Simulation-based Sensitivity Analysis for Non-ignorable Missing Data

Peng Yin and Jian Qing Shi[*]

School of Mathematics & Statistics, Newcastle University

November 19, 2018

Abstract

Sensitivity analysis is popular in dealing with missing data problems particularly for non-ignorable missingness. It analyses how sensitively the conclusions may depend on assumptions about missing data e.g. missing data mechanism (MDM). We called models under certain assumptions sensitivity models. To make sensitivity analysis useful in practice we need to define some simple and interpretable statistical quantities to assess the sensitivity models. However, the assessment is difficult when the missing data mechanism is missing not at random (MNAR). We propose a novel approach in this paper on attempting to investigate those assumptions based on the nearest-neighbour (KNN) distances of simulated datasets from various MNAR models. The method is generic and it has been applied successfully to several specific models in this paper including meta-analysis model with publication bias, analysis of incomplete longitudinal data and regression analysis with non-ignorable missing covariates.

Keywords: Bias parameter / sensitivity parameter, Bias model / Sensitivity model, Incomplete longitudinal data, Missing data mechanism, K- nearest-neighbour (KNN) distance, Non-ignorable missing data, Publication bias, Simulation-based sensitivity analysis

[*]School of Mathematics & Statistics, Newcastle University, Newcastle NE1 7RU, UK. Email: j.q.shi@ncl.ac.uk

1 Introduction

Model uncertainty and incomplete data are common issues in statistical analysis. An assessment of uncertainty due to incomplete data or model misspecification has attracted many researchers' attention for several decades (see e.g Cornfield et al., 1959; Vemuri et al., 1969; Draper, 1995; Copas and Li, 1997). Let θ be the parameter of interest and D a set of observations. Conventional inference employs a model $f(D;\theta)$ and this usually provides a consistent estimate for θ. However, if part of the data are unobserved, inference based on the observed data may result in bias. In this case bias analysis can be carried out by using a so-called bias model $f(D;\theta,\eta)$ associated with different missing data mechanisms (MDM) (see e.g. Greenland, 2005), where η is called the bias parameter. It is usually difficult to estimate η due to the lack of data or prior knowledge. For example, if the missingness is under so-called non-ignorable (or missing not at random) MDM (Little and Rubin, 2002) and η is the parameter involved in a MDM model, then η cannot be estimated as it depends on the unobserved data. In some cases it is possible to resort to follow-up studies in order to estimate η (see e.g. Kim and Yu, 2011); in many instances, however, such investigations are inherently difficult to conduct due to the nature of the study as it occurs, for example, in epidemiological studies. Also extra bias may exist due to the lack of randomization and the independence problem between former observations and follow-up samples.

Sensitivity analysis is one of the commonly used approaches for assessing uncertainty via a *bias parameter* η or the related *bias model*; hence η is also known as the *sensitivity parameter* and its associated model the *sensitivity model.* As such, we will refer to η the sensitivity parameter in the remainder of the paper. The use of these models is prolific in many and varied areas, for example, Copas (1999) and Copas and Shi (2000, 2001) to explore publication bias in meta-analysis using the Heckman model (Hechman, 1979), Molenberghs et al. (2001) to investigate incomplete contingency tables and Copas and Eguchi (2001, 2005) to look at local sensitivity analysis. Those discussions characterize the sensitivity analysis in different ways, but their aims are essentially the same: to examine the influence of an individual point on model-based inference. A different approach considers all possible sources of uncertainty by defining a prior density coupled with a Monte Carlo sensitivity analysis to sample '*bias* *parameters and then inverts the bias model to provide a distribution of bias-corrected estimates*' (Greenland, 2005, p.269). However it is usually difficult to choose and justify a prior density.

Copas (2013) pointed out that '*a sensitivity analysis is essentially a warning of how sensitively the conclusions of a meta-analysis may depend on key assumptions about the study selection process*'. Using the idea in a general sensitivity model $f(D;\theta,\eta)$, we need to analyse how the conclusions change when η changes. It is usually more useful for practitioner if we can find some simple and interpretable statistical quantities to assess the sensitivity models instead of using η directly. For example Copas and Shi (2001) used the P-value for the goodness of fit to the funnel plot and the estimated number of unpublished studies. In spite of being very useful, those methods are limited to special cases and it is not straightforward to extend the idea to other problems. In this paper we propose a generic method based on the nearest-neighbour distance between the observed data and the data simulated from different sensitivity models. We attempt to achieve two aims: (i) exclude the models which are 'far away' from the true model. We do so by comparing the KNN distances with a critical value based on a permutation test; (ii) in the set of plausible sensitivity models resulting from (i), find the most plausible model or a set of the most plausible models in terms of minimal distance. This will result in a set of most plausible estimates in a small range for the parameter of interest θ, which would provide useful information to help practitioners to draw conclusions. We will refer to this method as *simulation-based sensitivity analysis* (SSA).

