty \),

\[
\frac{1}{\sqrt{N}} \frac{\partial p_{\hat{\theta}}}{\partial \theta} \to \text{Normal}(0, V) \quad \text{in distribution},
\]

\[
- \frac{1}{N} \frac{\partial^2 p_{\hat{\theta}}}{\partial \theta^2} \to U \quad \text{in probability},
\]

where \( U \) and \( V \) are real matrices. Then \( U=V \).
To show result (i), a first-order Taylor expansion of $\frac{\partial p_n(\hat{\theta})}{\partial \theta}$ yields

$$\hat{\theta} - \theta = -\left(\frac{\partial^2 p_n}{\partial \theta^2}\right)^{-1} \frac{\partial p_n}{\partial \theta} + o_p(N^{-1/2}),$$

which implies that $\sqrt{N}(\hat{\theta} - \theta)$ converges in distribution to $\text{Normal}(0, U^{-1})$, provided that $U=V$ is nonsingular.

To show result (ii), a second-order Taylor expansion of $p_n(\theta)$ yields

$$p_n(\hat{\theta}) - p_n(\theta) = -\frac{1}{2} (\hat{\theta} - \theta)^\top \left(\frac{\partial^2 p_n}{\partial \theta^2}\right) (\hat{\theta} - \theta) + o_p(1),$$

which implies that $-2(p_n(\hat{\theta}) - p_n(\theta))$ converges in distribution to a chi-squared distribution with the degrees of freedom equal to $d$.

3. Conditional likelihoods

In this section, we study nonparametric likelihood estimation using only the phase-two data $(Y_1, X_1), \ldots, (Y_n, X_n)$. The stratum information is not retained for the units that are not completely observed. Not even the value of $N$ is retained. For each of the expressions below, $N$ (if appears) will be cancelled out.

3.1. Known selection probability

Consider the situation where the selection probabilities $(\pi_1, \ldots, \pi_j)$ are known. The conditional log-likelihood for $(Y, X)$ given $R_i = 1, i = 1, \ldots, n$, is

$$\ell_c^i(\theta, G) = \sum_{i=1}^n \log[f(Y_i|X_i; \theta) G((X_i))] - n \log \left(\sum_{j=1}^J \pi_j Q_j\right).$$

For fixed $\theta$, the distribution $\hat{G}(\theta)$ maximizing $\ell_c^i(\theta, G)$ is given by

$$\hat{G}(X_i; \theta) = \frac{\sum_j \pi_j \hat{Q}_j}{n \sum_j \pi_j Q_j(X_i; \theta)}, \quad i = 1, \ldots, n,$

where $\hat{Q}_j = \sum_{i=1}^n Q_i(X_i) \hat{G}(X_i))$. The profile log-likelihood at $\theta$ is therefore

$$p_{n}^{C}(\theta) = \sum_{i=1}^n \log \left[ \frac{f(Y_i|X_i; \theta)}{n \sum_j \pi_j Q_j(X_i; \theta)} \right],$$

free of $(Q_1, \ldots, Q_J)$. Note that $p_{n}^{C}(\theta)$ is identical to the conditional log-likelihood for $Y_i$ given $(X_i, R_i = 1), i = 1, \ldots, n$, as used in Manski and McFadden (1981).

Likewise for $p_n(\theta)$, the Bartlett identity holds for $p_{n}^{C}(\theta)$:

$$\frac{1}{\sqrt{N}} \frac{\partial p_{n}^{C}(\theta)}{\partial \theta} \rightarrow \text{Normal}(0, V_1) \quad \text{in distribution},$$

$$-\frac{1}{N} \frac{\partial^2 p_{n}^{C}(\theta)}{\partial \theta^2} \rightarrow U_1 \quad \text{in probability},$$

where $U_1 = V_1 = B$ defined in the Appendix. Therefore, $p_{n}^{C}(\theta)$ enjoys the prototypical properties of profile log-likelihood described in Section 2.2.

These results are discussed in Lawless et al. (1999, Section 3.1.2). However, Lawless et al. considered $\ell^i_c$ as the resulting log-likelihood for the phase-two data (or complete-data), whereas we consider $\ell^i_c$ as the conditional log-likelihood in the case where the selection probabilities $\pi_j$ are known. This assumption is necessary in order for $\ell^i_c$ to be a function of $(\theta, G)$ only. An alternative conditional likelihood exists when the selection probabilities $\pi_j$ are unknown as we discuss in Section 3.2.

3.2. Unknown selection probability

Consider the situation where the selection probabilities $(\pi_1, \ldots, \pi_j)$ are unknown. The conditional likelihood for $(Y, X)$ given $S_i, i = 1, \ldots, n$, is

$$\ell^i_c(\theta, G) = \sum_{i=1}^n \log[f(Y_i|X_i; \theta) G((X_i))] - \sum_{j=1}^J n_j \log Q_j.$$


Effectively, the data \( \{Y_1, X_1\}, \ldots, \{Y_n, X_n\} \) form a stratified sample from the \( J \) strata with random stratum sizes due to Bernoulli sampling. Previously, nonparametric likelihood estimation has been studied for conditional density models in the case where sampling strata are defined in terms of only \( Y \) and stratified sampling is used with fixed stratum sizes. Anderson (1972) and Prentice and Pyke (1979) provided classical results on the MLE when \( Y \) is dichotomous or categorical and model (1) is logistic or multinomial logistic with intercept terms, in which case only the slope parameters in \( \theta \) can be identified. For general regression models (1), Cosslett (1981, Section 2.10; 1993) derived the profile likelihood and the MLE, and established the asymptotic properties of the MLE. However, the asymptotic properties of the profile likelihood, particularly properties (i) and (ii) described in Section 2.2, remain to be addressed.

Zhou et al. (2002) considered a hybrid sampling design in which a \( Y \)-stratified sample is generated with fixed stratum sizes together with an overall simple random sample. They derived the profile likelihood and the MLE, and established the asymptotic properties of the MLE. However, it remains unclear whether property (i) of profile likelihood in Section 2.2 holds, that is, whether the inverse Fisher information is a consistent estimator of the asymptotic variance of the MLE (see Theorem 1, Zhou et al., 2002).

We focus on the two-phase sampling design without an overall simple random sample, and present our approach for the computation and the large sample theory of profile likelihood. Similar results can be obtained for two-phase sampling designs supplemented with a simple random sample as in Zhou et al. (2002). Our investigation allows the sampling strata to be determined in terms of both \( X \) and \( Y \).

By regarding \((\theta, Q)\) as free arguments, define
\[
\kappa^G_n(\theta, Q) = \sum_{i=1}^n \log \left[ \frac{f(Y_i|X_i; \theta)}{K^G_n(Y_i|X_i; \theta)} \right] - \sum_{j=1}^J n_j \log Q_j,
\]
where \( v_j = n_j/(NQ_j) \), \( j = 1, \ldots, J \). Proposition 3 provides a computational formula for the profile log-likelihood based on \( \ell^G_n(\theta, G) \).

**Proposition 3.**

(i) For fixed \( \theta \), the conditional log-likelihood \( \ell^G_n(\theta, G) \) achieves a maximum over all distributions at \( \hat{G}(\theta) \) with
\[
\hat{G}(X_i; \theta) = \frac{N^{-1}}{\sum_{k=1}^K v_k Q_k(X_i; \theta)}, \quad i = 1, \ldots, n,
\]
where \( v_j = n_j/(NQ_j) \), and \( \hat{Q} = \hat{Q}(\theta) \) satisfies
\[
Q_j = \frac{1}{N} \sum_{i=1}^n \frac{Q(X_i; \theta)}{v_k Q_k(X_i; \theta)}, \quad j = 1, \ldots, J.
\]
The profile log-likelihood \( \ell^G_n(\theta) \) is given by \( \kappa^G_n(\theta, \hat{Q}(\theta)) \).

