Scandinavian
Journal of Statistics Theory and Applications

doi: 10.1111/j.1467-9469.2012.00811.x

Evaluating Statistical Hypotheses Using Weakly-Identifiable Estimating Functions

GUANQUN CAO

Department of Statistics and Probability, Michigan State University

DAVID TODEM

Department of Epidemiology and Biostatistics, Michigan State University

LIJIAN YANG

Center for Advanced Statistics and Econometrics Research, Soochow University; and Department of Statistics and Probability, Michigan State University

JASON P. FINE

Deparment of Biostatistics, University of North Carolina

ABSTRACT. Many statistical models arising in applications contain non- and weakly-identified parameters. Due to identifiability concerns, tests concerning the parameters of interest may not be able to use conventional theories and it may not be clear how to assess statistical significance. This paper extends the literature by developing a testing procedure that can be used to evaluate hypotheses under non- and weakly-identifiable semiparametric models. The test statistic is constructed from a general estimating function of a finite dimensional parameter model representing the population characteristics of interest, but other characteristics which may be described by infinite dimensional parameters, and viewed as nuisance, are left completely unspecified. We derive the limiting distribution of this statistic and propose theoretically justified resampling approaches to approximate its asymptotic distribution. The methodology's practical utility is illustrated in simulations and an analysis of quality-of-life outcomes from a longitudinal study on breast cancer.

Key words: estimating equations, global sensitivity analysis, infimum and supremum statistics, missing not at random, model misspecification, pseudolikelihood

1. Introduction

Data in statistical research are often well described by models, in which the scientific questions of interest are described by an unknown, finite-dimensional parameter vector. Such models may be either fully parametric or semiparametric, where other aspects of the model may be described by infinite dimensional parameters which are completely unspecified. In such settings, it is often of interest to use the observed data in order to draw inferences about the parameters of interest. Standard inferential techniques may be applied if the parameters of interest can be well estimated by minimizing a parametric loss function or more generally by solving a parametric estimating function which does not involve infinite dimensional nuisance parameters. In many situations, however, these parameters may be non-identifiable or at best weakly identifiable from the estimating function so that the standard inferential theories may not be valid. The objective of this paper is to develop hypothesis tests for scenarios in which the model parameters are weakly identifiable. Conceptually, the term weak identifiability refers to the situations where data contain some information about model parameters but not enough to identify them uniquely.

To illustrate the problem quite sharply, we consider a simple theoretical example where a fully parametric model is indexed by an unknown parameter vector (θ, β) for an observable

random quantity Y. We assume that realizations $\{Y_i\}_{i=1}^n$ of Y are independent and identically distributed (i.i.d.) normal $\mathcal{N}(\theta+\beta, 1)$ variates. The objective is to evaluate the hypothesis $H_0 : \theta_0 = 0$, where θ_0 is the true value of θ. Using only observed data and assuming that β_0, the true value of β is unknown, inferences for θ_0 may not be conducted using standard techniques due to identifiability problems arising from the mean model being overparameterized.

Another interesting, more practical illustration of this problem comes from the missing data literature where weakly-identifiable models are frequently encountered. Specific examples include the study of publication bias in meta-analysis (Chambers & Welsh, 1993; Copas & Li, 1997; Copas, 1999) and the analysis of longitudinal data subject to non-random non-responses (Scharfstein *et al.*, 1999; Kenward *et al.*, 2001; Rotnitzky *et al.*, 2001; Little & Rubin, 2002). Identifiability issues commonly arise with non-random missing data, where the parameters in the model for the missingness may not be jointly identifiable with those in the model for the outcomes of interest using only the observed data, particularly with semiparametric models, where some of the nuisance parameters may be infinite dimensional. Analyses which assume identifiability may be unreliable, with the joint selection and outcome model yielding flat 'estimation' surfaces potentially having multiple modes. These phenomena have previously been reported by several authors in modelling potentially non-ignorable missing data models (Scharfstein *et al.*, 1999; Todem *et al.*, 2010).

In section 3, we consider these missing data issues when analyzing longitudinal data with informative dropout employing the model of Troxel *et al.* (1998b). The model is semiparametric, with the parameter being estimated denoted by (θ, β), where β is the selection parameter that measures the extent of non-randomness of the missing data mechanism and θ consists of the remaining finite dimensional parameters of the selection and outcome models. The hypotheses of interest concern covariate effects on the outcome, which are contained in θ. In Troxel *et al.* (1998b), a so-called pseudo-likelihood analysis, described in detail in section 3, was carried out under the assumption of parameter identifiability. The resulting estimating function only involves (θ, β), with the longitudinal dependence in the outcomes completely unspecified and not estimated. We investigated the parameter identifiability assumption in a reanalysis of the cancer data from Troxel *et al.* (1998b) by profiling the pseudo-likelihood analysis in β (Fig. 1). The profile pseudolikelihood is flat in β, suggesting a model that is at

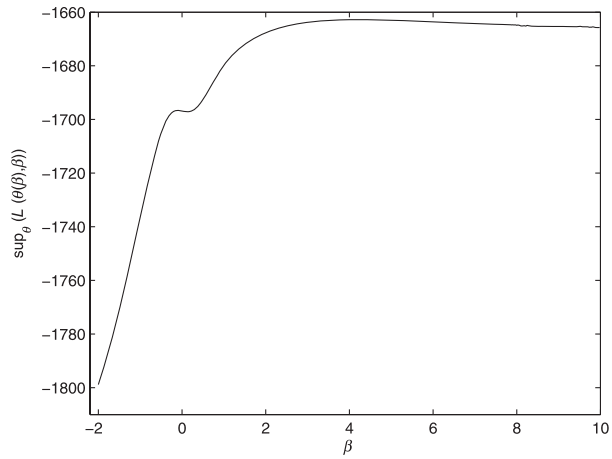

Fig. 1. Supremum of the pseudo-likelihood function profiled across β, the parameter measuring the extent of non-randomness of the missing data mechanism in the study.

best weakly identifiable. These results draw into question inferences which assume identifiability of θ and β.

Due to identifiability concerns, tests concerning the model parameters cannot use conventional theory to assess statistical significance. Essentially, the standard estimation and inference techniques may fail due to the models being overparameterized. A natural remedy is to partition the parameter indexing the estimating function into certain parameters of interest and other parameters which may be viewed as secondary parameters. For the theoretical example discussed earlier, where Y is normally distributed, the parameter of interest in light of the hypothesis under study is θ, while β is the secondary parameter. In the missing data application (Troxel *et al.*, 1998b), the parameter β which describes the informativeness may be viewed as the secondary parameter, while the covariate effects in θ may be of primary interest in hypothesis testing. In practice, the choice of θ and β will depend on the application.

Various approaches to the problem of non-identifiable parameters that have appeared in the literature focused primarily on maximum likelihood based procedures. Almost all previous works in hypothesis testing deal with the case where non-identifiability only occurs under the null hypothesis. Examples include Davies (1977, 1987), Hansen (1996), Ritz & Skovgaard (2005) and Song *et al.* (2009). Generally, this requires that the model is identifiable under the alternative hypothesis. In sensitivity analysis, the testing problem has a different formulation. The model may not be identifiable under either the null or the alternative hypothesis. Moreover, even after fixing a set of parameters, it may not be clear whether the parameters of interest can be consistently estimated under the null hypothesis. To be concrete, in the normal example, for each value of β, the maximum likelihood estimator of θ consistently estimates $\theta_0+\beta-\beta_0$, where θ_0 and β_0 are the true values of θ and β. This only equals θ_0 when $\beta=\beta_0$. Our approach to inference about the parameters of interest is to adapt the profiling strategy from the earlier works described above. Because the testing problem is fundamentally different, the resulting developments are non-standard, with relatively little work in the literature on this problem. Since the model may not be identifiable even after profiling, we need to consider the behaviour of the profile estimator under model misspecification under the null.

This inferential strategy poses substantial technical challenges beyond those encountered with supremum tests which assume identifiability under the alternative. In missing data applications used to motivate the sensitivity analysis, rigorous results for full likelihood analyses have been established (Lu & Copas, 2004), essentially requiring model identifiability. More recently, Todem *et al.* (2010) demonstrated how to conduct likelihood inference via infimum tests, including a precise analysis of the behaviour of the profile estimators under model misspecification and the distribution of the corresponding infimum test. Such tests are particularly important when the quantity being tested does not increase or decrease monotonically as the non-identified parameters are increased or decreased. Under monotonicity, it is only necessary to perform the tests at the limits of the non-identified parameter space. They developed simultaneous confidence bands which enable identification of those values of the sensitivity parameter for which significant results are obtained. Although these likelihood-based methods are useful, they require a full distribution specification for the data. This can be a difficult task in practice, especially when observed data do not have enough information to fully identify the parameter of interest.