The remainder of the paper is organized as follows. Section 2 will first describe the idea of simulation-based sensitivity analysis, followed by a detailed discussion on how to evaluate sensitivity models using the nearest-neighbour distance and a permutation test. The method is applied to a missing data problem in meta-analysis and a longitudinal study in Section 3; both examples are illustrated with numerical results from real data problems. Finally, a discussion is given in Section 4 with further examples presented in Appendix.

2 Simulation-based sensitivity analysis

2.1 Sensitivity models and sensitivity parameters

Let $\mathcal{F}$ be the population of complete data from which we wish to infer the parameter of interest θ using model $f(\mathcal{F};\theta)$. If a sample D is drawn randomly from $\mathcal{F}$, $\hat{\theta}$ calculated from the model $f(D;\theta)$ is usually unbiased. However, observed data, denoted by D_{obs}, are often not a representative of the complete data. Conventional inference usually employs a model $f(D_{obs};\theta)$ under assumptions of missing at random (MAR) and this may result in bias since those assumptions are often invalid under some 'imperfect' situations such as publication bias in meta-analysis, measurement error with non-ignorable missingness or the use of a misspecified MDM model. The effect of bias source may be modeled with a bias model parametrized with a bias parameter.

Let $D = (D_{obs}, D_{mis})$ be a set of complete data including both observed and unobserved data and R a missingness indicator vector which takes 1 if data is observed or 0 otherwise. The complete data model can be factorized into an extrapolation model and an observed data model as follows.

$$f(D, R|\Theta) = f(D_{mis}|D_{obs}, R, \Theta_m)f(D_{obs}, R|\Theta_o). \tag{1}$$

Here, Θ_m and Θ_o denote parameters indexing the models for missing data and observed data respectively. The item $f(D_{mis}|D_{obs}, R, \Theta_m)$ is called extrapolation model since the missing data are often outside the range of the observed data. Usually the parameter of interest θ is a subset of Θ_0 and the remaining components are the nuisance parameters. To simplify notations we will not distinguish Θ_0 from θ and will always use the latter. The extrapolation model cannot usually be identified unless extra assumptions are made. The *bias* or *sensitivity parameters* (Daniels and Hogan, 2008; Greenland, 2005), denoted by η, are a function of the parameter Θ_m. These parameters are usually inestimable. But when their values are given, the full data model $f(D, R|\theta, \boldsymbol{\Theta}_m)$ in (1) is identified. In another words, the estimate of θ depends on the value of η, and it may be biased if the value of η is given wrongly. We assume one value exists, denoted by η_0, for which the model $f(D, R|\theta, \boldsymbol{\Theta}_m)$ would provide an unbiased estimate of θ given D_{obs} and η_0.

We now illustrate the meaning of sensitivity model and sensitivity parameters using an example which assesses the relationship between passive smoking and lung cancer (Hackshaw et al., 1997). A total of 37 published epidemiological studies are selected to investigate the risk of lung cancer in female non-smokers whose spouses/partners did or did not smoke. Suppose y_i is the estimated log odds ratio reported in the ith study and s_i is the corresponding standard error for $i = 1, \ldots, n$. A random-effects meta-analysis model is given by

$$y_i = \mu_i + \sigma_i \epsilon_i,\ \mu_i \sim N(\mu, \tau^2),\ \epsilon_i \sim N(0, 1),\ i = 1, \ldots, n. \tag{2}$$

Here μ is the overall mean effect which is the parameter of interest, τ^2 is the variance measuring heterogeneity while σ_i^2 is the within-study variance and is usually replaced by the sample variance s_i^2. Figure 1 presents the funnel plot of the log odds ratios showing the sign of publication bias, i.e. smaller studies give more positive results than the larger studies. In other words smaller studies with inconclusive results are less likely to be selected; see the detailed discussion in Copas and Shi (2000).

Without assuming publication bias, the maximum likelihood estimate of the overall mean is $\hat{\mu} = 0.22$. The relative risk is therefore 1.246 or the excess risk is 24.6%, which implies that people who live with smokers have a 24.6% excess risk to have lung cancer when compared with those living with non-smokers. This is however an over-estimate due to publication bias as argued by Copas and Shi (2000). To address the problem they used the following selection model

$$z_i = a + b/s_i + \delta_i,\ \delta_i \sim N(0, 1),\ \mathrm{corr}(\epsilon_i, \delta_i) = \rho. \tag{3}$$

A study is selected when $z_i > 0$ (i.e. $R_i = 1$). In this model $\eta = (a, b)$ are the sensitivity parameters. If $\eta = (a, b)$ is given, $P(z_i > 0|y_i)$ can be calculated from the selection model whereas the parameter of interest μ and the nuisance parameters (τ^2, ρ) can be estimated from the meta-analysis model (2) and the selection model (3) (Copas and Shi, 2000) by maximizing the following log-likelihood.