(ii) Furthermore, for any \( G \) and any \( Q \) (free of \( G \)) satisfying \( Q_j > 0 \) for \( j = 1, \ldots, J \),
\[
\max_G \ell^G_n(\theta, G) \leq \max_Q \kappa^G_n(\theta, Q)
\]
where the equality holds at \( \hat{G}(\theta) \) and \( \hat{Q}(\theta) \).

The result (i) follows from that of Vardi (1985) for nonparametric likelihood estimation in biased sampling models. However, an important generalization allowed here is that \( \hat{G}(\theta) \) and \( \hat{Q}(\theta) \) may not be unique [see the discussion below on the uniqueness of \( \hat{G}(\theta) \) and \( \hat{Q}(\theta) \)]. The observations \( \{X_i, S_i = j, i = 1, \ldots, n\} \) constitute an iid biased sample from \( G \):
\[
dP(X = x|S = j) = \frac{Q(x)}{Q_j} dG(x),
\]
where \( Q(x) \) serves as a weight function, and \( Q_j \) the normalizing constant. Eqs. (10) and (11) correspond to those (see equations 3.6 and 3.7, Vardi, 1985) for estimating the baseline distribution and the normalizing constants in a biased sampling model. Similar equations are obtained in a related model for Monte Carlo integration (see equations 3.4 and 3.5, Kong et al., 2003).

Eqs. (10) and (11) are algebraically similar to (5) and (6) with \( v_j = n_j/(NQ_j) \) in place of \( \gamma_j = 1-(N_j-n_j)/(NQ_j) \). But unlike (6), Eq. (11) does not imply \( \sum_{j=1}^J Q_j = 1 \). The \( J \) equations of (11) are linearly dependent, because multiplying (11) by \( v_j \) and summing over \( j \) yield \( n/N \) on both sides. Therefore, Eq. (11) need to be considered with the constraint \( \sum_{j=1}^J Q_j = 1 \) simultaneously imposed.

A necessary and sufficient condition for the uniqueness of \( \hat{G}(\theta) \) and \( \hat{Q}(\theta) \) is Vardi’s (1985) connectedness condition: for each pair \( 1 \leq k, j \leq J \), either \( Q_k(X_j) > 0 \) and \( Q_k(X_j) > 0 \) for some \( X_j \) or there exists a chain of such relationships connecting the pair. Equivalently, the union of \( X_j \) such that \( Q_k(X_j) > 0 \) for a subset of strata \( j \) should overlap with that for the complementary subset of strata. Otherwise, let \( J' \subset \{1, \ldots, J\} \) indicate such a subset of strata and \( I \subset \{1, \ldots, n\} \) indicate \( X_i \)
such that $Q_j(X_i) > 0$ for $j \in \mathcal{J}$. Then the subsystem of (11) in $Q_j$, $j \in \mathcal{J}$, becomes self-contained:

$$Q_j = \frac{1}{N} \sum_{i \in I} \frac{Q_j(X_i)}{\sum_{k \in \mathcal{J}} v_k Q_j(X_i)}, \quad j \in \mathcal{J}. \quad (13)$$

These equations are linearly dependent, because multiplying (13) by $v_j$ and summing over $j \in \mathcal{J}$ yield $\sum_{j \in \mathcal{J}} n_j / N$ on both sides. Assume that the subset $\mathcal{J}$ is connected (or a subset of $\mathcal{J}$ can be considered). Then $Q_j$, $j \in \mathcal{J}$, are uniquely determined by (13) up to a positive multiple, and so is the restriction of $G$ to $\{X_i : i \in I\}$ by (10). The contribution of $X_i$, $i \in I$, to the profile log-likelihood is given by

$$\sum_{i \in I} \log \left[ \frac{f(Y_i | X_i; \theta)}{N \sum_{k \in \mathcal{J}} v_k Q_k(X_i; \theta)} \right] - \sum_{j \in \mathcal{J}} n_j \log \hat{Q}_j,$$

and is invariant to the scaling of $\hat{Q}_j$, $j \in \mathcal{J}$. Therefore, the profile log-likelihood is uniquely determined whether or not the connectedness condition holds.

Suppose that $f(y|x; \theta) > 0$ everywhere in the product space of $(y, x)$. The connectedness condition is satisfied when the strata are defined by a classification of $Y$ only, but not when the strata are defined by a cross-classification of $(Y, V)$ with $V$ a subvector of $X$. In the latter case, the subset of strata within each class of $V$ is connected in itself, but disconnected from other strata. Nevertheless, the condition can be satisfied when the strata are obtained by grouping those initially defined by a cross-classification. For example, suppose that $Y$ has three levels (low, medium, high) and $V$ has two levels (low, high), where the levels can be intervals if $Y$ or $V$ is continuous. The connectedness condition is satisfied when the set of strata consists of $(Y, V) = (\text{low, low}), (\text{low, high}), (\text{medium, low or high}), (\text{high, low}), (\text{high, high})$. Here, cross-classification by $Y$ and $V$ is employed only for $Y$ in the low or high level (or in the two tails).

If the connectedness condition holds, Eq. (11) with $\sum_{j=1}^J Q_j = 1$ can be solved by using the Newton–Raphson algorithm. Otherwise, individual subsystems like (13) need to be solved separately. Alternatively, Eqs. (10) and (11) can be solved by using the iterative proportional scaling (IPS) algorithm (Darroch and Ratcliff, 1972; Deming and Stephan, 1940), as discussed in Kong et al. (2003) for similar equations. For the $n \times J$ array $\{n_j Q(X_i)\}$, the condition given by (10) and (11) indicates that after rescaling, $n_j Q(X_i) \rightarrow n Q(X_i) G_i / q_j$, each row total is 1 and the $j$th column total is $n_j$. In addition, the normalization condition requires that $\sum_{i=1}^N G(X_i) = 1$ or equivalently $\sum_{j=1}^J Q_j = 1$. The IPS iteration is therefore

$$G^{(t+1)}(X_i) = \frac{N^{-1} \sum_{k=1}^J v_k Q(X_i)}{\sum_{i=1}^N \sum_{k=1}^J v_k Q(X_i)} = \frac{n^{-1} \sum_{k=1}^J v_k Q(X_i)}{\sum_{i=1}^N \sum_{k=1}^J v_k Q(X_i)}.$$

In terms of $Q_j^{(t)}$, a full iteration is given by

$$Q_j^{(t+1)} = \frac{1}{N} \sum_{i=1}^n Q_j(X_i) G^{(t+1)}(X_i).$$

The sequence $Q_j^{(t)}, j = 1, \ldots, J$, converges to a solution of (11) as $t \rightarrow \infty$ whether or not the connectedness condition holds. In our simulations, the IPS algorithm is very stable and converges reasonably fast. Like the hybrid scheme in Section 2.1, applying the IPS algorithm and the Newton–Raphson algorithm sequentially can be used to gain the speed and the stability of convergence.

The large sample behavior of $pl_\infty^C(\theta)$ can be studied similarly as in Section 2.2. The formulas (8) and (9) for the gradient and Hessian are applicable. Proposition 4 shows that the Bartlett identity holds. As a result, $pl_\infty^C(\theta)$, like $pl_\infty(\theta)$ and $pl_\infty^S(\theta)$, enjoys the prototypical properties of profile log-likelihood under standard regularity conditions (including the nonsingularity of $U_2 = V_2$).