In this paper, we extend the profiling idea to arbitrary estimating functions involving θ and β but which do not require a complete parametric model specification. Our set-up includes the likelihood score functions as a special case. The generalization of the infimum test and confidence bands to non-likelihood settings is non-trivial. The infimum test has the advantage

that it is simply defined directly in terms of contrasts whereas the supremum tests are obtained through non-trivial derivations using the log-likelihood functions (Dacunha-Castelle & Gassiat, 1999). We present generic conditions which establish the large sample properties of the estimating function for θ profiled on β, including the uniform consistency and weak convergence of the θ estimator as a function of β. To our knowledge, these theoretical results are novel, with issues related to non-identifiable estimating functions not having been studied rigorously, previously. We accommodate misspecification and uniformity in β in a general paradigm which permits the profiling to be carried out with respect to any suitable estimating function. Owing to the complexity of the asymptotic distributions of the infimum test and confidence bands, resampling is needed. A theoretically justified procedure is discussed for approximating such distributions.

The rest of this article is organized as follows. In section 2, we present the general framework of the problem, the proposed test and the resampling procedure, along with a proof of the key asymptotic properties. In section 3, the methodology is exhibited using the cancer dataset in Troxel *et al.* (1998b) and in simulations, where the naive Wald test may have either inflated type I error rate or reduced power. Some remaining issues are discussed in section 4.

2. The method

2.1. The general framework

We consider a model involving a finite dimensional parameter $\varpi \in \Omega$ for an observable random quantity Y. The parameter ϖ may not completely determine the distribution of Y, that is, there may be other aspects of the model which are unspecified. The interest is drawing inferences about ϖ with i.i.d. realizations $\{Y_i\}_{i=1}^n$ of Y and a general estimating function $S_Y(\varpi)$. Denote by ϖ_0 the true value of ϖ. If $E\{S_Y(\varpi_0)\} = 0$, then an estimator $\hat{\varpi}$ of ϖ_0 usually can be obtained by solving the estimating equation, $S_Y(\varpi) = 0$; see chapter 5 of van der Vaart (2000b) for an overview of Z-estimators. If S_Y identifies ϖ_0, then under other mild regularity conditions, this estimating equation yields a consistent and asymptotically normal parameter estimator. Under such regularity conditions, inferences about ϖ_0 can be conducted using the large sample properties of $\hat{\varpi}$. Problems may occur if the model as a function of ϖ is 'overparameterized', with multiple values of ϖ satisfying $E\{S_Y(\varpi)\} = 0$. In this case, the estimator may not have the usual asymptotic properties.

Non-identifiability can be addressed by fixing some components of ϖ, conditional upon which the remaining parameters are uniquely defined by S_Y. One may partition $\varpi = (\theta, \beta)$, where θ, a p-dimensional vector, is assumed to be 'identifiable' for a fixed q-dimensional vector β, as defined in section 2.2. If the true value β_0 of the non-identified parameter β is known, the estimator $\hat{\theta}_0$ at $\beta = \beta_0$ can be used to conduct reliable inferences about θ_0, the true value of θ. This estimator is readily available by solving the estimating equation $S_Y(\theta, \beta_0) = 0$, for fixed and known β_0. The approach is unfeasible, as the true value β_0 is usually unknown to the analyst in practice. A common strategy is to fix β and study the estimator of θ at various values of $\beta \in \Xi$. To highlight the dependence on β, we denote by $\hat{\theta}(\beta)$, the estimator of θ for a fixed β. The estimator of θ when $\beta = \beta_0$ is $\hat{\theta}_0 = \hat{\theta}(\beta_0)$.

For the simple normal example, $\hat{\theta}(\beta) = \bar{Y} - \beta$ and $\hat{\theta}(\beta_0) = \bar{Y} - \beta_0$, where $\bar{Y}$ is the sample mean. This estimator is normally distributed with mean $\theta_0 + \beta_0 - \beta$ and variance n^{-1}, uniformly in β, for each fixed n. Of course, in general, it is not possible to obtain clean finite sample results and large sample approximations are needed. In the subsection below, we study the uniform asymptotic properties of $\hat{\theta}(\beta)$ for $\beta \in \Xi$.

2.2. Large sample properties of $\hat{\theta}(\beta)$

When β is fixed at its true value β_0, it is well established that for an estimating function $S_Y(\theta, \beta_0)$ which is smooth in θ, the estimator $\hat{\theta}$ is consistent and approximately normal under mild regularity conditions (see, for example, van der Vaart & Wellner (2000a)). That is, $n^{\frac{1}{2}}\{\hat{\theta}(\beta_0) - \theta_0\} \to_d \mathcal{N}(0, \Sigma_0)$, where $\Sigma_0 = (D(\theta_0))^{-1}\mathrm{var}(S_Y(\theta_0, \beta_0))(D^{-1}(\theta_0))^{\mathrm{T}}$, with $D(\theta_0)$ being the expected value of the first-order derivative of $S_Y(\theta, \beta_0)$ with respect to θ. These properties of $\hat{\theta}(\beta_0)$ can be used to conduct large-sample inferences about θ_0.

For a given β, the estimator $\hat{\theta}(\beta)$ will converge to a quantity $\theta^*(\beta)$, which is generally different from θ_0 if $\beta \neq \beta_0$. For the simple normal example, $\theta^*(\beta) = \theta_0 + \beta_0 - \beta$. This contrasts with set-ups on testing with non-identifiability under the null (Davies, 1977, 1987), where it is generally assumed that $\theta^*(\beta) = \theta_0$ for all β. Moreover, appropriately standardized, $\hat{\theta}(\beta)$ will be asymptotically normal, with variance which may be estimated using a sandwich variance approach. This is an extension of standard pointwise asymptotic theory for maximum likelihood estimation with misspecified models, originating in the seminal work of Huber (1967) and White (1982). We study below the uniform convergence of this estimator across all values of $\beta \in \Xi$.

Suppose the data consist of i.i.d. realizations $\{Y_i\}_{i=1}^n$ of Y. Let $s_{Y_i}(\theta, \beta)$ be the contribution of subject i to the estimating function $S_Y(\theta, \beta)$. Define $S_n(\theta, \beta) = n^{-1}\sum_{i=1}^n s_{Y_i}(\theta, \beta)$ and $\tilde{S}(\theta, \beta) = E\{s_{Y_1}(\theta, \beta)\}$. Let $g_{Y_i}(\theta, \beta) = \partial s_{Y_i}(\theta, \beta)/\partial\theta$, $W_Y(\theta, \beta) = n^{-1}\sum_{i=1}^n g_{Y_i}(\theta, \beta)$ and $\tilde{W}(\theta, \beta) = E\{g_{Y_1}(\theta, \beta)\}$. For any given $\beta \in \Xi$, let $\hat{\theta}(\beta)$ denote the solution to $S_Y(\theta, \beta) = 0$, that is $S_Y(\hat{\theta}(\beta), \beta) = 0$. The 'least false' (White, 1982) parameter $\theta^*(\beta)$, satisfies $\tilde{S}(\theta^*(\beta), \beta) = 0$. Define $\mathcal{G}_1 = \{s_{Y_i}(\theta, \beta) : i = 1, \ldots, n, \theta \in \Theta, \beta \in \Xi\}$ and $\mathcal{G}_2 = \{g_{Y_i}(\theta, \beta) : i = 1, \ldots, n, \theta \in \Theta, \beta \in \Xi\}$.

We assume the following regularity conditions:

- (C1) The sets $\Theta \subset \mathbb{R}^p$ and $\Xi \subset \mathbb{R}^q$ are compact and $\theta^*(\beta)$ is an interior point of Θ for any $\beta \in \Xi$.
- (C2) The function classes, $\mathcal{G}_1$ and $\mathcal{G}_2$, are pointwise measurable and satisfy the uniform entropy condition (van der Vaart & Wellner, 2000a).
- (C3) $\inf_{\theta \in \Theta,\, \beta \in \Xi} \lambda_{\min}\{-\tilde{W}(\theta, \beta)\} > 0$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix.
- (C4) The estimating function $S_Y(\theta, \beta)$ has continuous first-order derivatives with respect to θ for any given $\beta \in \Xi$.

Condition C1 defines the parameter space for the implied parameter $\theta^*(\beta)$ for a given β. Because $\theta^*(\beta)$ may be non-constant in β, the parameter space for $\theta^*(\beta)$ across β is contained in a suitably defined functional space. Conditions C2 and C3 give conditions under which uniform asymptotic results for $\theta^*(\beta)$ may be obtained. The entropy condition C2 ensures that the estimating function is well behaved across all β. The condition is satisfied by functions which are uniformly bounded and uniformly Lipschitz of order $> \{\dim(\theta) + \dim(\beta)\}/2$, where $\dim(\cdot)$ denotes the dimension of a vector. Condition C3 guarantees the identifiability of $\theta^*(\beta)$ for all β. The longitudinal data model presented in section 3 meets these requirements. Note that the smoothness specified in condition C4 only applies to θ. Differentiability in β is not assumed. Non-smoothness in θ could be accommodated under stronger assumptions.

The proof of theorem 1 is provided in the Appendix.