$$\begin{aligned} L(\mu, \tau^2, \rho) &= \sum_{i=1}^{n} [\log f(y_i|z_i > 0)] \\ &= \sum_{i=1}^{n} [\log f(y_i) + \log P(z_i > 0|y_i) - \log P(z_i > 0)]. \end{aligned} \tag{4}$$

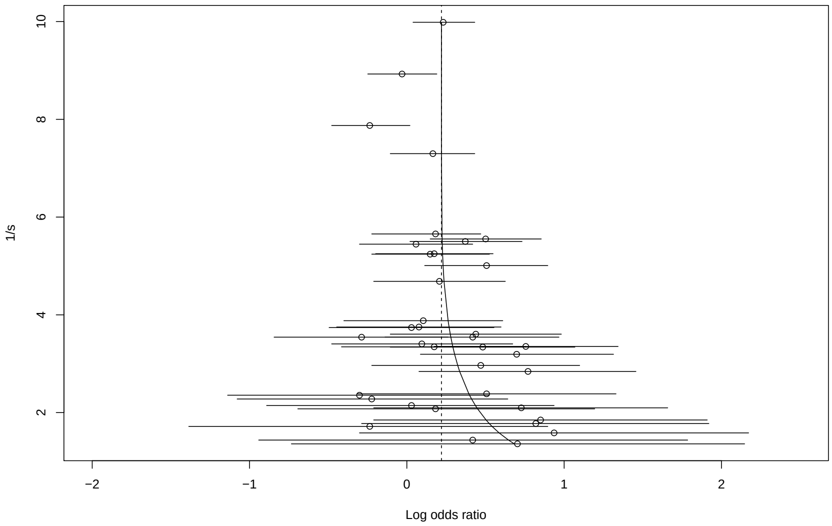


Figure 1: Passive smoking and lung cancer data: log odds ratios and their 95% confidence interval versus the inverse of the sample standard errors for 37 studies; the dashed line represents $\hat{\mu} = 0.22$, the estimate without assuming selectivity; the solid line represents fitted values obtained from the most plausible model with $(a, b, \hat{\mu}) = (-2.6, 0.8, 0.165)$.

However, since the number of unselected studies is unknown, $\eta = (a, b)$ is inestimable. In this example, uncertainty exists for the sensitivity parameters (a, b) as well as the assumed selection model (3). A sensitivity analysis investigates how sensitively the estimate of θ depends both on the choice of the selection model and the choice of the sensitivity parameter η.

The sensitivity parameter η in this example has no clear physical meaning. It is difficult to draw any meaningful conclusion if it is used directly. To make sensitivity analysis useful we need to develop some simple and interpretable quantities. Copas and Shi (2000) used the p-value of the goodness-of-fit to funnel plots and Copas (2013) used the overall selection probability. Those quantities have clear physical interpretation from which meaningful conclusions can be drawn; this idea, however, is not easily expandable to other missing data problems. In this paper, we propose a more generic method which works by comparing the observations with datasets simulated from candidates of the sensitivity model $f(D; \theta, \eta)$. The nearest-neighbour distance is used to evaluate those models. In principle, this method can be applied to any missing data problem; we provide further details about the procedure in the next two subsections.

2.2 Evaluating sensitivity models

As we have discussed previously, $f(D, R; \theta, \eta)$ is the sensitivity model, where $D = (D_{obs}, D_{mis})$, θ is the parameter of interest and η is the sensitivity parameter. There is uncertainty for η, but θ can be estimated when η is given. For example, we may estimate θ through a marginal density

$$f(D_{obs}, R|\theta, \eta) = \int f(D_{obs}, D_{mis}, R|\theta, \eta) dD_{mis}. \tag{5}$$

The meta-analysis example discussed in the previous subsection follows this approach, and θ is estimated from the marginal likelihood (4) for the observed data when $\eta = (a, b)$ is given.

Let η_0 be the true value. The observed data D_{obs} are therefore generated from either the full model $f(D, R; \theta, \eta_0)$ or its corresponding marginal model $f(D_{obs}, R; \theta, \eta_0)$. Hence, if the value η_0 is known, an unbiased estimate for θ can be determined from the model and the observed data D_{obs}; see the discussion on the related topics in **?**. Unfortunately η_0 is usually unknown and it cannot be estimated from D_{obs} under non-ignorable missingness. Our idea considers a set of sensitivity parameters denoted by Γ and investigate each η in Γ by comparing $f(D, R; \theta, \eta)$ with $f(D, R; \theta, \eta_0)$. If they are very close to each other, the candidate model would result in a (nearly) unbiased estimate under certain regularity conditions. Since η_0 is unknown, the comparison cannot be made directly; an indirect approach simulates datasets from the candidate model and proceeds further by comparing the simulated and the observed data (bearing in mind that the observed data D_{obs} come from $f(D, R; \theta, \eta_0)$).