**Proposition 4.** Suppose that as $N \rightarrow \infty$,

$$\frac{1}{\sqrt{N}} \frac{\partial pl_\infty^C}{\partial \theta} \rightarrow \text{Normal}(0, V_2) \quad \text{in distribution},$$

$$- \frac{1}{N} \frac{\partial^2 pl_\infty^C}{\partial \theta^2} \rightarrow U_2 \quad \text{in probability},$$

where $U_2$ and $V_2$ are real matrices. Then $U_2 = V_2$.

Note that if $Y$ is dichotomous or categorical and model (1) is logistic or multinomial logistic with intercept terms, then $pl_\infty^C(\theta)$ is flat along the direction of each intercept in $\theta$, and hence $U_2 = V_2$ is singular. Nevertheless, $pl_\infty^S(\theta)$ with fixed intercept parameters in this case is a proper profile log-likelihood of the slope parameters alone with the prototypical properties (Prentice and Pyke, 1979; Cosslett, 1981).
4. Auxiliary information

In this section, we study nonparametric likelihood estimation when the stratum probabilities \((Q_1, \ldots, Q_J)\) are known. With this auxiliary information, the likelihood principle implies that inference should be free of whether the stratum information \(S\) is retained for the phase-one sample. Both \(l_m(\theta, G)\) and \(l^0_m(\theta, G)\) are parallel to

\[
l^0_m(\theta, G) = \sum_{i=1}^{n} \log \left( f(Y_i, X_i; \theta) G(X_i) \right),
\]

where \(Q_j = \int Q_j(x; \theta) \, dG\) is known, \(j = 1, \ldots, J\). For each of the expressions below, \(N\) (if appears) will be cancelled out as in Section 3.

The data \((Y_i, X_i), \ldots, (Y_n, X_n)\) form a stratified sample with random stratum sizes as in Section 3.2. Similarly, nonparametric likelihood estimation has been studied for conditional density models with \(Y\)-stratified sampling when \((Q_1, \ldots, Q_J)\) are known. Cosslett (1981, Section 2.16; 1993) derived the profile likelihood and the MLE, and established the asymptotic properties of the MLE. However, the profile likelihood as a useful inferential tool has not been explicitly investigated.

We present our approach for the computation and the large sample theory of profile likelihood, and provide results for the general case where the sampling strata are determined in term of both \(X\) and \(Y\). For \(Y\)-stratified sampling, if the joint probabilities are known for a cross-classification of \((Y, V)\) with \(V\) a subvector of \(X\), then post-stratification on both \(Y\) and \(V\) can be employed to make use of such knowledge.

By regarding \((\theta, \pi)\) as free arguments, define

\[
k^\pi_m(\theta, \pi) = \sum_{i=1}^{n} \log \left( \frac{f(Y_i, X_i; \theta)}{n \sum_{k=1}^{J} \pi_k Q_k(X_i; \theta)} \right) + n \sum_{j=1}^{J} \pi_j Q_j.
\]

Proposition 5 provides a formula for the profile log-likelihood based on \(l^0_m(\theta, G)\).

**Proposition 5.**

(i) For fixed \(\theta\), the log-likelihood \(l^0_m(\theta, G)\) achieves a maximum over distributions supported on \(\{X_1, \ldots, X_n\}\) at \(G(\theta)\) with

\[
\hat{G}(X_i; \theta) = \frac{N^{-1}}{\sum_{k=1}^{J} \pi_k Q_k(X_i; \theta)}, \quad i = 1, \ldots, n,
\]

where \(\hat{\pi} = \hat{\pi}(\theta)\) maximizes, subject to \(\sum_{j=1}^{J} \pi_j Q_j = n/N\),

\[
\sum_{i=1}^{n} \log \left( \sum_{k=1}^{J} \pi_k Q_k(X_i; \theta) \right).
\]

The search over \(\pi\) is restricted to \(\Pi_n = \{\pi: \sum_{j=1}^{J} \pi_j Q(X; \theta) > 0, i = 1, \ldots, n\}\). The profile log-likelihood \(p_{n}^0(\theta)\) is given by \(k^\pi_m(\hat{\theta}(\hat{\pi}))\).

(ii) Furthermore, for any \(G\) supported on \(\{X_1, \ldots, X_n\}\) and any \(\pi \in \Pi_n\),

\[
\max_G l_{n}^0(\theta, G) \leq \min_{\pi} k^\pi_m(\theta, \pi),
\]

where the equality holds at \(G(\hat{\theta})\) and \(\hat{\pi}(\theta)\).

The results (i) and (ii) can be proved similarly as in Tan (2004) for nonparametric likelihood estimation in linear submodels for Monte Carlo integration. The fact that \(Q_j\) are known adds the linear constraints, \(f Q_j(x) \, dG = Q_j\), to the (full) biased sampling model (12). The definition of \(\hat{\pi}\) suggests that \(\hat{\pi}\) is obtained by maximum likelihood in fitting to the data \(\{X_1, \ldots, X_n\}\) the mixture model with \(J\) components \(Q_j(x) \, dG\) and mixture coefficients \(\propto \pi_j Q_j\). Interestingly, this estimation is required even if the true selection probabilities \(\pi_j\) are known. The coefficients \(\pi_j Q_j\) can be 0 or negative as long as \(\pi \in \Pi_n\). The estimates \(\hat{\pi}_j Q_j\) are close to, but are not necessarily equal to, \(\pi_j / N\), and both of them converge to \(\pi_j Q_j\) (that is, \(\hat{\pi}_j \to \pi_j\)) as \(N \to \infty\).

The critical condition for maximizing (15) indicates that \(\hat{\pi}\) satisfies

\[
Q_j = \frac{1}{N} \sum_{i=1}^{n} \frac{Q_j(X_i; \theta)}{\sum_{k=1}^{J} \pi_k Q_k(X_i; \theta)}, \quad j = 1, \ldots, J,
\]

which implies \(\sum_{j=1}^{J} \pi_j Q_j = n/N\). Eqs. (14) and (16) are algebraically similar to (5) and (6) and to (10) and (11), with subtle differences [see Imbens, 1992, for a related discussion on (11) and (16)]. For (6) or (11), \(Q_j\) are unknown and \(\gamma_j\) or \(\eta_j\) are (known) functions of \(Q_j\). For (16), \(Q_j\) are known and \(\pi_j\) are unknown, free of \(Q_j\). Nevertheless, \(\gamma_j\), \(\eta_j\), and \(\hat{\pi}_j\) converge to the same limits, that is, the true selection probability \(\pi_j\) as \(N \to \infty\). Therefore, (5), (10), and (14) are estimators of \(G\).
Section 3.2. Let the connectedness condition fail, for example, when the strata are defined by a cross-classification of (Fletcher, 1987, Section 5.1). Furthermore, the dimensionality of the optimization can be effectively reduced when the strata are defined by a cross-classification of (Y, V) as discussed in Section 3.2. Let J ⊂ {1, ..., J} indicate a subset of strata that is connected in itself but disconnected from other strata, and I ⊂ {1, ..., n} indicate X such that Q(Xi) > 0 for i ∈ J. Then (15) becomes a sum of two components in πj, j ∈ J, and in πj, j /∈ J, respectively. The component in πj, j ∈ J, is given by

\[
\sum_{i \in J} \log \left( \sum_{k \in J} \pi_k Q_k(X_i) \right).
\]

The subsystem of (16) in πj, j ∈ J, becomes self-contained:

\[
Q_j = \frac{1}{N} \sum_{i \in J} \sum_{k \in J} \pi_k Q_k(X_i), \quad j \in J.
\]

Therefore, πj, j ∈ J, can be obtained by maximizing (17) subject to \( \sum_{k \in J} \pi_k Q_k = n/N \). For the cross-classification example, the dimension of the vector of πj, j ∈ J, is the number of levels of Y, but that of π = (π1, ..., πJ) \( ^T \) is the product of those of Y and V.