Theorem 1. *Under conditions* C1–C4, $\sup_{\beta \in \Xi} \|\hat{\theta}(\beta) - \theta^*(\beta)\| \to_p 0$, *where* $\|\cdot\|$ *represents the Euclidean norm. Furthermore,* $n^{\frac{1}{2}}(\hat{\theta}(\beta) - \theta^*(\beta))$ *converge weakly to a tight Gaussian process with positive definite covariance function*

$$\Sigma^*(\beta_1,\beta_2)=\lim_{n\to\infty}\operatorname{cov}\{n^{\frac{1}{2}}(\hat\theta(\beta_1)-\theta^*(\beta_1)),n^{\frac{1}{2}}(\hat\theta(\beta_2)-\theta^*(\beta_2))\}$$
$$=[\{\tilde W(\theta^*(\beta_1),\beta_1)\}^{-1}]^{\mathrm T}E\{s_{Y_1}(\theta,\beta_1)s_{Y_1}^{\mathrm T}(\theta,\beta_2)\}\{\tilde W(\theta^*(\beta_2),\beta_2)\}^{-1}.$$

For fixed β,

$$\Sigma^*(\beta,\beta)=\lim_{n\to\infty}\operatorname{var}\{n^{\frac{1}{2}}(\hat\theta(\beta)-\theta^*(\beta))\}$$
$$=[\{\tilde W(\theta^*(\beta),\beta)\}^{-1}]^{\mathrm T}E\{s_{Y_1}(\theta,\beta)s_{Y_1}^{\mathrm T}(\theta,\beta)\}\{\tilde W(\theta^*(\beta),\beta)\}^{-1}.$$

The covariance function may be easily estimated using a robust sandwich variance estimator along the lines of White (1982), which is valid under model misspecification. This estimator may be used to construct pointwise confidence intervals for $\theta^*(\beta)$ at fixed β using the pointwise asymptotic normality of $\hat\theta(\beta)$. However, for the testing and confidence band procedures described below, the complexity of the limiting distribution across β is prohibitive for conducting inference, even with variance estimation. For such scenarios, we suggest resampling to approximate the distribution of the estimator.

It can easily be shown that the regularity conditions are satisfied for the simple normal example. Interestingly, $\hat\theta(\beta)-\theta^*(\beta)=\bar Y-\theta_0-\beta_0$, which does not depend on β. This greatly simplifies the results of theorem 1, since the standardized estimators are identical for all β, which is not generally true. One should note that the form of the mean model is critical. If we assumed that $E(Y)=\theta\beta$, then the eigenvalue condition, C3, would be violated at $\beta=0$ and the uniform convergence in theorem 1 would fail to hold on intervals containing zero.

2.3. Global sensitivity testing

Suppose we are interested in evaluating the null hypothesis: $\mathbf{H}_0: C\theta_0=c$, where θ_0 is the true value of θ and C an $r\times\dim(\theta_0)$ contrast matrix for assessing single and multiple linear combinations of model parameters. For example, when testing the jth component of θ, one takes C to be $1\times\dim(\theta)$ vector with a one at the jth position and zeros elsewhere. Under non-identifiability, the above hypothesis cannot be tested without imposing unverifiable restrictions. If the true sensitivity parameter β_0 is known, then $\mathbf{H}_0: C\theta^*(\beta_0)=c$, where $\theta^*(\beta_0)=\theta_0$.

In practice, where β_0 is unknown, one may consider the process $\theta^*(\beta)$, observing that the trivial inequality,

$$0\le\inf_{\beta\in\Xi}\|C\theta^*(\beta)-c\|\le\|C\theta^*(\beta_0)-c\|\le\sup_{\beta\in\Xi}\|C\theta^*(\beta)-c\|,$$

permits a conservative assessment of $\mathbf{H}_0$. To do so, we formulate the infimum hypothesis: $\mathbf{H}_{\inf}:\inf_{\beta\in\Xi}\|C\theta^*(\beta)-c\|=0$.

The infimum statistic $\mathcal{T}_{\inf}=\inf_{\beta\in\Xi}\|C\hat\theta(\beta)-c\|$ can be used to evaluate this hypothesis. The distribution of this statistic can be derived analytically in some simple situations. As an example, we revisit the normal scenario discussed earlier where the interest is in evaluating the hypothesis, $\mathbf{H}_0:\theta_0=0$, using the processes $\hat\theta(\beta)=\bar Y-\beta$. For ease of illustration, assume $\Xi=[0,1]$, such that the infimum statistic becomes $\mathcal{T}_{\inf}=\inf_{\beta\in[0,1]}|\bar Y-\beta|$. This is a mixture of a point mass at 0 with probability $\Pr(\bar Y\in[0,1])$ and two truncated normal distributions. Specifically,

$$\inf_{\beta\in[0,1]}|\bar Y-\beta|=\begin{cases}-\bar Y & \text{if } \bar Y<0,\\ 0 & \text{if } \bar Y\in[0,1],\\ \bar Y-1 & \text{if } \bar Y>1.\end{cases}$$

The corresponding cumulative distribution function-CDF $F_{\inf}(x) = \Pr(\inf_{\beta\in[0,1]} |\bar{Y} - \beta| \le x)$ is

$$F_{\inf}(x) = \Pr(\bar{Y} \le x+1) - \Pr(\bar{Y} \le -x), \text{ for } x \ge 0. \quad (1)$$

In particular, we have $F_{\inf}(0) = \Pr(\bar{Y} \le 1) - \Pr(\bar{Y} \le 0) = \Pr(\bar{Y} \in [0,1])$, reflecting the point mass at 0 for $\mathcal{T}_{\inf}$.

In general, because of the complexity of the limiting distribution of the infimum of the test process, simple general analytic results do not appear tractable. Instead, resampling may be utilized. A simple non-parametric bootstrap (Efron & Tibshirani, 1993) may be used to compute variance estimators, and to carry out the simultaneous inferences necessary for the infimum tests and the confidence bands, described below. The validity of the bootstrap follows automatically from empirical process theory under the regularity conditions given in van der Vaart & Wellner (2000b) even under model misspecification. This requires the boundedness of the estimating function for fixed $\beta \in \Xi$. A difficulty with the non-parametric bootstrap is that it requires solving the estimating function for all β in each bootstrap sample, which may be computationally demanding. An alternative resampling technique which does not require repeatedly solving the estimating function may be constructed. The basic idea is to generate realizations directly from the limiting distribution of $\hat{\theta}(\beta)$ and to use these realizations to approximate the distribution of the infimum test and confidence bands. This resampling technique has been extensively used in the literature when the true asymptotic distribution is hard if not impossible to derive analytically (see for example, Parzen *et al.*, 1994 and Zhu & Zhang, 2006). To do this, one fixes the estimator based on the observed data and then 'perturbs' this estimator using a disturbance which conditionally on data has mean zero and variance-covariance in β equalling that of $\hat{\theta}(\beta)$ in theorem 1. The procedure is given by the following steps:

Step 1. Generate n i.i.d. random variables from a standard normal model ζ, denoted $\{\zeta_1^{(b)}, \ldots, \zeta_n^{(b)}\}$, where superscript (b) represents replications.

Step 2. Given the realizations of the data, $\{Y_i\}_{i=1}^n$, and values of $\beta \in \Xi$, calculate $\tilde{\theta}^{(b)}(\beta)$ using the simulated $\{\zeta_1^{(b)}, \ldots, \zeta_n^{(b)}\}$ and the equation,

$$\tilde{\theta}^{(b)}(\beta) = \hat{\theta}(\beta) + \left[n^{-1} \sum_{i=1}^{n} s_{Y_i}(\hat{\theta}(\beta), \beta) \zeta_i^{(b)} \right] W_Y^{-1}(\hat{\theta}(\beta), \beta), \quad (2)$$

where the statistic $\hat{\theta}(\beta)$ takes value $\hat{\theta}^o(\beta)$ for observed data $\{Y_i\}_{i=1}^n$.

Step 3. Calculate $\mathcal{T}_{\inf}^{(b)} = \inf_{\beta\in\Xi} \| C\tilde{\theta}^{(b)}(\beta) - c \|$ using $\tilde{\theta}^{(b)}(\beta), \beta \in \Xi$.

By repeatedly generating the normal variates $\{\zeta_j\}_{j=1}^n$, B times, and repeating steps 2 and 3 for each generated sample, we obtain the empirical distribution of $\mathcal{T}_{\inf}^{(b)}$ given observed data. Theorem 2 below establishes that this empirical distribution converges to the marginal asymptotic distribution of $\mathcal{T}_{\inf}$ as $n \to \infty$. Let $\mathbf{1}(\mathcal{E})$ be the indicator function for event $\mathcal{E}$. The p-value of the test is then $B^{-1} \sum_{b=1}^{B} \mathbf{1}(\mathcal{T}_{\inf}^{(b)} \ge \mathcal{T}_{\inf}^{o})$, the proportion of these bootstrap observations which exceed $\mathcal{T}_{\inf}^{o}$ the observed value of the statistic.

For the simple normal example, we compare the resampling null distribution of $\mathcal{T}_{\inf}^{(b)}$ to the analytical distribution $F_{\inf}(\cdot)$ in (1) for a finite sample size. Setting $\theta_0 = 0$ under the null and $\beta_0 = 0.5$, we generate $\{Y_i\}_{i=1}^n$ from a normal distribution $\mathcal{N}(0.5, 10^2)$. Furthermore, we take $\Xi = [0,1]$ and for each resample $b = 1, \ldots, B$, we compute $\mathcal{T}_{\inf}^{(b)} = \inf_{\beta\in[0,1]} |\tilde{\theta}^{(b)}(\beta)|$, where $\tilde{\theta}^{(b)}(\beta) = \bar{Y} - \beta - n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) \zeta_i^{(b)}$. Results with $n = 100$ and $B = 10{,}000$ resamples are plotted in Fig. 2. The resampling distribution provides a good approximation to the analytical distribution for this simple hypothetical example.