The steps of this simulation-based sensitivity analysis (SSA) procedure are as follows.

(i) Select one η in Γ or generate it from a prior distribution $p(\eta)$ if we have some prior knowledge about η;

(ii) Estimate $\hat{\theta}_\eta$ using the sensitivity model $f(D, R; \theta, \eta)$ given the selected η;

(iii) Simulate an incomplete sample D_{η}^{obs} from the model $f(D, R; \hat{\theta}_{\eta}, \eta)$;

(iv) Calculate distance $s(D_{\eta}^{obs}, D_{obs})$;

(v) Repeat Steps (i) to (iv) for each candidate η in Γ.

We have demonstrated how to calculate $\hat{\theta}_{\eta}$ from $f(D, R; \theta, \eta)$ given η in Step (ii) using the meta-analysis and publication bias example in the previous subsection. More examples will be discussed later in this paper.

In Step (iii) we first sample a complete dataset D_{η} from $f(D, R; \hat{\theta}_{\eta}, \eta)$ and then remove some elements from D_{η} in order to generate the incomplete observations D_{η}^{obs}; this latter dataset is comparable to the actual observations, D_{obs}. This process requires a missing data mechanism to be specified by η. As there is no an unified approach to simulate D_{η} or D_{η}^{obs}, we discuss some viable techniques in Section 3 and Appendix A.3.

Step (iv) calculates the distance between the simulated dataset D_{η}^{obs} and the actual observations D_{obs}. The key here is defining a proper measure in order to measure the 'closeness' or 'similarity' between the sets. This is particularly important for the large-dimensional case. To measure the similarity or dissimilarity between two clusters various statistical distances have been investigated in the literature. A common approach calculates the distance between each pair of data points in D_{η}^{obs} and D_{obs}, and then use either the minimum (single linkage by Sneath (1957)), the maximum (complete linkage by Sorensen (1948)) or the average distance (Sokal and Michener, 1958). In our experience, the first two distances do not work very well in SSA while the average distance works well for some examples but the performance is not consistent. As an alternative, we use the K nearest-neighbour (KNN) method, first introduced by Fix and Hodges (1951) as a nonparametric measure; further details are given in Appendix A.1. This measure works well in most of the cases we have tried.

We will discuss how to use the distance $s(D_{\eta}^{obs}, D_{obs})$ to conduct a sensitivity analysis in the next subsection.

Remark 1. The sensitivity model, either $f(D, R|\theta, \eta)$ or $f(D_{obs}, R|\theta, \eta)$, depends on the sensitivity parameter η as well as the hierarchical structure (see equation (1)); so does D_{η}^{obs}. Hence, the distance $s(D_{\eta}^{obs}, D_{obs})$ can be used to investigate both the plausibility of η and the model structure. This method is therefore appropriate for the study of misspecified models and other related problems.

Remark 2. Numerical problems may arise in step (iv) when calculating the distance between D_{obs} and one set of D_η^{obs}. A strategy that works in those instances makes use of an average distance by sampling more than one set from D_η^{obs} for a given η. We will be referring to the number of replicates taken to build the average distance as a Monte Carlo sample size or simply MC size.

Remark 3. KNN distance is just a way to measure similarity between D_{obs} and D_η^{obs}. It performs well for the examples discussed in this paper. However, some other distances or measures may be used for different types of data or problems.

2.3 Sensitivity analysis

The distance $s(D_{obs}, D_\eta^{obs})$ measures the 'closeness' between the sensitivity model and the 'true' model. Those models giving rise to large values of this metric can simply be discarded as they are very unlikely. In order to define a test criterion more formally, we borrow the idea of an 'achieved significance level' (ASL) introduced by (Fisher, 1971) using permutation tests. Let now D_1 and D_2 denote a pair of permutation samples drawn from the combined dataset $D^* = (D_{obs}, D_\eta^{obs})$ whose distance is $s^* = s(D_1, D_2)$. We define the ASL for SSA as:

$$ASL_\eta = \Pr_{H_0}\{s^* \geq s(D_{obs}, D_\eta^{obs})\}, \tag{6}$$

where H_0 is the null hypothesis stating that D_{obs} and D_η^{obs} come from the same distribution. When we need to resort to average distances to tackle numerical instabilities as per Remark 2, permutation samples are generated over the multiple D_η^{obs} with a given MC size. The ASL is the proportion of the pairs of the permutation samples for which the distances are larger than the distance between D_{obs} and D_η^{obs}. Thus ASL works similarly to a p-value with larger values taken as being in favour of H_0.