The large sample behavior of \( \hat{\pi}_n^A(\theta) \) can be studied similarly as in Sections 2.2 and 3.2. The formulas (8) and (9) for the gradient and Hessian are applicable. Proposition 6 shows that the Bartlett identity holds. As a result, \( \hat{\pi}_n^A(\theta) \), like \( \hat{\pi}_n(\theta) \), \( \hat{\pi}_n^{C_1}(\theta) \), and \( \hat{\pi}_n^{C_2}(\theta) \), enjoys the prototypical properties of profile log-likelihood under standard regularity conditions (including the nonsingularity of \( U_0 = V_0 \)).

**Proposition 6.** Suppose that as \( N \to \infty, \)

\[
\frac{1}{\sqrt{N}} \frac{\partial \hat{\pi}_n^A}{\partial \theta} \to \text{Normal (0, V_0)} \quad \text{in distribution,}
\]

\[
\frac{-1}{N} \frac{\partial^2 \hat{\pi}_n^A}{\partial \theta^2} \to U_0 \quad \text{in probability,}
\]

where \( U_0 \) and \( V_0 \) are real matrices. Then \( U_0 = V_0 \).

5. **Comparison**

**Efficiency comparison:** In Sections 2–4, we have examined the profile log-likelihoods \( \hat{\pi}_n(\theta) \), \( \hat{\pi}_n^{C_1}(\theta) \), \( \hat{\pi}_n^{C_2}(\theta) \), and \( \hat{\pi}_n^A(\theta) \), and the corresponding MLEs \( \hat{\theta}, \hat{\theta}^{C_1}, \hat{\theta}^{C_2}, \) and \( \hat{\theta}^A \). These likelihoods and estimators are based on increasingly less informative data and assumptions in the following order:

(i) \( \hat{\theta}^A \) is based on the phase-one and the phase-two data (or the phase-two data alone) under the assumption that the stratum probabilities \( Q \) are known, and is not affected by whether the selection probabilities \( \pi_j \) are known;

(ii) \( \hat{\theta} \) is based on both the phase-one and the phase-two data, and is not affected by whether the selection probabilities \( \pi_j \) are known;

(iii) \( \hat{\theta}^{C_1} \) is based on only the phase-two data under the assumption that the selection probabilities \( \pi_j \) are known;

(iv) \( \hat{\theta}^{C_2} \) is based on only the phase-two data under the assumption that the selection probabilities \( \pi_j \) are unknown.

Interestingly, we show in the Appendix that the efficiencies of the estimators obey the same order of magnitude:

\[
\text{eff}(\hat{\theta}^A) \geq \text{eff}(\hat{\theta}) \geq \text{eff}(\hat{\theta}^{C_1}) \geq \text{eff}(\hat{\theta}^{C_2}),
\]

where eff denotes the inverse of asymptotic variance, and \( O_1 \geq O_2 \) indicates that \( O_1 - O_2 \) is nonnegative definite for symmetric matrices \( O_1 \) and \( O_2 \). In this sense, likelihood serves as an effective tool for capturing information (in terms of data and assumptions) and drawing inference for the present situations. These results agree with those in the classical likelihood theory on the effectiveness of using likelihood for capturing data and assumptions. First, the MLE in the presence of missing data is no more efficient than the MLE given the complete data. Second, the MLE under a submodel with additional assumptions is no more efficient than the MLE under the full model.
**Numerical example:** For illustration, consider a surrogate covariate problem as described in Lawless et al. (1999). The data-generating process is as follows: (i) generate \( X_1 \sim \text{Normal}(0, 1), X_2 \sim \text{Normal}(0, 1) \) correlated .9 with \( X_1 \), and a six-level surrogate \( V \) for \( X_2 \) by discretizing \( X_2 \) at (-2, -1, 0, 1, 2); (ii) generate

\[
Y = \theta_0 + \theta_1 X_1 + \epsilon,
\]

where \( \epsilon \sim \text{Normal}(0, \sigma^2) \), independent of \( (X_1, X_2) \); (iii) define three strata, \( \{Y \leq a_1\}, \{a_1 < Y \leq a_2\}, \) and \( \{Y > a_2\} \), and select units at phase two with probability \( \pi_1, \pi_2, \) or \( \pi_3 \) in the three strata. As in Lawless et al. (1999, Section 6.3), \( a_1 \) and \( a_2 \) are fixed at the 5% and 95% quantiles of \( Y \), and \( \pi_1 = \pi_2 = 0.25, \pi_3 = 0.014 \), so that the numbers of units selected at phase two in the three strata are approximately the same. We investigate estimation of \( (\theta_0, \theta_1, \sigma) \) for model (18) with \( X = (X_1, V) \). In the simulations discussed below, \( \theta_0 = \theta_1 = 1, \sigma = 1, \) and \( N = 20,000 \).

Suppose that the surrogate \( V \) is available for all the units at phase one, but \( X_1 \) is measured only for the units in the phase-two sample. The likelihoods studied in Sections 2–4 can be applied in the following two manners: (i) the surrogate \( V \) is ignored, and \( S \) indicates the three strata of \( Y \); (ii) the surrogate \( V \) is used for post-stratification, and \( S \) indicates the 18 cross-classified strata of \( (Y, V) \). In the latter case, the corner strata may be empty even in the phase-one sample (\( N = 0 \)).

For a simulated dataset, Fig. 1 draws the profile log-likelihoods (divided by \( N \)) marginally in \( \theta_1 \) at the MLEs of \( \theta \) with the maxima aligned at 0. The profile log-likelihoods \( p_{\text{full}}(\theta) \), \( p_{\text{cond}}(\theta) \), \( p_{\text{aux}}(\theta) \), and \( p_{\text{aux}}^2(\theta) \) are increasingly more flatly peaked (with decreasing curvatures). Furthermore, \( p_{\text{full}}(\theta) \) involving the surrogate \( V \) is more sharply peaked than that ignoring the surrogate \( V \). Additional information about \( \theta \) is provided by the phase-one data of \( V \), even though the conditional density of \( Y \) given \( (X_1, V) \) is free of \( V \). Similarly, \( p_{\text{full}}(\theta) \) involving the surrogate \( V \) is more sharply peaked than that ignoring the surrogate \( V \). In contrast, \( p_{\text{aux}}^2(\theta) \) involving the surrogate \( V \) is more flatly peaked than that ignoring the surrogate \( V \). A stratified sample

---

**Table 1**

<table>
<thead>
<tr>
<th></th>
<th>No surrogate</th>
<th>Surrogate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full</td>
<td>Cond. 1</td>
</tr>
<tr>
<td><strong>Bias</strong></td>
<td>0.0021</td>
<td>-0.0015</td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td>0.00103</td>
<td>0.00117</td>
</tr>
<tr>
<td><strong>Est. var.</strong></td>
<td>0.000975</td>
<td>0.00111</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>0.896</td>
<td>0.891</td>
</tr>
</tbody>
</table>

**Note:** Variance is the empirical variance, Est. var. is the average of the variance estimates, and Coverage is the coverage frequency of 90% CI.
from a cross-classification of \((Y, V)\) is less informative about \(\theta\) than that from a classification of \(Y\) alone, because the stratum fractions are unknown by assumption. Finally, \(p_{in}^{C_1}(\hat{\theta})\) involving the surrogate \(V\) is the same as that ignoring the surrogate \(V\).