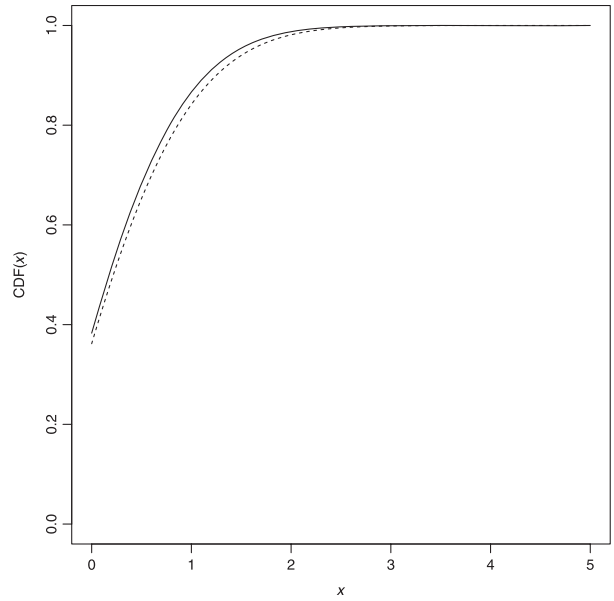

Fig. 2. Plot of the exact (solid line) and the resampled (dashed line) CDF ($\mathrm{CDF}(x) = \Pr(\inf_{\beta\in[0,1]} |\bar{Y} - \beta| \le x)$) of the infimum test statistic under the null $\theta_0 = 0$ for the simple normal example, assuming the true parameter $\beta_0 = 0.5$, sample size $n = 100$ and $B = 10{,}000$ resamples.

If the infimum (null) hypothesis cannot be rejected, then a supremum test or equivalently a simultaneous confidence region may be used to check whether $\|C\theta^*(\beta) - c\| > 0$ in some regions of Ξ. The supremum hypothesis $\mathbf{H}_{\sup}$ may be tested with the statistic $\mathcal{T}^o_{\sup} = \sup_{\beta\in\Xi} \|C\hat{\theta}(\beta) - c\|$ using the bootstrap realizations of $\tilde{\theta}^{(b)}(\beta), \beta \in \Xi$. The p-value of the supremum test is then $B^{-1}\sum_{b=1}^{B} \mathbf{1}(\mathcal{T}^{(b)}_{\sup} \ge \mathcal{T}^o_{\sup})$, where $\mathcal{T}^{(b)}_{\sup}$ are the bootstrap realizations of the statistic. Alternatively, a simultaneous confidence region for $C\theta^*(\beta) - c$ across all values of β may be constructed. Let $0 < \varphi < 1$. A simultaneous confidence region for $C\theta^*(\beta) - c$, $\beta \in \Xi$ is given by $\{\vartheta(\beta) : \Xi \to R^r;\ \|\vartheta(\beta) - C\hat{\theta}(\beta) + c\| < \rho_\varphi\}$, where ρ_φ is the $(1-\varphi)$th empirical percentile of $\{\sup_{\beta\in\Xi} \|C\tilde{\theta}^{(b)}(\beta) - C\hat{\theta}^o(\beta)\|\}_{b=1}^{B}$, with $\hat{\theta}^o(\beta)$ being the value of the statistic $\hat{\theta}(\beta)$ for observed data $\{Y_i\}_{i=1}^n$.

Theorem 2 supports the validity of the resampling based infimum test and confidence bands.

Theorem 2. *Under conditions* C1–C4, *the conditional distribution of the process* $n^{1/2}\{\tilde{\theta}(\beta) - \hat{\theta}^o(\beta)\}$ *given realizations* $\{Y_i\}_{i=1}^n$ *of* Y, *is asymptotically equivalent to the unconditional distribution of the process* $n^{1/2}\{\hat{\theta}(\beta) - \theta^*(\beta)\}$, $\beta \in \Xi$.

Theorem 2 (proof provided in the Appendix) coupled with a continuous mapping theorem gives that the infimum and supremum tests can be carried out using this resampling procedure. For the simple normal example, $n^{1/2}\{\hat{\theta}(\beta) - \theta^*(\beta)\} = n^{1/2}(\bar{Y} - \theta_0 - \beta_0)$ and $n^{1/2}\{\tilde{\theta}(\beta) - \hat{\theta}^o(\beta)\} = -n^{-1/2}\sum_{i=1}^n (Y_i - \bar{Y})\zeta_i$, which do not depend on β. The random quantity $n^{1/2}(\bar{Y} - \theta_0 - \beta_0)$ is normally distributed with mean 0 and variance 1, uniformly in β, for each fixed n. Given observed data $\{Y_i\}_{i=1}^n$, $-n^{-1/2}\sum_{i=1}^n (Y_i - \bar{Y})\zeta_i$ is also normally distributed with mean 0 and variance $n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ which converges almost surely to 1 as $n \to \infty$.

The choice of the support Ξ of β is critically important in performing the test in practice. If values of β are selected in some data-driven fashion, the limiting distribution in theorem 1 will be invalid. This is similar to Hansen (1996) for the case where the model is identifiable under the null after profiling on β, that is, when $\theta^*(\beta)=\theta_0$, $\forall\beta\in\Xi$. On the other hand, an approach which ignores sample information about Ξ may be unnecessarily conservative and potentially sacrifices power. One possible solution is to consult with subject-matter experts on the choice of Ξ. This choice ideally should be based on prior studies, as in the breast cancer analysis in section 3, where closely related datasets were used to select the range for the sensitivity parameter. From a technical standpoint, this choice should also be computationally feasible.

3. Numerical studies

3.1. Application to a pseudo-likelihood model for missing data in longitudinal studies

We consider the data set-up and model described in Troxel *et al.* (1998b) for potentially non-random missing data in longitudinal studies. The model will be referred to as the TLH model. The data arise from a longitudinal study where each subject $i\,(i=1,\ldots,n)$, is to be observed at K occasions. For subject i, we have a $K\times 1$ response vector, $Y_i^*=(Y_{i1}^*,\ldots,Y_{iK}^*)^{\mathrm{T}}$ which may not be fully observed. To accommodate missingness, subject i has a vector of missing data indicators $R_i=(R_{i1},\ldots,R_{iK})^{\mathrm{T}}$, where $R_{it}=1$ if Y_{it}^* is observed and 0 otherwise. Let $Y_{i,\mathrm{obs}}^*$ and $Y_{i,\mathrm{miss}}^*$ denote the observed and missing components of Y_i^*, respectively. Each individual also has a $K\times J$ covariate matrix X_i, which is assumed fully observed. The response Y_i in our general formulation is $\{Y_{i,\mathrm{obs}}^*, R_i, X_i\}$.

The key idea of the TLH methodology is to model the time point pair (Y_{it}^*, R_{it}), without accounting for the dependence on other time points. Let $f(u\,|\,w)$ denote the density function of random quantity u conditional on possibly non-random quantity w. We assume a simple selection model given by, $f(Y_{it}^*, R_{it}\,|\,X_{it},\varpi)=f(Y_{it}^*\,|\,X_{it},\varpi)f(R_{it}\,|\,Y_{it}^*,X_{it},\varpi)$, where ϖ is a finite but unknown parameter and X_{it} may contain both time dependent and independent covariates.

The TLH model assumes that density $f(Y_{it}^*\,|\,X_{it},\varpi)$ is that of normal $\mathcal{N}(\mu_{it},\sigma_t)$, where μ_{it} and $\sigma_t\,(t=1,\ldots,K)$ are elements of ϖ. The missing data process is assumed to satisfy $R_{it}\sim Bernoulli(1-\pi_{it})$ where the failure probability $\pi_{it}=\Pr(R_{it}=0\,|\,Y_{it}^*,X_{it},\varpi)$. We assume a logistic regression model relating the missing data probability to potentially unobserved responses, that is,

$$\mathrm{logit}(\pi_{it})=\gamma_{0t}+\gamma_{1t}X_{it}+\beta_t Y_{it}^*, \tag{3}$$

where γ_{jt} and $\beta_t\,(j=0,1;t=1,\ldots,K)$ are unknown parameters and elements of ϖ. The parameter β_t measures the extent of non-randomness of the missing data mechanism in the study at time t. Specifically, $\exp\{\beta_t\}$ represents the odds ratio for missing response at time t for each additional unit increase of the hypothetical response Y_{it}^*. Here, π_{it} in (3) depends on Y_{it}^* and not on previous elements of Y_i^*. Following warnings by Troxel *et al.* (1998b), we emphasize that this model could suffer from misspecification if the approximation of the logistic link function to the true link function fails.