Other approach to work out the ASL uses the concept of 'internal distance'; in that case, the permutation sample (D_1, D_2) is drawn from the D_η^{obs} instead of the combined dataset D^*.

For a given significance level (e.g. $\alpha = 5\%$), a plausible set of sensitivity parameters is defined as $\Gamma_\alpha = \{\eta : ASL_\eta > \alpha\}$. Models not included in Γ_α should be avoided.

For all the models within Γ_α, the distance calculated for each η can be used as a further guide in deriving an estimate for θ. For instance, we define the 'most plausible value' of η that gives rise to the smallest distance and term the corresponding estimate of θ as the 'most plausible estimate'. It is important to emphasize, however, that this estimate may not be consistent regardless of the sample size. Hence, caution should be exercised even when the most plausible model has been chosen; we provide further discussion about this topic in the examples.

3 Examples

We will discuss two missing data problems in this section, including publication bias problem in meta-analysis and incomplete longitudinal data. Techniques on how to generate D_η from $f(D;\theta,\eta)$ and how to sample D_η^{obs} from D_η will also be discussed. More examples will be given in Appendix, including mean estimation with non-ignorable missing data and non-ignorable missing covariates.

3.1 Publication bias in meta-analysis

We carry out the meta-analysis for the example of passive smoking and lung cancer discussed in Section 2.1. In this example, μ is the parameter of interest, (τ^2,ρ) are the nuisance parameters and $\eta=(a,b)$ are sensitivity parameters which are inestimable. When $\eta=(a,b)$ are known, the marginal likelihood for the observed data is given by (4), i.e.

$$\begin{aligned} L(\mu,\tau,\rho|a,b) &= \sum_{i=1}^{n}[\log f(y_i|z_i>0,s_i)] \\ &= \sum_{i=1}^{n}\left[-\frac{1}{2}\log(\tau^2+\sigma_i^2)-\frac{(y_i-\mu)^2}{2(\tau^2+\sigma_i^2)}-\log\Phi(a+\frac{b}{s_i})+\log\Phi(v_i)\right] \end{aligned}$$

where

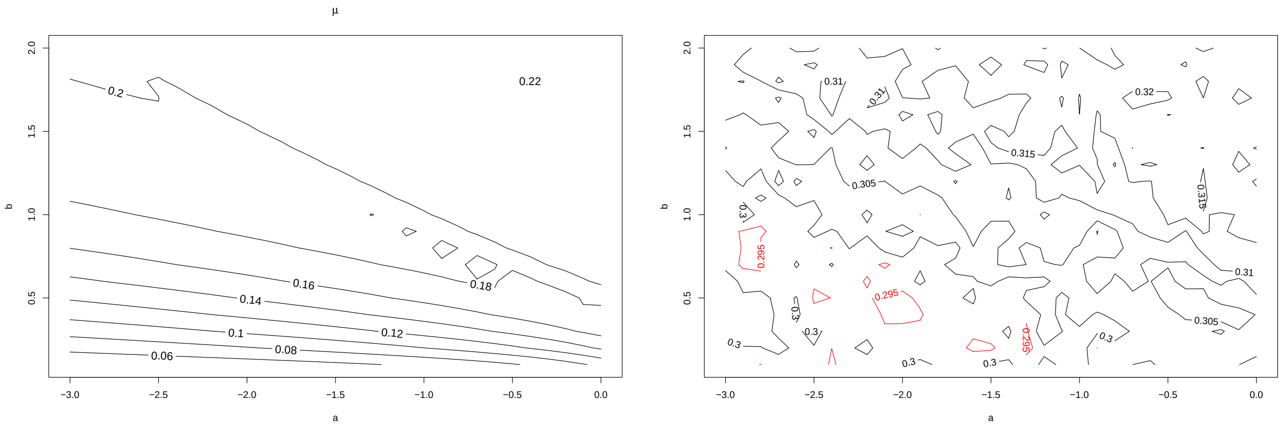
$$v_i=\frac{a+\frac{b}{s_i}+\tilde\rho_i\frac{y_i-\mu}{(\tau^2+\sigma_i^2)^{1/2}}}{(1-\tilde\rho_i^2)^{1/2}},\ \tilde\rho_i=\frac{\sigma_i}{(\tau^2+\sigma_i^2)^{1/2}}\rho,$$

and $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Parameters (μ,τ^2,ρ) can be estimated by maximizing the above marginal likelihood given the observed data.