Table 1 summarizes estimates of \(\theta_i\) from 1000 repeated simulations. The estimators \(\hat{\theta}^A\), \(\hat{\theta}^C\), and \(\hat{\theta}^{C_2}\) are increasingly less precise. Furthermore, \(\hat{\theta}^A\) involving the surrogate \(V\) are more precise and \(\hat{\theta}^{C_1}\) involving the surrogate \(V\) is less precise than those ignoring the surrogate \(V\), whereas \(\hat{\theta}^{C_1}\) involving the surrogate \(V\) is the same as that ignoring the surrogate \(V\). The averages of the variance estimates from the Fisher information agree well with the empirical variances of the estimates. The coverage frequencies of 90\% two-sided confidence intervals (CIs) are close to the nominal level.

6. Discussion

Profile likelihoods and MLEs: For each situation of Sections 2–4, the profile likelihood approach is to first derive the profile likelihood as a stand-alone function, and then find the MLE as a maximizer to that function. This approach is straightforward and useful for developing both the computation and the large sample theory, and highlights a common structure in the profile likelihoods for these different situations.

First, the computation consists of two nested levels, with iterations typically needed within each level. The first (or inner) level is to maximize the profile likelihood and find the MLE. Second, the formulas for the gradient and Hessian of the profile likelihood can be directly used to obtain the Bartlett identity, which can then be used to establish the large sample results for the profile likelihood and the MLE as in the classical likelihood theory.

Third, our formulas for \(p_{pl}(\theta), p_{pl}^{C_1}(\theta), \) and \(p_{pl}^{C_2}(\theta)\) share a common structure, in which the “duality functions” \(\kappa_n(\theta, Q), \) \(\kappa_n^{C_1}(\theta, Q), \) and \(\kappa_n^{C_2}(\theta, \pi)\) are defined such that

(i) \(\kappa_n(\theta, Q)\) evaluated at \(Q = \hat{Q}(\theta)\) gives \(p_{pl}(\theta),\) \(\kappa_n^{C_1}(\theta, Q)\) at \(Q = \hat{Q}(\theta)\) gives \(p_{pl}^{C_1}(\theta),\) and \(\kappa_n^{C_2}(\theta, \pi)\) at \(\pi = \hat{\pi}(\theta)\) gives \(p_{pl}^{C_2}(\theta);\)

(ii) \(\partial \kappa_n/\partial Q\) evaluated at \(Q = \hat{Q}(\theta)\) gives \(0, \partial \kappa_n^{C_1}/\partial Q\) at \(Q = \hat{Q}(\theta)\) gives \(0,\) and \(\partial \kappa_n^{C_2}/\partial \pi\) at \(\pi = \hat{\pi}(\theta)\) gives \(0.\)

Condition (ii) is important, because it underlies Eqs. (8) and (9) for the gradient and Hessian of \(p_{pl}(\theta)\) and those for \(p_{pl}^{C_1}(\theta)\) and \(p_{pl}^{C_2}(\theta),\) and hence simplifies the proofs of Propositions 2, 4, and 6 considerably. For example, by condition (i) alone, \(\kappa_n^{C_1}(\theta, \pi)\) can be replaced by

\[
\sum_{i=1}^n \log \left[ \frac{f(Y_i|X_i; \theta)}{N \sum_{k=1}^K \pi_k Q_k(X_i; \theta)} \right],
\]

which, evaluated at \(\pi = \hat{\pi}(\theta)\), gives \(p_{pl}^{C_2}(\theta).\) But this function does not satisfy condition (ii), and therefore is less desirable than \(\kappa_n^{C_1}(\theta, \pi)\) currently defined.

The functions \(\kappa_n(\theta, Q), \kappa_n^{C_1}(\theta, Q), \) and \(\kappa_n^{C_2}(\theta, \pi)\) are in duality with the log-likelihoods \(l_n(\theta, G), \) \(l_n^{C_1}(\theta, G), \) and \(l_n^{C_2}(\theta, G)\) respectively, as captured by Propositions 1(ii), 3(ii), and 5(ii). These inequalities directly imply that conditions (i) and (ii) are satisfied, and make it clear that the MLE \((\hat{\theta}, \hat{Q})\) is a saddle point of \(\kappa_n(\theta, Q),\) \((\hat{\theta}^{C_1}, \hat{Q}^{C_1})\) a maximum point of \(\kappa_n^{C_1}(\theta, Q),\) and \((\hat{\theta}^{C_2}, \hat{\pi})\) a saddle point of \(\kappa_n^{C_2}(\theta, \pi),\) as previously discussed by Cossett (1981) and Scott and Wild (2006). The function \(\kappa_n(\theta, Q), \kappa_n^{C_1}(\theta, Q), \) or \(\kappa_n^{C_2}(\theta, \pi)\) seems to be (but is not) the log-likelihood function of \((\theta, Q)\) or \((\theta, \pi).\) The MLE in each case is a solution to the critical conditions jointly in \((\theta, Q)\) or \((\theta, \pi).\) But solving nonlinear equations in this way is computationally challenging. Our two-level computational scheme provides a simple, effective solution, involving optimization within each level and making it possible to use techniques such as the EM, IPS, and trust-region algorithms.

Pseudo-likelihoods: Lawless et al. (1999, Section 3.2) reviewed several methods using pseudo-likelihoods. Here, we discuss a log pseudo-likelihood related to the profile log-likelihood \(p_{pl}^{C_1}(\theta)\) and highlight their different properties.

Consider the log pseudo-likelihood function

\[
\kappa_n^{C_1}(\theta) = \sum_{i=1}^n \log \left[ \frac{f(Y_i|X_i; \theta)}{n \sum_{k=1}^K \pi_k Q_k(X_i; \theta)} \right]
\]

obtained by substituting \(\pi_j = n_j/N\) for \(\pi_j\) in \(p_{pl}^{C_1}(\theta).\) This function is no longer a profile log-likelihood. It is straightforward to show that the Bartlett identity fails as expected. Therefore, \(\kappa_n^{C_1}(\theta)\) lacks the prototypical properties of profile log-likelihood described in Section 2.2. Next, consider the estimator \(\hat{\theta}^{C}\) that maximizes \(\kappa_n^{C}(\theta)\) or equivalently solves \(\partial \kappa_n^{C}/\partial \theta = 0.\) In the Appendix, we show that this estimator is sandwiched between \(\hat{\theta}\) and \(\hat{\theta}^{C_1}\) in terms of efficiency:

\[
\text{eff}(\hat{\theta}) \geq \text{eff}(\hat{\theta}^{C}) \geq \text{eff}(\hat{\theta}^{C_1}).
\]

Informally, the first inequality holds because \(\hat{\theta}\) is the MLE using the phase-one and the phase-two data, whereas the second inequality holds because \(\hat{\theta}^{C}\) uses the phase-one and the phase-two data but \(\hat{\theta}^{C_1}\) uses only the phase-two data.
Continuous phase-one data: In the setting of Section 2, additional information besides the stratum variable S can be available at phase one. For example, this situation occurs when Y is continuous and measured for the phase-one sample, and S is defined by a discretization of Y. Then the log-likelihood is given by

\[ L_n(\theta, G) = \sum_{i=1}^{n} \log[f(Y_i|X_i, \theta)G(X_i)] + \sum_{j=n+1}^{N} \log(f(Y_j)), \]

where \( f(Y_j) = \int f(Y_j|x, \theta) \, dG \) plays the role of \( Q_j \). Maximizing \( L_n(\theta, G) \) over \( G \) supported on \( \{X_1, \ldots, X_n\} \) leads to equations similar to (5) and (6), but the unknowns involved are \( f(Y_j), j = n+1, \ldots, N \). A challenge is that the number of unknowns increases as the sample size increases. Song et al. (2009) showed that the resulting profile likelihood still has prototypical asymptotic properties as in Section 2.2, and the MLE is semiparametric efficient. Their proof is based on empirical process

Acknowledgment

The author thanks a referee for helpful comments. This research was supported by the U.S. National Science Foundation.