The TLH model lends itself to a pseudo-likelihood analysis (Gong & Samaniego, 1981), where the longitudinal association is naively ignored in the likelihood construction. Specifically, the independence pseudo-likelihood function based on observed data $\{Y_{i,\mathrm{obs}}^*, R_i, X_i\}_{i=1}^n$ is

$$\begin{aligned}
\ell_{\text{ind}}(\varpi) &= \prod_{i=1}^{n} f(Y^*_{i,\text{obs}}, R_i \mid \varpi) = \prod_{i=1}^{n} \int \cdots \int f(Y^*_{i,\text{obs}}, Y^*_{i,\text{miss}}, R_i \mid \varpi)\, \mathrm{d}Y^*_{i,\text{miss}} \\
&= \prod_{i=1}^{n} \prod_{t=1}^{K} \{f(Y^*_{it}, R_{it} \mid \varpi)\}^{R_{it}} \left\{ \int f(Y^*_{it}, R_{it} \mid \varpi)\, \mathrm{d}Y^*_{it} \right\}^{1-R_{it}} \\
&= \prod_{i=1}^{n} \prod_{t=1}^{K} \{f(Y^*_{it} \mid \varpi) f(R_{it} \mid Y^*_{it}, \varpi)\}^{R_{it}} \left\{ \int f(Y^*_{it} \mid \varpi) f(R_{it} \mid Y^*_{it}, \varpi)\, \mathrm{d}Y^*_{it} \right\}^{1-R_{it}} \\
&= \prod_{i=1}^{n} \prod_{t=1}^{K} \{f(Y^*_{it} \mid \varpi)(1-\pi_{it})\}^{R_{it}} \left\{ \int f(Y^*_{it} \mid \varpi) \pi_{it}\, \mathrm{d}Y^*_{it} \right\}^{1-R_{it}}.
\end{aligned}$$

We have suppressed the dependence on covariates in $\ell_{\text{ind}}(\varpi)$. As a pseudo-likelihood model, conditions C1–C4 are easily verified and the asymptotic results hold. The densities in the TLH model are normal and Bernoulli, which are smooth functions of the unknown parameters.

3.2. Real data analysis

To illustrate our methodology, we consider data from the International Breast Cancer Study Group-IBCSG, previously reported by Hürny *et al.* (1992); and Troxel *et al.* (1998b). This is a group of randomized breast cancer studies with primary endpoints being survival and relapse; and quality of life being a secondary endpoint. One study, Study VI, is a randomized trial of adjuvant chemotherapy following surgery for the treatment of breast cancer. In this study, four treatments (A, B, C and D) were randomly assigned to 431 pre-menopausal cancer patients and several domains of quality of life were assessed. In this paper, we focus on three quality-of-life domains; (i) PACIS (perceived adjustment to chronic illness scale), (ii) Mood and (iii) Appetite. These variables were originally measured on a 0–100 scale, but are normalized using a square-root transformation as recommended by Troxel *et al.* (1998b). Questionnaires for the quality of life assessment were administered to study patients at baseline and every three months for two years. Our analysis employs the first three time points, with rates of missing data equalling 16 per cent, 33 per cent and 37 per cent for PACIS, 16 per cent, 33 per cent and 38 per cent for Mood, and 15 per cent, 33 per cent and 38 per cent for Appetite. A full description of Study VI and other IBCSG trials may be found elsewhere (Hürny *et al.*, 1992; Troxel *et al.*, 1998a).

As in earlier analyses of Study VI, we consider the following model for the measurement outcome,

$$\mu_{it} = \mu_{0t} + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i}, \quad (t = 1, 2, 3),$$

where μ_{0t} is a time-dependent intercept and α_j is a slope associated with $X_{ji}, j = 1, 2, 3$. Here $(X_{1i}, X_{2i}, X_{3i}) = \begin{cases} (1,0,0) & \text{if treatment A}, \quad (0,0,1) \quad \text{if treatment C} \\ (0,1,0) & \text{if treatment B}, \quad (0,0,0) \quad \text{if treatment D}. \end{cases}$

The missing data model is

$$\text{logit}(\pi_{it}) = \gamma_{0t} + \beta Y^*_{it} \quad (i = 1, \ldots, 431; t = 1, 2, 3),$$

where γ_{0t} is a time-dependent intercept and β is a slope associated with Y^*_{it}. As discussed previously, β quantifies the non-randomness of the missing data process. A constant σ_t is assumed across time.

Our objective is to assess the treatment and time effects on the mean quality of life. Under the assumed model, the hypotheses of interest are $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and $\mu_{01} = \mu_{02} = \mu_{03}$ for the

Table 1. *p-values for evaluating the Treatment and Time effects using data from Study VI of the IBCSG trials*

	Treatment effect				Time effect			
	Wald test				Wald test			
Responses	(TLH)	COM†	IGN‡	Infimum test	(TLH)	COM†	IGN‡	Infimum test
PACIS	0.281	0.170	0.303	0.231(0.522*)	0.216	0.302	0.142	0.134 (<0.001*)
Mood	0.008	0.041	0.011	0.025	0.552	0.030	0.006	0.063 (<0.001*)
Appetite	0.229	0.370	0.376	0.281(0.369*)	<0.001	0.163	0.023	<0.001

**p*-value of supremum test.
†Complete cases.
‡Ignorable cases.

treatment and time effects, respectively. As a preliminary analysis, we first evaluated these hypotheses under identifiability assumptions. Specifically, we fit the TLH model by simultaneously estimating both β and $\theta=(\alpha_1, \alpha_2, \alpha_3, \mu_{01}, \mu_{02}, \mu_{03}, \gamma_{01}, \gamma_{02}, \gamma_{03})$ via the independence pseudolikelihood estimating function. A Wald test based on the sandwich estimator of the covariance matrix of the regression parameter estimates was performed to evaluate the hypotheses of interest. *p*-values of these Wald tests for the three responses are given in Table 1. In brief, these inferences suggest that there is no treatment effect on PACIS and Appetite and no time effect on PACIS and Mood. The treatment effect on Mood and the time effect on Appetite are significant at 1 per cent level. In addition to these analyses, we also conducted two crude analyses that do not explicitly model the missing data mechanism. The first analysis used only subjects with complete data sequences, therefore removing subjects with incomplete data profiles. The second analysis ignored the missing data and conduct the so-called ignorable (missing at random) inferences by forcing β, the non-randomness missing data parameter, to zero. Results of these analyses are also summarized in Table 1. From these additional exploratory analyses, the treatment and time effects are found to be statistically significant for Mood at 5 per cent level. The ignorable analysis also appears to yield a statistically significant time effects on Appetite. Of course, these crude analyses may not be reliable as they rely on assumptions that are not verifiable using observed data at hand.

The Wald tests conducted under the assumption of identifiability may not have desirable properties if identifiability is violated. As illustrated in Fig. 1, the model was at best weakly-identifiable for the outcome PACIS. Model identifiability was also a concern for β for the other two responses. We performed the infimum test to conservatively evaluate the treatment and time effects on the three quality of life domains. To conduct these tests, the set Ξ for the range of β was obtained from an independent source. We considered data on post-menopausal cancer patients from Study VII of the IBCSG trials. Objectives of this study were similar to those of Study VI, except that the menopausal status of study participants differed. The joint model appeared to be identifiable when applied to Study VII data. Based on these results, we derived 99 per cent confidence intervals to use as ranges for β in the infimum tests for Study VI. The ranges for PACIS, Mood and Appetite were $[-4, 0]$, $[-3, 0]$ and $[-5.6, -1.6]$, respectively. Recall that in the missing data model, $\exp\{-\beta\}$ represents the odds ratio of being observed at any time point for each additional unit increase of the hypothetical response Y_{it}^*. Since Y_{it}^* takes values in the range 0 – 10 on a square-root scale, for the selected ranges, the odds ratio may be as high as: $\exp\{4\}=54.60$ for PACIS, $\exp\{3\}=20.09$ for Mood, and $\exp\{5.6\}=270.43$ for Appetite. One might criticize these upper bounds as being scientifically unreasonable. However, permitting such extreme scenarios provides for a conservative test, which is in the spirit of sensitivity analysis. For computational feasibility,

the ranges were approximated on fine grids with equally spaced points of 0.02. p-values of the infimum tests are given in Table 1.

The infimum hypothesis for the treatment effect was rejected for Mood at the 5 per cent level (p-value = 0.025), but not for PACIS (p-value = 0.281) and Appetite (p-value = 0.231). For the time variable, a strongly significant effect was detected only for Appetite (p-value < 0.001). For non-significant infimum test results, a supremum test was conducted to see if one could not reject the null hypothesis for all values $\beta \in \Xi$. The supremum test for the treatment effect was not rejected on PACIS (p-value = 0.522) and Appetite (p-value = 0.369), but was strongly rejected for the time effect on PACIS (p-value < 0.001) and Mood (p-value < 0.001).