A sensitivity analysis is conducted to investigate how the estimates of μ change for different values of $\eta = (a, b)$. The selection of possible values of a and b is based on a range of selection probabilities for the largest and the smallest studies as suggested in Copas and Shi (2001). This region is $a \in (-3, 0)$ and $b \in (0.1, 2)$ as shown in Figure 2 (a), from which a grid of values can be chosen accordingly to conduct the sensitivity analysis. The area towards the bottom left indicates a high potential for publication bias; this reduces as we move towards the top right of the figure. For each selected pair of (a, b), μ is estimated using the maximum likelihood method. A contour plot is presented in Figure 2 (a). At the top right $\hat{\mu} = 0.22$ which agrees with the estimate from a model without assuming publication bias but this value falls as the sensitivity parameters move towards the bottom left corner. A question which arises naturally is: which estimate should we use? In this regard, Copas and Shi (2000) and Copas and Shi (2001) have suggested several statistical quantities which are interpretable and from which some answers can be found to the aforementioned question.

We now use the SSA method to conduct a sensitivity analysis. For each pair of (a, b) we calculate the estimates $(\hat{\mu}, \hat{\tau}^2, \hat{\rho})$ and simulate a complete dataset using meta-analysis model (2); denoted the simulated data by $D = \{(y_j, s_j), j = 1, \ldots, N\}$. Given the known values of (a, b) and the estimates of the parameters, we subsequently draw random numbers of z from the selection model (3). Note that z is correlated with y and the study is not selected if $z < 0$. The selected studies form the observed data $D^{obs}_{(a,b)}$. This dataset is compared with the actual dataset D_{obs} by using KNN distance $s(D^{obs}_{(a,b)}, D_{obs})$. In accordance with Remark 2, Section 2.2, we have used the average distance with 1000 replicates for each pair of (a, b) in order to remove any numerical instability (i.e. the MC size is 1000). The contours of the average distances are shown in Figure 2 (b).

Using the permutation method, discussed in Section 3.3, we can calculate the ASL for each value of η and discard all pairs of sensitivity parameters whose ASL values are smaller than 0.05. The remaining pairs form a plausible set of sensitivity parameters which results in estimates for $\hat{\mu}$ ranging from 0.043 to 0.197. The upper bound for this estimate indicates an excess risk of 21.7%. The conventional method (without assuming publication bias) results in an excess risk of 24.6%, thus overestimating the



(a) $\hat{\mu}$

(b) average distance

Figure 2: Contours plots for passive smoking and lung cancer data. In (b), the area in red corresponds to the most plausible values for η with the average distances being all very close to the smallest distance of 0.295.

SSA upper bound by 13%.

The most plausible value (smallest distance) is achieved when $(a, b) = (-2.6, 0.8)$ which leads to the estimate of $\hat{\mu} = 0.165$ or an excess risk of 17.9%. The fitted line is shown in Figure 1, showing a much better fit than the dashed line which is obtained from the model without considering publication bias.

As we have mentioned before hand, caution is needed when working with the most plausible value. This can be seen by looking at Figure 2(b), where the average KNN distances are all very close to the shortest distance of 0.295 for several values of the sensitivity parameters. The estimates of μ for those sets range between 0.086 and 0.183 which lead to excess risks in the interval between 9% and 20%.

Based on the 37 studies collected in the meta-analysis and the simulation-based sensitivity analysis, we can conclude that the excess risk is not likely to be higher than 21.7% or smaller than 4.4%. Most likely it will take values between 9% and 20%. This result is consistent with the findings in Copas and Shi (2000). The latter is based on goodness-of-fit test but the method is not easily extendable to other missing data problems. The SSA method proposed in this paper is more general and can be used to deal with different types of problems.

3.2 Incomplete Longitudinal Example

We next consider a follow-up randomized study with M scheduled repeated measures. Let Y_m be the outcome measured at visit m; we use the notation $\boldsymbol{Y}_m^- = (Y_1, \ldots, Y_m)$ to denote the history of the outcomes up to visit m and $\boldsymbol{Y}_m^+ = (Y_{m+1}, \ldots, Y_M)$ as the future outcomes after visit m. Under monotone missingness, if the outcome at visit m is missing then all outcome after that visit are all missing. We denote R_k as the indicator that Y_k is observed, which equals to 1 if Y_k is observed and 0 otherwise. We assume that Y_1 is observed on all individuals ($R_1 = 1$) and let d be the last visit on which a subject has a measure; it follows that the observed data for that subject is $\boldsymbol{Y}_d^- = (Y_1, \ldots, Y_d)$, where $d \leq M$. We are interested in the inference of the mean $\mu = E(Y_M)$, the intended outcome at the final scheduled visit.

The probability of dropping out at visit k can be modeled by the following logistic model:

$$\text{logit}\{1 - P(d = k | d \geq k, \boldsymbol{Y}_k^-, \boldsymbol{Y}_k^+)\} = h(\boldsymbol{Y}_k^-) + \eta Y_{k+1}$$

where $h(\cdot)$ is an unknown function. Either parametric or nonparametric approaches can be used. A special case is the following linear model: $h(\boldsymbol{Y}_k^-) = \eta_0 + \eta_1 Y_k$. This model describes the probability of dropping out at visit $k+1$ by a logistic linear model which depends on the latest recorded data Y_k and the current unobserved data Y_{k+1}. If $\eta = 0$ this is a MAR model. One would make this choice if there is evidence or belief that the missingness associated to Y_{k+1} can be entirely encoded by the recorded history observations $\boldsymbol{Y}_k^-$. This is certainly a strong assumption in longitudinal study and usually very difficult to justify. We now use the SSA method proposed in this paper to investigate the model under a MNAR mechanism, and try to find a plausible value of η in the neighbourhood of 0 (where 0 corresponds to the MAR model).