Appendix A

Proof of Proposition 1.

(i) We restrict \( G \) to distributions supported on \( \{X_i : i = 1, \ldots, n\} \), and write \( w_i = G(X_i) \). For fixed \( \theta \), the log-likelihood is up to a constant

\[ \sum_{i=1}^{n} \log(w_i) + \sum_{j=1}^{f} (N_j - n_j) \log\left(\sum_{i=1}^{n} Q_i(X_i)w_i\right). \]

By introducing a Lagrange multiplier \( \lambda \) for the constraint \( \sum_{i=1}^{n} w_i = 1 \) and setting the derivative with respect to \( w_i \) equal to 0, we obtain

\[ \frac{1}{w_i} + \sum_{j=1}^{f} (N_j - n_j) \frac{Q_j(X_i)}{Q_j} - \lambda = 0, \quad i = 1, \ldots, n. \]

Multiplying through by \( w_i \) and summing over \( i \) yields that \( \lambda = N \). Therefore,

\[ w_i = \frac{N^{-1}}{\sum_j \gamma_j Q_j(X_j)}, \quad i = 1, \ldots, n. \]

The resulting \( Q = (Q_1, \ldots, Q_f)^T \) satisfies Eq. (6).

(ii) Let \( G \) be a distribution supported on \( \{X_1, \ldots, X_n\} \) and write \( Q_j = \int Q_j(x) \, dG \) for \( j = 1, \ldots, f \). Let \( Q' = (Q'_1, \ldots, Q'_f)^T \) be a vector of positive numbers such that \( \sum_{j=1}^{f} \gamma'_j Q'_j(X_i) > 0 \) for \( i = 1, \ldots, n \), where \( \gamma'_j = 1 - (N_j - n_j)/NQ'_j \). Then \( L_n(\theta, G) \leq \kappa_n(\theta, Q') \), because by Jensen's inequality,

\[ L_n(\theta, G) - \kappa_n(\theta, Q') \leq N \log \left[ \sum_{i=1}^{n} G(X_i) \left( \sum_{k=1}^{f} \gamma'_k Q'_k(X_i) + \frac{1}{N} \sum_{j=1}^{f} (N_j - n_j) Q'_j / Q'_j \right) \right] = 0. \]

Proof of Proposition 2. Write \( \square f = \partial f(y|x; \theta)/\partial \theta \) and \( \square Q_k = \partial Q_k(x; \theta)/\partial \theta \).

(i) By direct calculations, we find

\[ \frac{1}{N} \square \kappa_n = \frac{1}{N} \sum_{i=1}^{n} \left[ \frac{\nabla f(Y_i|X_i)}{f(Y_i|X_i)} - \frac{\sum_{k=1}^{f} \gamma_k \nabla Q_k(X_i)}{\sum_{k=1}^{f} \gamma_k Q_k(X_i)} \right]. \]
where $\hat{A} = \text{diag}((N_j - n_j)/(NQ^2_j))$ and $Z \otimes^2 = ZZ^\top$ for a vector $Z$. Therefore,

$$-\frac{1}{N} \begin{pmatrix} \frac{\partial^2 \kappa_n}{\partial \theta^2} & \frac{\partial^2 \kappa_n}{\partial \theta \partial Q^\top} \\
\frac{\partial^2 \kappa_n}{\partial Q \partial \theta^\top} & \frac{\partial^2 \kappa_n}{\partial Q^2} \end{pmatrix} \rightarrow \begin{pmatrix} B & -DA \\
-AD^\top & -(A + ACA) \end{pmatrix}$$
in probability,

where $A = \text{diag}((1 - \pi_j)/Q_j)$ and

$$B = \mathbb{E} \left[ R \frac{(\nabla f(Y|X))^\otimes^2}{f^2(Y|X)} - R \left( \frac{\sum_{j=1}^p \pi_k \nabla Q_k(X)}{\sum_{j=1}^p \pi_k Q_k(X)} \right) \right],$$

$$C = \mathbb{E} \left[ R \frac{Q(X)^\otimes^2}{\left( \sum_{j=1}^p \pi_k Q_k(X) \right)^2} \right],$$

$$D = \mathbb{E} \left[ R \frac{\left( \sum_{j=1}^p \pi_k \nabla Q_k(X) \right) Q^\top(X) \left( \sum_{j=1}^p \pi_k Q_k(X) \right)^2}{\left( \sum_{j=1}^p \pi_k Q_k(X) \right)^2} \right].$$

(ii) By Taylor expansions for $\gamma' = (\gamma_1, \ldots, \gamma^\top)$ about $\pi = (\pi_1, \ldots, \pi^\top)$, we find

$$\frac{1}{N} \frac{\partial \kappa_n}{\partial \theta} = \frac{1}{N} \sum_{i=1}^n \left[ \frac{\nabla f(Y_i|X_i)}{f(Y_i|X_i)} - \frac{\sum_{j=1}^p \pi_k \nabla Q_k(X_i)}{\sum_{j=1}^p \pi_k Q_k(X_i)} \right] + \frac{1}{N} \sum_{i=1}^n \left( \frac{\sum_{j=1}^p \pi_k \nabla Q_k(X_i)}{\sum_{j=1}^p \pi_k Q_k(X_i)} \right)^2 (\gamma - \pi) + o_p(N^{-1/2}),$$

$$\frac{1}{N} \frac{\partial \kappa_n}{\partial Q} = -\hat{A} \left[ \frac{1}{N} \sum_{i=1}^n \frac{Q(X_i)}{\sum_{j=1}^p \pi_k Q_k(X_i)} - Q \right] + \hat{A} \frac{1}{N} \sum_{i=1}^n \left( \frac{\sum_{j=1}^p \pi_k Q_k(X_i)}{\sum_{j=1}^p \pi_k Q_k(X_i)} \right)^2 (\gamma - \pi) + o_p(N^{-1/2}).$$

Note that $\gamma_1 - \pi_1 = -(NQ_1)^{-1} \sum_{i=1}^N (1 - R_i)1[S_i = j] - (1 - \pi_1)Q_j$ and $\sqrt{N}(\gamma_1 - \pi_1, \ldots, \gamma_j - \pi_j)^\top$ converges to multivariate normal with mean 0 and variance

$$\Pi = \text{diag}((1/Q)^{-1} \sum_{i=1}^N (1 - R_i)1[S_i = j] - (1 - \pi_1)Q_j) - ((1 - \pi)^\otimes^2) \text{diag}(1/Q) - (1 - \pi)^\otimes^2.$$