When the supremum test was rejected, a sensitivity analysis was conducted using a simultaneous 95 per cent confidence band approach to identify regions of β for which the pointwise null hypotheses are rejected. Plots of these analyses for contrasts $\mu^*_{01}(\beta) - \mu^*_{02}(\beta)$ and $\mu^*_{02}(\beta) - \mu^*_{03}(\beta)$ for PACIS and Mood are given in Fig. 3. For PACIS, the 95 per cent simultaneous confidence band for $\mu^*_{02}(\beta) - \mu^*_{01}(\beta)$, $-4 \le \beta < -0.4$; and $\mu^*_{03}(\beta) - \mu^*_{02}(\beta)$, $-4 \le \beta \le -3$, did not contain 0. Similar analyses for Mood revealed that a 95 per cent simultaneous confidence band for $\mu^*_{02}(\beta) - \mu^*_{01}(\beta)$, $-3 \le \beta < -0.7$, did not contain 0. The confidence band for $\mu^*_{03}(\beta) - \mu^*_{02}(\beta)$ did not exclude 0 over the selected range of β ($-3 \le \beta < 0$). The Wald tests which assume identifiability were non-significant for all pairwise comparisons at the 5 per cent level.

3.3. Simulation study

Here, we report results of a simulation study comparing the performance of the infimum test to that of the naive Wald test derived under identifiability assumptions. The simulations were conducted under a TLH model specified so as to roughly approximate data from Study

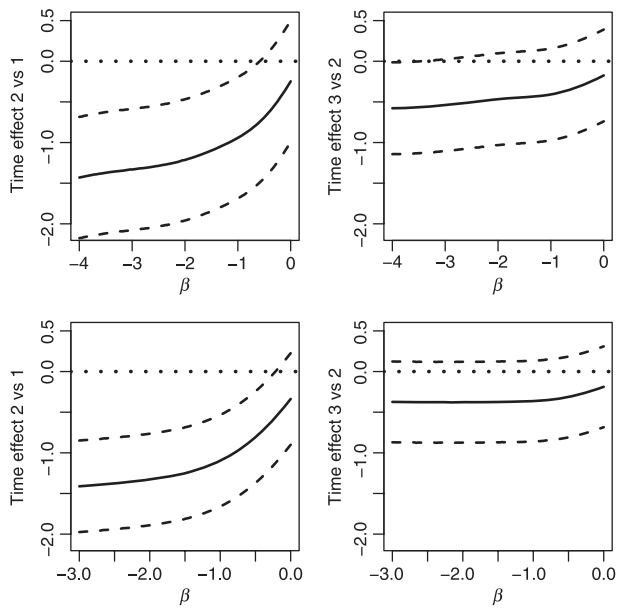

Fig. 3. The top panel corresponds to PACIS; and the bottom panel to Mood. In each panel, the solid lines represent $\hat{\mu}_{02}(\beta) - \hat{\mu}_{01}(\beta)$ (on the left) and $\hat{\mu}_{03}(\beta) - \hat{\mu}_{02}(\beta)$ (on the right) for fixed values of the parameter β. The dashed lines are the corresponding 95 per cent simultaneous confidence bands and the dotted lines are the null values.

VI of the IBCSG trials. For simplicity, only two treatments (A and B) and two time points ($K=2$) were considered. The outcome vector (Y_{i1}^*, Y_{i2}^*), assuming dependence on subject i, was generated from a two-dimensional normal distribution with univariate mean models,

$$\mu_{it} = \mu_{0t} + \alpha_t X_{1i}, \quad t = 1, 2,$$

and time-point variances $\sigma_t, t=1,2,$ and correlation coefficient ρ. The parameters μ_{0t} and α_t are time-dependent intercepts and slopes associated with covariate X_{1i}, which equals 1 if treatment B and 0 otherwise. We reparameterized μ_{0t} and α_t as, $\mu_{0t} = \tilde{\alpha}_0 + \tilde{\alpha}_1 I(t=2)$ and $\alpha_t = \tilde{\alpha}_2 + \tilde{\alpha}_3 I(t=2)$, where $I(t=2)$ is an indicator variable taking value 1 at the second time point. Throughout our simulations, we fixed the variances σ_t, $t=1,2,$ to 1 and the correlation coefficient ρ to 0.4. Missing observations were generated using a logistic model relating the dropout probability π_{it} to the response Y_{it}^* as,

$$\text{logit}(\pi_{it}) = \gamma_{0t} + \gamma_{1t} X_{1i} + \beta Y_{it}^*,$$

where γ_{0t}, γ_{1t} and β are respectively the intercept and slopes associated with X_{1i} and Y_{it}^*. Time-dependent parameters γ_{0t} and γ_{1t} were reparameterized as, $\gamma_{0t} = \tilde{\gamma}_0 + \tilde{\gamma}_1 I(t=2)$ and $\gamma_{1t} = \tilde{\gamma}_2 + \tilde{\gamma}_3 I(t=2), t=1,2$.

We study the size and power of the infimum and Wald tests for $\tilde{\alpha}_3$, the parameter that captures the interaction effect of time and treatment on the mean response. We set $\tilde{\alpha}_3 = 0$ and $\tilde{\alpha}_3 = 1$ for the size and power of the test, respectively. Additionally, $(\tilde{\alpha}_1, \tilde{\alpha}_2) = (0, 0)$ and $(\tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\gamma}_3) = (0.5, -2, 0.2)$ when evaluating size, and $(\tilde{\alpha}_1, \tilde{\alpha}_2) = (0.1, 1)$ and $(\tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\gamma}_3) = (1, -3, 1)$ when evaluating power.

The parameter $\tilde{\gamma}_0$ was varied throughout our simulations to produce different missing data rates. Specifically, to study the size of the test $\tilde{\gamma}_0$ was fixed to 0.5, to produce about 15 per cent and 22 per cent missing observations at the first and second time point, respectively, and at 1.8 to produce about 33 per cent and 43 per cent missing observations at the first and second time point, respectively. For the power, $\tilde{\gamma}_0$ was fixed to 0.5, producing rates of missing observations roughly 14 per cent and 26 per cent at the first and second time point, respectively, and to 2, producing rates of missing observations roughly 32 per cent and 46 per cent at the first and second time point, respectively. Finally, throughout our simulations, we set the true β to -1.

One thousand datasets were generated with sample sizes 100 and 300. Equal proportions of subjects were assigned to treatment A and B. The infimum tests were performed on the interval $\Xi = [-2, 0]$. To ensure computational feasibility, a fine grid of equally spaced points of 0.02, was considered. We used 1000 resamples from the alternative resampling scheme discussed in section 2.3 to approximate the null distribution of the infimum test.

The infimum and Wald tests were performed using working regression models having the same form as those used to generate data. These models saturate the number of parameters, leading to potential non-identifiability as a result of overparameterization. Table 2 shows the rejection rates for nominal test levels 0.01, 0.05 and 0.1. Asymptotic standard errors (as the number of Monte Carlo iterations tends to infinity) are reported in the last row of the table. Overall, the infimum tests perform well, with the resampling distribution of the test providing a reasonable approximation to the nominal level. The Wald test appears to be very liberal when compared to the infimum test. The anti-conservativeness of the Wald test does not diminish as the sample size increases. Based on these results, our recommendation is to avoid the Wald test when identifiability is of concern. Because the empirical type I error rate of the infimum test and that of the Wald test are different, comparing their empirical powers is not appropriate. Nevertheless for both methods, a larger sample size improves the power of

Table 2. *Empirical type I error and power of the infimum test* and Wald test (in parenthesis) for evaluating the interaction effect represented by the parameter* $\tilde{\alpha}_3$

True value	Missing† data rate	$n=100$ Nominal test level 0.1	0.05	0.01	$n=300$ Nominal test level 0.1	0.05	0.01
$\tilde{\alpha}_3=0$	15%, 22% ($\tilde{\gamma}_0=0.5$)	0.099 (0.147)	0.049 (0.092)	0.011 (0.034)	0.103 (0.158)	0.057 (0.117)	0.010 (0.059)
	33%, 43% ($\tilde{\gamma}_0=1.8$)	0.110 (0.149)	0.044 (0.097)	0.010 (0.051)	0.122 (0.165)	0.062 (0.113)	0.012 (0.073)
$\tilde{\alpha}_3=1$	14%, 26% ($\tilde{\gamma}_0=0.5$)	0.982 (0.961)	0.959 (0.944)	0.870 (0.848)	>0.999 (0.948)	>0.999 (0.942)	>0.999 (0.934)
	32%, 46% ($\tilde{\gamma}_0=2$)	0.905 (0.889)	0.846 (0.827)	0.675 (0.659)	0.999 (0.961)	0.999 (0.949)	0.993 (0.926)
Monte Carlo SE		0.003	0.007	0.009	0.003	0.007	0.009

*Test performed using $\Xi=[-2,0]$.
†First and second time point missing data rate.

Table 3. *Empirical power of the infimum test to detect the interaction effect* $\tilde{\alpha}_3=1$ *for two ranges* Ξ *of* β *with true value being* $\beta_0=-1$

Ξ	Missing* data rate	$n=100$ Nominal test level 0.10	0.05	0.01	$n=300$ Nominal test level 0.10	0.05	0.01
$[-3,3]$	14%, 26%	0.912	0.865	0.720	0.999	0.997	0.987
	32%, 46%	0.783	0.706	0.529	0.981	0.960	0.905
$[-5,5]$	14%, 26%	0.899	0.836	0.655	0.997	0.990	0.970
	32%, 46%	0.767	0.695	0.492	0.953	0.925	0.858
Monte Carlo SE		0.003	0.007	0.009	0.003	0.007	0.009

*First and second time point missing data rate.

detecting the alternatives under consideration, a finding consistent with the literature. Moreover, the power decreases with increasing missing data rates.