Let us first consider a simple case. Suppose that we just have two visits, i.e. $M = 2$. There is no missing data for the first visit, i.e., Y_1 is observed for all the subjects. But part of the data are missing in the second visit. Let R be the missing indicator for Y_2. A semi-parametric logistic regression model can be defined by:

$$P(R = 1 | Y_1, Y_2) = \text{expit}(h(Y_1) + \eta Y_2), \tag{7}$$

where $h(Y_1)$ is a nonparametric model and $\text{expit}(t) = e^t/(1 + e^t)$. The parameter η is not identifiable from the observed data only since it also depends on the missing part of Y_2.

We need the following lemma to apply SSA to the above problem.

Lemma 1 *Let T be the statistical variables which are partly observed. Given the observed data D_{obs}, the conditional density of the missing part of T (i.e. when $R = 0$) can be expressed by*

$$f(T|D_{obs}, R = 0) = f(T|D_{obs}, R = 1)\frac{Q(D)}{E(Q(D)|D_{obs}, R = 1)}, \tag{8}$$

where

$$Q(D) = \frac{P(R = 0|D)}{P(R = 1|D)} \tag{9}$$

is the conditional odds of missing data.

The proof is given in Appendix A.2.

Using (7) and the above lemma, we have

$$f(Y_2|D_{obs}, R = 0) = f(Y_2|D_{obs}, R = 1)\frac{\exp(-\eta Y_2)}{\mathrm{E}(\exp(-\eta Y_2)|D_{obs}, R = 1)}. \tag{10}$$

In the above formula, the conditional density of the observed part of Y_2, $f(Y_2|D_{obs}, R = 1)$, can be obtained parametrically or non-parametrically, for example, using a generalized additive model (Hastie and Tibshirani, 1990) or Gaussian process regression (Shi and Choi, 2013). Thus the only uncertainty in formula (10) is the sensitivity parameter η which cannot be estimated from the observed data.

The missing data can therefore be generated from (10) for each given η. This forms a complete dataset for Y_2. We can further select a subset using the selection model (7) which we denote as $Y_{2,\eta}^{obs}$. And, finally, this subset can be compared with the real observed data Y_2^{obs} by using the KNN distance. Thus we can conduct a sensitivity analysis for η. The details of this procedure are worked out in the next subsection with a real example.

Let us now move to discuss the general case with M visits. Assume that the data have been recorded for all the subjects up to the k-th visit but with some missing data afterwards. Let R_{k+1} be the missing data indicator for the $(k+1)$-th visit. We use a logistic regression model similar to (7) as the MDM model:

$$P(R_{k+1} = 1|\boldsymbol{Y}_k^-, Y_{k+1}) = \mathrm{expit}(h(\boldsymbol{Y}_k^-) + \eta_{k+1}Y_{k+1}). \tag{11}$$

Using Lemma 1, we have

$$f(Y_{k+1}|D_{obs}, R_{k+1} = 0) = f(Y_{k+1}|D_{obs}, R_{k+1} = 1)\frac{\exp(-\eta_{k+1}Y_{k+1})}{\mathrm{E}(\exp(-\eta_{k+1}Y_{k+1})|D_{obs}, R = 1)}. \tag{12}$$

We can generate the missing part of Y_{k+1} using the above density function for any given η_{k+1} and then use those generated data to fill in the missing observations in Y_{k+1}; we denote the data up to the $(k+1)$-th visit as $\boldsymbol{Y}^{-}_{k+1,\eta_{k+1}}$. Note that the imputed data for missing part of Y_{k+1} depends on the sensitivity parameter, so the notation is sub-indexed by η_{k+1}.

The procedure goes on to generate the missing data for the $(k + 2)$-th visit in a similar way using formulas (11) and (12) slightly adapted by

$$P(R_{k+2} = 1|\boldsymbol{Y}^{-}_{k+1,\eta_{k+1}}, Y_{k+2}) = \mathrm{expit}(h(\boldsymbol{Y}^{-}_{k+1,\eta_{k+1}}) + \eta_{k+2}Y_{k+2}),$$

and

$$f(Y_{k+2}|D_{obs,\eta_{k+1}}, R_{k+2} = 0) = f(Y_{k+2}|D_{obs,\eta_{k+1}}, R_{k+2} = 1)\frac{\exp(-\eta_{k+2}Y_{k+2})}{\mathrm{E}(\exp(-\eta_{k+2}Y_{k+2})|D_{obs,\eta_{k+1}}, R = 1)},$$

where $D_{obs,\eta_{k+1}}$ includes all the observed data and the inputed data for missing part in the $(k + 1)$-th visit. The above density function can be used to generate missing data and fill in the missing part for the $(k+2)$-th visit. Keeping up with the notation, η_{k+2} is the sensitivity parameter in this step.