Therefore, $N^{-1/2} \frac{\partial \kappa_n}{\partial \theta}$ and $N^{-1/2} \frac{\partial \kappa_n}{\partial Q^\top}$ converge to multivariate normal with mean 0 and variance

$$V' = \begin{pmatrix} B & 0 \\
0 & A(C - Q\otimes^2 A) \end{pmatrix} \Pi(D^\top D)^{-1} \begin{pmatrix} B & 0 \\
0 & A(C - Q\otimes^2 A) \end{pmatrix} + \begin{pmatrix} D & -(1 - \pi)^\otimes^2 \\
-(1 - \pi)^\otimes^2 D^\top & -(1 - \pi)^\otimes^2 C - A(C - \pi)^\otimes^2 \end{pmatrix}.$$
where the simplification is due to $A(A + A^2)^{-1} A = (I + A)^{-1} A$ and $D - DA(A + A^2)^{-1} A = D(I + A)^{-1}$. Throughout, $I_d$ and $J_d$ are identity matrices of size $d$ and $J$. Next, by (i) and (9), $-N^{-1}(\partial^2 p_n / \partial \theta^2)$ converges in probability to

$$U = B + DA(A + A^2)^{-1} A = B + D(I + A)^{-1}.$$ 

It follows that $U = V$, because $A(C - Q \otimes^2) A + II + 2(1 - \pi)^2 = A + A^2$. □

**Proof of Proposition 3.** (i) The result follows from Vardi (1985).

(ii) Write $Q_j = \int Q_j(x) dG$ for $j = 1, \ldots, J$. It is sufficient to show $\ell_{n_i}^2(\theta, G) = \ell_{n_i}^2(\theta, Q)$. In fact, by Jensen’s inequality,

$$\ell_{n_i}^2(\theta, Q) - \ell_{n_i}^2(\theta, Q) = \sum_{i=1}^n \log \left[ G_i(X_i) / \sum_{k=1}^J v_k Q_k(X_i) \right] \leq n \log \left[ \sum_{i=1}^n \left( \sum_{k=1}^J v_k Q_k(X_i) \right) \right] = 0. \quad \Box$$

**Proof of Proposition 4.** The proof is similar to that of Proposition 2. (i) By direct calculations, we find

$$\frac{1}{N} \frac{\partial \ell_{n_i}^2}{\partial \theta} = \frac{1}{N} \sum_{i=1}^n \left[ \frac{\partial f_i(X_i)}{f_i(X_i)} \right] - \frac{1}{N} \sum_{i=1}^n \left( \sum_{k=1}^J v_k Q_k(X_i) \right) \frac{\partial^2 (\sum_{k=1}^J v_k Q_k(X_i))}{\partial \theta} + o_p(1),$$

$$\frac{1}{N} \frac{\partial^2 \ell_{n_i}^2}{\partial \theta^2} = \frac{1}{N} \sum_{i=1}^n \left( \sum_{k=1}^J v_k Q_k(X_i) \right)^2 \frac{\partial^2 (\sum_{k=1}^J v_k Q_k(X_i))}{\partial \theta^2} + o_p(1),$$

$$\frac{1}{N} \frac{\partial^2 \ell_{n_i}^2}{\partial \theta \partial Q} = \frac{1}{N} \sum_{i=1}^n \left( \sum_{k=1}^J v_k Q_k(X_i) \right)^2 \frac{\partial^2 (\sum_{k=1}^J v_k Q_k(X_i))}{\partial \theta \partial Q} + o_p(1),$$

where $H = \text{diag}(\pi_i/(NQ_j^2))$. Therefore,

$$\frac{1}{N} \left( \begin{array}{ccc} \frac{\partial^2 \ell_{n_i}^2}{\partial \theta^2} & \frac{\partial^2 \ell_{n_i}^2}{\partial \theta \partial Q} \\ \frac{\partial^2 \ell_{n_i}^2}{\partial \theta \partial Q} & \frac{\partial^2 \ell_{n_i}^2}{\partial Q^2} \end{array} \right) \rightarrow \left( \begin{array}{cc} B & DH \\ HD & H-H \end{array} \right)$$

in probability,

where $H = \text{diag}(\pi_i/Q_j)$, and $B$, $C$, and $D$ are defined in the proof of Proposition 2. (ii) By Taylor expansions for $v = (v_1, \ldots, v_j)^T$ about $\pi = (\pi_1, \ldots, \pi_j)^T$, we find

$$\frac{1}{N} \frac{\partial \ell_{n_i}^2}{\partial \theta} = \frac{1}{N} \sum_{i=1}^n \left[ \frac{\partial f_i(X_i)}{f_i(X_i)} \right] - \frac{1}{N} \sum_{i=1}^n \left( \sum_{k=1}^J v_k Q_k(X_i) \right) \frac{\partial^2 (\sum_{k=1}^J v_k Q_k(X_i))}{\partial \theta} + o_p(N^{-1/2}),$$

$$\frac{1}{N} \frac{\partial^2 \ell_{n_i}^2}{\partial \theta^2} = \frac{1}{N} \sum_{i=1}^n \left( \sum_{k=1}^J v_k Q_k(X_i) \right)^2 \frac{\partial^2 (\sum_{k=1}^J v_k Q_k(X_i))}{\partial \theta^2} + o_p(N^{-1/2}),$$

$$\frac{1}{N} \frac{\partial^2 \ell_{n_i}^2}{\partial \theta \partial Q} = \frac{1}{N} \sum_{i=1}^n \left( \sum_{k=1}^J v_k Q_k(X_i) \right)^2 \frac{\partial^2 (\sum_{k=1}^J v_k Q_k(X_i))}{\partial \theta \partial Q} + o_p(N^{-1/2}).$$

Note that $v - \pi = (NQ_j)^{-1} \sum_{i=1}^N [R_i S_i = j] - \pi Q_j$ and $\sqrt{N}(v_i - \pi_1, \ldots, v_j - \pi_j)^T$ converges to multivariate normal with mean 0 and variance

$$\Sigma = \text{diag}(1/Q_1) - \text{diag}(\pi Q)^{\otimes^2},$$

Therefore, $N^{-1/2}(\partial \ell_{n_i}^2 / \partial \theta^T, \partial \ell_{n_i}^2 / \partial Q^T)^T$ converges to multivariate normal with mean 0 and variance

$$\Sigma = \text{diag}(1/Q_1) - \text{diag}(\pi Q)^{\otimes^2},$$

Therefore, $N^{-1/2}(\partial \ell_{n_i}^2 / \partial \theta^T, \partial \ell_{n_i}^2 / \partial Q^T)^T$ converges to multivariate normal with mean 0 and variance

$$\Sigma = \text{diag}(1/Q_1) - \text{diag}(\pi Q)^{\otimes^2}.$$
(iii) By a Taylor expansion, we find
\[
\frac{1}{N} \frac{\partial \kappa_n^A}{\partial \pi} = \frac{1}{N} \left[ \frac{\partial \kappa_n^A}{\partial \theta} \frac{\partial \kappa_n^A}{\partial \theta^2} + \left( \frac{\partial^2 \kappa_n^A}{\partial \theta^2 \partial Q} - \frac{\partial^2 \kappa_n^A}{\partial \theta \partial Q^2} \right) \right] + o_p(N^{-1/2}).
\]