While the ability to choose an appropriate support set Ξ of β to perform the infimum tests is highly desirable in practice, our simulations (results not shown) indicate that only a minimal inflation of type I error rate is observed under a modest misspecification of the set Ξ. For example, when Ξ does not contain the true β, but β_0 is not far away from the boundaries of the set, close to the nominal level is still achieved under the null hypothesis. As an example, we performed the infimum test on the interval $[0,2]$, which does not contain $\beta_0=-1$. For this range of β, the infimum tests nearly maintain their sizes at all significance levels. However, when $[10,12]$ was selected for the range of β, the infimum tests were overly anti-conservative.

Another simulation study was conducted to evaluate the effects of the choice of the set Ξ on the power of the infimum tests. Specifically, we generated data as before, but performed the infimum tests on wider intervals, namely $[-3,3]$ and $[-5,5]$. Results of this simulation study are given in Table 3. As expected, the power decreases as the interval widens, which occurs regardless of the missing data rate. Following a referee's recommendation, further simulations were conducted to evaluate the loss of power when the infimum test is performed on a given support set of β compared to the ideal set $\Xi=\{\beta_0\}$. For this, we generated the data as before with the only difference that $\tilde{\alpha}_3=0.7$. We then performed the infimum test using $\Xi=[-2,0]$ and $\Xi=\{-1\}$. Results revealed a minor loss of power of the infimum test on $\Xi=[-2,0]$ compared to the ideal set $\Xi=\{-1\}$ (see Table 4).

Table 4. *Empirical power of the infimum test to detect the interaction effect* $\tilde{\alpha}_3 = 0.7$ *for two sets* Ξ *of* β *with true value being* $\beta_0 = -1$

		$n = 100$			$n = 300$		
	Missing*	Nominal test level			Nominal test level		
Ξ	data rate	0.10	0.05	0.01	0.10	0.05	0.01
$[-2, 0]$	13%, 23%	0.806	0.699	0.470	0.996	0.989	0.949
$\{-1\}$	13%, 23%	0.829	0.725	0.469	0.998	0.998	0.982
$[-2, 0]$	31%, 44%	0.677	0.570	0.342	0.978	0.957	0.834
$\{-1\}$	31%, 44%	0.707	0.589	0.348	0.992	0.974	0.915
Monte Carlo SE		0.003	0.007	0.009	0.003	0.007	0.009

*First and second time point missing data rate.

4. Discussion

While hypothesis testing under non-identifiability has been previously considered, the framework is often too restrictive for sensitivity analyses. In a sensitivity analysis, the model may not be identifiable under either the null or alternative hypothesis, and profiling may not lead to consistent estimation of the parameter of interest under the null. As a result, the supremum test may not be appropriate. As discussed in this paper, a theoretically rigorous approach to this testing problem may be based on infimum statistics, whose distribution must be carefully considered under model misspecification under the null hypothesis.

The infimum testing approach was previously studied for likelihood analyses of parametric models (Todem *et al.*, 2010). In this paper, we have extended these results to general estimating functions for parametric models. This includes limiting results for the profile estimators and the infimum test and confidence bands, as well as the validity of the bootstrap procedure. Such results are critically important in sensitivity analyses of complex data arising in longitudinal studies, where full model specification may be difficult and partially specified models may be more easily analyzed using non-likelihood based approaches.

Acknowledgements

The authors are grateful to Karen Price and Richard Gelber of IBCSG for their permission to use the quality-of-life data. The authors would like to thank the editor and assistant editor, as well as two anonymous referees for their constructive and insightful comments that greatly improved the manuscript. This work was partially supported by the second author's NCI/NIH K-award, its supplement from the 2009 American Recovery and Reinvestment Act funding mechanism and the third author's NSF awards and funding from Jiangsu Specially-Appointed Professor Program, Jiangsu, China.

References

Chambers, R. L. & Welsh, A. H. (1993). Log-linear models for survey data with non-ignorable non-response. *J. Roy. Statist. Soc., Series B: Statistical Methodology* **55**, 157–170.

Copas, J. (1999). What works?: Selectivity models and meta-analysis. *J. Roy. Statist. Soc., Ser. A: Statist. Soc*. **162**, 95–109.

Copas, J. B. & Li, H. G. (1997). Inference for non-random samples (Disc: P77-95). *J. Roy. Statist. Soc., B: Methodol*. **59**, 55–77.

Dacunha-Castelle, D. & Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.* **27**, 1178–1209.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43.
Efron, B. & Tibshirani, R. (1993). *An introduction to the Bootstrap.* Chapman & Hall Ltd, London.
Gong, G. & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.* **9**, 861–869.
Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **61**, 413–430.
Huber, P. J. (1967). The behavior of the maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley Symposium in mathematical Statistics and Probability* (eds L. M. De Cam & J. Neyman). University of California Press, Berkeley. 221–233.
Hürny, C., Bernhard, J., Gelber, R., Coates, A., Castiglione, M., Isley, M., Dreher, D., Peterson, H., Goldhirsch, A. & Senn, H. (1992). Quality of life measures for patients receiving adjuvant therapy for breast cancer: an international trial. *Eur.J. Cancer* **28**, 118–124.
Kenward, M. G., Goetghebeur, E. J. T. & Molenberghs, G. (2001). Sensitivity for incomplete categorical data. *Statist. Model.* **1**, 31–48.
Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons, New York.
Lu, G. & Copas, J. B. (2004). Missing at random, likelihood ignorability and model completeness. *Ann. Statist.* **32**, 754–765.
Parzen, M. I., Wei, L. J. & Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341–350.
Ritz, C. & Skovgaard, I. M. (2005). Likelihood ratio tests in curved exponential families with nuisance parameters present only under the alternative. *Biometrika* **92**, 507–517.
Rotnitzky, A., Scharfstein, D., Su, T.-L. & Robins, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57**, 103–113.
Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (C/R: P1121-1146). *J. Amer. Statist. Assoc.* **94**, 1096–1120.
Song, R., Kosorok, R. & Fine, J. (2009). On asymptotically optimal tests under loss of identifiability in semiparametric models. *Ann. Statist.* **37**, 2409–2444.
Todem, D., Fine, J. & Peng, L. (2010). A global sensitivity test for evaluating statistical hypotheses with nonidentifiable models. *Biometrics* **66**, 558–566.
Troxel, A. B., Harrington, D. P. & Lipsitz, S. R. (1998a). Analysis of longitudinal data with non-ignorable non-monotone missing values. *J. Roy. Statist. Soc. Ser. C: Appl. Statist.* **47**, 425–438.
Troxel, A. B., Lipsitz, S. R. & Harrington, D. P. (1998b). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* **85**, 661–672.
van der Vaart, A. W. & Wellner, J. A. (2000a). *Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli theorems. High dimensional probability II*. (eds E. Giné, D. M. Mason, J. A. Wellner). Birkhäuser, Boston. 115–133.
van der Vaart, A. W. & Wellner, J. A. (2000b). *Weak convergence and empirical processes*. Springer, New York.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 141–161.
Zhu, H. T. & Zhang, H. P. (2006). Generalized score test of homogeneity for mixed effects models. *Ann. Statist.* **34**, 1545–1569.

Received August 2011, in final form June 2012

David Todem, Division of Biostatistics, Department of Epidemiology and Biostatistics, 909 Fee Road, Room B601, Michigan State University, East Lansing, MI 48824, USA.
E-mail: todem@msu.edu

Appendix

Proof of theorem 1. (i) We show that $\sup_{\beta\in\Xi}\|\hat{\theta}(\beta)-\theta^{*}(\beta)\|\to_p 0$.