Therefore, \(N^{-1/2}(\partial \kappa_n^A/\partial \theta)\) converges to multivariate normal with mean 0 and variance
\[
V_2 = \begin{pmatrix} I_d & - \mathbf{D} \mathbf{H}^{-1} (\mathbf{P} + \mathbf{Q}) \mathbf{H}^{-1} \mathbf{D}^T \end{pmatrix} \begin{pmatrix} I_d & - \mathbf{D} \mathbf{H}^{-1} (\mathbf{P} + \mathbf{Q}) \mathbf{H}^{-1} \mathbf{D}^T \end{pmatrix}^{-1},
\]

where the simplification is due to \(\mathbf{H}(\mathbf{P} + \mathbf{Q})\mathbf{H}^{-1} \mathbf{D}^T\). On the other hand, \(-N^{-1}(\partial^2 \kappa_n^A/\partial \theta^2)\) converges in probability to
\[
U_2 = \mathbf{B} - \mathbf{D} \mathbf{H}^{-1} (\mathbf{P} + \mathbf{Q}) \mathbf{H}^{-1} \mathbf{D}^T.
\]

It follows that \(U_2 = V_2\), because \(\mathbf{H}(\mathbf{P} + \mathbf{Q})\mathbf{H}^{-1} \mathbf{D}^T\). □

**Proof of Proposition 5.** See the proof of Theorem 2 in Tan (2004). □

**Proof of Proposition 6.**

(i) By direct calculations, we find
\[
\frac{1}{N} \frac{\partial \kappa_n^A}{\partial \pi} = \frac{1}{N} \sum_{i=1}^{n} \frac{Q(X_i)}{\sum_{k=1}^{J} \pi_k Q_k(X_i)} + \frac{n}{N} \sum_{k=1}^{J} \pi_k Q_k,
\]
\[
\frac{1}{N} \frac{\partial \kappa_n^A}{\partial \theta} = \frac{1}{N} \sum_{i=1}^{n} \frac{\nabla f(Y_i | X_i)}{f(Y_i | X_i)} - \frac{\sum_{k=1}^{J} \pi_k \nabla Q_k(X_i)}{\sum_{k=1}^{J} \pi_k Q_k(X_i)},
\]
\[
\frac{1}{N} \frac{\partial^2 \kappa_n^A}{\partial \theta^2} = -\frac{1}{N} \sum_{i=1}^{n} \left[ \frac{(\nabla f(Y_i | X_i))^2}{f(Y_i | X_i)} - \frac{\sum_{k=1}^{J} \pi_k \nabla Q_k(X_i)}{\sum_{k=1}^{J} \pi_k Q_k(X_i)} \right] + o_p(1),
\]
\[
\frac{1}{N} \frac{\partial^2 \kappa_n^A}{\partial \pi \partial \pi} = \frac{1}{N} \sum_{i=1}^{n} \left( \sum_{k=1}^{J} \pi_k Q_k(X_i) \right)^2 - \frac{n}{N} \sum_{k=1}^{J} \pi_k Q_k \sum_{k=1}^{J} \pi_k.
\]
Note that \(n/N \to \sum_{k=1}^{J} \pi_k Q_k\). Therefore,
\[
-\frac{1}{N} \left( \frac{\partial^2 \kappa_n^A}{\partial \pi^2} \frac{\partial^2 \kappa_n^A}{\partial \pi^2} \right) \to \begin{pmatrix} \mathbf{B} & -\mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{B} & -\mathbf{D} \end{pmatrix}^{-1} \quad \text{in probability},
\]

where \(\mathbf{B}, \mathbf{C},\) and \(\mathbf{D}\) are defined in the proof of Proposition 2. (ii) \(N^{-1/2}(\partial \kappa_n^A/\partial \theta^T, \partial \kappa_n^A/\partial \pi^T)^T\) converges to multivariate normal with mean 0 and variance
\[
V_0 = \begin{pmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{C} - \frac{\mathbf{Q} \mathbf{Q}^T}{\sum_{k=1}^{J} \pi_k Q_k} \end{pmatrix}.
\]

(iii) By a Taylor expansion, we find
\[
\frac{1}{N} \frac{\partial^2 \kappa_n^A}{\partial \theta^2} = \frac{1}{N} \left[ \frac{\partial^2 \kappa_n^A}{\partial \theta^2} - \frac{\partial \kappa_n^A}{\partial \theta} \frac{\partial \kappa_n^A}{\partial \theta^2} \right] + o_p(N^{-1/2}).
\]

Therefore, \(N^{-1/2}(\partial \kappa_n^A/\partial \theta)\) converges to multivariate normal with mean 0 and variance
\[
V_0 = \begin{pmatrix} I_d & - \mathbf{D} \left( \frac{\mathbf{Q} \mathbf{Q}^T}{\mathbf{Q} \mathbf{Q}^T} \right) \end{pmatrix} \begin{pmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{C} - \frac{\mathbf{Q} \mathbf{Q}^T}{\sum_{k=1}^{J} \pi_k Q_k} \end{pmatrix} \begin{pmatrix} I_d & - \mathbf{D} \left( \frac{\mathbf{Q} \mathbf{Q}^T}{\mathbf{Q} \mathbf{Q}^T} \right) \end{pmatrix} + \mathbf{B} + \mathbf{D} \left( \frac{\mathbf{Q} \mathbf{Q}^T}{\sum_{k=1}^{J} \pi_k Q_k} \right) \mathbf{D}^T.\]
On the other hand, \(-N^{-1}(\hat{\varphi}^2 p_j^B/\hat{\vartheta} ^2_j)\) converges in probability to
\[
U_0 = B + D\left(C - \frac{Q \otimes 2}{\sum_{k=1}^r \pi_k Q_k} \right)^{-1} D^\top.
\]
It follows immediately that \(U_0 = V_0\). □

**Efficiency comparisons.** The efficiency comparison between \(\hat{\vartheta}, \hat{\vartheta}^C\), and \(\hat{\vartheta}^A\) follows, because \(U_0 \geq U \geq U_1(=B) \geq U_2\). The third inequality holds, because \(\text{diag}(Q/\pi) - C \geq D^\top B^{-1} D \geq 0\) (see below) and hence \(H^{-1} H C \geq 0\).

Next, the asymptotic variance of \(N(\hat{\vartheta} - \vartheta)\) is
\[
B^{-1} (B - D \text{diag}(\pi(1 - \pi)/Q) D^\top) B^{-1}.
\]
The asymptotic variance of \(N(\hat{\vartheta} - \vartheta)\) is
\[
\left( B + D A (A + C A)^{-1} D^\top \right)^{-1} = B^{-1} - B^{-1} D A (A + C A + D^\top B^{-1} D) A^{-1} D^\top B^{-1}.
\]
For the efficiency comparison between \(\hat{\vartheta}, \hat{\vartheta}^C\), and \(\hat{\vartheta}^A\), it remains to show that \(C + D^\top B^{-1} D \leq \text{diag}(Q/\pi)\) or equivalently \(\text{diag}(Q/\pi) - C \geq D^\top B^{-1} D\).

This inequality follows by taking
\[
Z_1 = R \left[ \frac{\pi Q(X)}{\sum_{k=1}^r \pi_k Q_k(X)} - (1(S = 1), \ldots, 1(S = J)) \right]^\top,
\]
\[
Z_2 = R \left[ \frac{\nabla f(Y|X)}{\nabla f(Y|X)} - \sum_{k=1}^r \pi_k \nabla Q_k(X) \right]^\top
\]
in the Cauchy–Schwartz inequality \(\text{var}(Z_1) \geq \text{cov}(Z_1, Z_2)\text{var}^{-1}(Z_2)\text{cov}^{-1}(Z_1, Z_2)\). □

**References**