Condition C2 implies that $\mathcal{G}_1$ and $\mathcal{G}_2$ are Donsker and hence Glivenko-Cantelli (van der Vaart & Wellner, 2000a,b). Therefore,

$$\sup_{\theta\in\Theta,\,\beta\in\Xi}\|S_Y(\theta,\beta)-\tilde{S}(\theta,\beta)\|\to_p 0 \quad \text{and} \quad \sup_{\theta\in\Theta,\,\beta\in\Xi}\|W_Y(\theta,\beta)-\tilde{W}(\theta,\beta)\|\to_p 0. \tag{4}$$

The definitions of $\hat{\theta}(\beta)$ and $\theta^*(\beta)$ and condition C4 imply that

$$\begin{aligned}0 &= S_Y(\hat{\theta}(\beta),\beta) - \tilde{S}(\theta^*(\beta),\beta)\\ &= (S_Y(\hat{\theta}(\beta),\beta) - S_Y(\theta^*(\beta),\beta)) + (S_Y(\theta^*(\beta),\beta) - \tilde{S}(\theta^*(\beta),\beta))\\ &= \tilde{W}(\check{\theta}(\beta),\beta)(\hat{\theta}(\beta)-\theta^*(\beta)) + v_{2n}(\beta)(\hat{\theta}(\beta)-\theta^*(\beta)) + v_{1n}(\beta), \end{aligned} \tag{5}$$

where $\check{\theta}(\beta)$ is on the line segment between $\hat{\theta}(\beta)$ and $\theta^*(\beta)$. Also, $\sup_{\beta\in\Xi}|v_{2n}(\beta)| \to_p 0$ and $v_{1n}(\beta) = S_Y(\theta^*(\beta),\beta) - \tilde{S}(\theta^*(\beta),\beta)$. From (5), we have,

$$\begin{aligned}\sup_{\beta\in\Xi}\|\hat{\theta}(\beta)-\theta^*(\beta)\| &= \sup_{\beta\in\Xi}\|-(\tilde{W}^{-1}(\check{\theta}(\beta),\beta)+v_{2n}(\beta))v_{1n}(\beta)\|\\ &\le \sup_{\beta\in\Xi}\|-\tilde{W}^{-1}(\check{\theta}(\beta),\beta)+v_{2n}(\beta)\|\sup_{\beta\in\Xi}\|v_{1n}(\beta)\|.\end{aligned}$$

Because of condition C3, for any $\theta\in\Theta$, for any $\beta\in\Xi$, there exists a positive number λ_1, such that $\lambda_{\min}(\beta) > \lambda_1 > 0$. For any $s\times s$ symmetric matrix $\mathbf{A}$, denote its Euclidean norm as $\|\mathbf{A}\| = \lambda_{\max}(\mathbf{A})$, where $\lambda_{\max}(\mathbf{A})$ is the largest eigenvalue of $\mathbf{A}$ and if $\mathbf{A}$ is also non-singular, $\|\mathbf{A}^{-1}\| = \lambda_{\min}^{-1}(\mathbf{A})$. Therefore, $\|-\tilde{W}^{-1}((\check{\theta}(\beta),\beta)\| = \lambda_{\min}^{-1}(\beta) \le \lambda_1^{-1}$, and $\sup_{\beta\in\Xi}\|\hat{\theta}(\beta)-\theta^*(\beta)\| \le \lambda_1^{-1}\sup_{\beta\in\Xi}\|v_n(\beta)\|$. The uniform consistency of $\hat{\theta}(\beta)$ to $\theta^*(\beta)$ follows from $\sup_{\beta\in\Xi}\|v_n(\beta)\| = \sup_{\beta\in\Xi}\|S_Y(\theta^*(\beta),\beta)-\tilde{S}(\theta^*(\beta),\beta)\| \le \sup_{\theta\in\Theta,\beta\in\Xi}\|S_Y(\theta,\beta)-\tilde{S}(\theta,\beta)\| \to_p 0$, according to (4).

(ii) We show that $n^{1/2}(\hat{\theta}(\beta)-\theta^*(\beta))$ converge weakly to a tight Gaussian process.

Based on the uniform consistency of $\hat{\theta}(\beta)$, and (4) and (5), applying the Taylor expansion to $S_Y(\hat{\theta}(\beta),\beta)$ around $S_Y(\theta^*(\beta),\beta)$ gives

$$\begin{aligned}n^{1/2}(\hat{\theta}(\beta)-\theta^*(\beta)) &\approx -n^{-1/2}\tilde{W}^{-1}(\theta^*(\beta),\beta)v_{1n}(\beta)\\ &= -n^{-1/2}\sum_{i=1}^{n}\tilde{W}^{-1}(\theta^*(\beta),\beta)(s_{Y_i}(\theta^*(\beta),\beta) - Es_{Y_1}(\theta^*(\beta),\beta))\\ &= -n^{-1/2}\sum_{i=1}^{n}\tilde{W}^{-1}(\theta^*(\beta),\beta)s_{Y_i}(\theta^*(\beta),\beta) \equiv -n^{-1/2}\sum_{i=1}^{n}\eta_i(\beta),\end{aligned}$$

where $\approx$ denotes asymptotic equivalence uniformly in $\beta\in\Xi$. Because condition C2 implies that $\mathcal{G}_1$ is Donsker and using previous results that $\tilde{W}^{-1}(\theta^*(\beta),\beta)$ is uniformly bounded for $\beta\in\Xi$, the function class $\{\tilde{W}^{-1}(\theta^*(\beta),\beta)s_{Y_i}(\theta^*(\beta),\beta), \beta\in\Xi, i=1,\ldots,n\}$ is Donsker. This permits the application of a functional central limit theory to establish the weak convergence of $\hat{\theta}(\beta)$. Therefore, $\lim_{n\to\infty}\mathrm{cov}\{n^{1/2}(\hat{\theta}(\beta_1)-\theta^*(\beta_1)), n^{1/2}(\hat{\theta}(\beta_2)-\theta^*(\beta_2))\} = E(\eta_1(\beta_1)\eta_1^{\mathrm{T}}(\beta_2)) = \Sigma^*(\beta_1,\beta_2)$. For a given β, $\mathrm{var}\{n^{1/2}(\hat{\theta}(\beta)-\theta^*(\beta))\} = E\left(\eta_1(\beta)\eta_1^{\mathrm{T}}(\beta)\right)$.

Proof of theorem 2. Applying the Taylor expansion to $s_{Y_i}(\theta^*(\beta),\beta)$ around $s_{Y_i}(\hat{\theta}^o(\beta),\beta)$ gives

$$n^{-1/2}\sum_{i=1}^{n}s_{Y_i}(\theta^*(\beta),\beta)\zeta_i = n^{-1/2}\sum_{i=1}^{n}s_{Y_i}(\hat{\theta}^o(\beta),\beta)\zeta_i + n^{-1/2}(\theta^*(\beta)-\hat{\theta}^o(\beta))\sum_{i=1}^{n}g_{Y_i}(\bar{\theta}(\beta),\beta)\zeta_i,$$

where $\bar{\theta}(\beta)$ is on the line segment between $\hat{\theta}^o(\beta)$ and $\theta^*(\beta)$. Given observations $\{Y_i\}_{i=1}^n$, condition C4 and $\sup_{\beta\in\Xi}\|\theta^*(\beta)-\hat{\theta}^o(\beta)\| = o_p(1)$, one has

$$n^{-1/2}\sum_{i=1}^{n}s_{Y_i}(\theta^*(\beta),\beta)\zeta_i \approx n^{-1/2}\sum_{i=1}^{n}s_{Y_i}(\hat{\theta}^o(\beta),\beta)\zeta_i.$$

Based on the definition of $\tilde{\theta}(\beta)$ in (2) and (4), one has

$$
\begin{aligned}
n^{1/2}(\tilde{\theta}(\beta)-\hat{\theta}^{o}(\beta)) &= -n^{-1/2}W_Y^{-1}(\hat{\theta}^{o}(\beta),\beta)\sum_{i=1}^{n} s_{Y_i}(\hat{\theta}^{o}(\beta),\beta)\zeta_i \\
&\approx -n^{-1/2}W_Y^{-1}(\hat{\theta}^{o}(\beta),\beta)\sum_{i=1}^{n} s_{Y_i}(\theta^{*}(\beta),\beta)\zeta_i \\
&\approx -n^{-1/2}\tilde{W}^{-1}(\theta^{*}(\beta),\beta)\sum_{i=1}^{n} s_{Y_i}(\theta^{*}(\beta),\beta)\zeta_i \\
&= -n^{-1/2}\sum_{i=1}^{n}\eta_i(\beta)\zeta_i.
\end{aligned}
$$

Hence, conditional on observations $\{Y_i\}_{i=1}^{n}$, $n^{1/2}(\tilde{\theta}(\beta)-\hat{\theta}^{o}(\beta))$ converges weakly to a Gaussian process with mean 0 and covariance function $\Sigma^{o}(\beta_1,\beta_2)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E(\eta_i(\beta_1)\times\zeta_i\zeta_i^{\mathrm{T}}\eta_i^{\mathrm{T}}(\beta_2)\mid Y)$. We also have

$$
\begin{aligned}
\Sigma^{o}(\beta_1,\beta_2)-\Sigma^{*}(\beta_1,\beta_2) &= \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\left[E(\eta_i(\beta_1)\zeta_i\zeta_i^{\mathrm{T}}\eta_i^{\mathrm{T}}(\beta_2)\mid Y)\right]-E(\eta_1(\beta_1)\eta_1^{\mathrm{T}}(\beta_2)) \\
&= \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}(\eta_i(\beta_1)\eta_i^{\mathrm{T}}(\beta_2))-E(\eta_1(\beta_1)\eta_1^{\mathrm{T}}(\beta_2))=0.
\end{aligned}
$$

Hence, $\lim_{n\to\infty}\mathrm{cov}\{n^{1/2}(\tilde{\theta}(\beta_1)-\hat{\theta}^{o}(\beta_1)),n^{1/2}(\tilde{\theta}(\beta_2)-\hat{\theta}^{o}(\beta_2))\}=\Sigma^{*}(\beta_1,\beta_2)$, the conditional distribution of $n^{1/2}(\tilde{\theta}(\beta)-\hat{\theta}^{o}(\beta))$ is asymptotically equivalent to the unconditional distribution of $n^{1/2}(\hat{\theta}(\beta)-\theta^{*}(\beta))$.