The procedure continuous until all the missing observations are filled in. In this case, the generated data depends on a sensitivity parameter vector $\boldsymbol{\eta} = (\eta_{k+1}, \ldots, \eta_M)^T$. For each visit starting from the $(k+1)$-th, we can form a subset of selected data based on the generated data and the selection model, and then compare it with the actual observations. This can be used to conduct a simulation-based sensitivity analysis for the parameter $\boldsymbol{\eta}$.

Remark 4. The aforementioned method may be very time consuming if the dimension of $\boldsymbol{\eta}$ is large. In this case, we may focus on a few visits which have a large missing rate and the MDM may be non-ignorable. We can then use SSA method for those visits by considering a grid of $\boldsymbol{\eta}$, and assume a MAR MDM and then use multiple imputation for the other visits.

3.2.1 Children weight data – a bivariate outcome with dropout

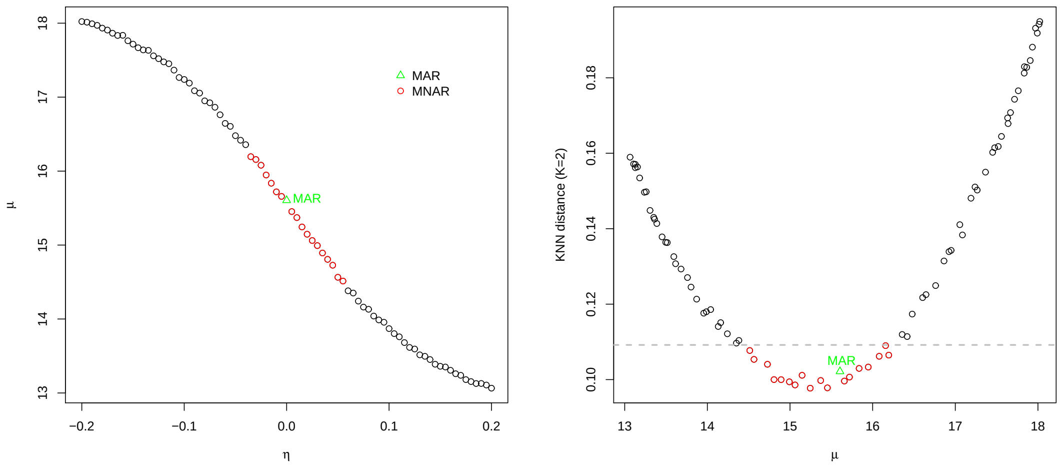
Here we will apply the simulation-based sensitivity analysis to a real example. The 1970 British Cohort Study (BCS70) follows the lives of more than 17,000 people born in England, Scotland and Wales. Two sub-samples in the first 5 years of birth were collected: a 22-month subsample (Chamberlain et al., 1970) and a 42-month subsample (Chamberlain, 1973). We are interested in the average Children' weight at 42-month from birth. The 22-month subsample has a total of 2348 individuals on which information was collected; the mean and standard deviation (SD) for those observations are 11.92 and 1.56kg respectively. Likewise, the 42-month subsample has 1856 observations (i.e. data on 492 people are missing for unknown reasons) with a mean and SD of 15.04 and 1.98kg respectively. We believe that the dropping out mechanism is associated with the first outcome (Y_1) and may be also associated with the current outcome (Y_2).

We now conduct a simulation-based sensitivity analysis as discussed in the previous subsection. We choose η within the interval (-0.2, 0.2) and consider any values beyond this range as not very likely. This is because for example, if $\eta < -0.2$, then at least 94% of the 494 unobserved Children would have weights greater than 20 kg which is not not a realistic outcome when considering the mean of 15.04 and standard deviation of 1.98 for the observed data.

In this example, we simply use a linear regression model to obtained the density function of observed Y_2 given the observed data, and a linear model: $h(Y_1) = \beta_0 + \beta_1 Y_1$ in (7). Then, the parameters (β_0, β_1) can be estimated from the observed data when the value of sensitivity parameter η is given.

The final results are shown in Figure 3. The plausible set indicates that the mean estimation of the 42 month weight is in the range of (14.51, 16.19) with the most plausible value of 15.22. The average distances are presented in Figure 3b, providing further evidence that the sensitivity models near the bottom (corresponding to the most plausible model) are much closer to the true model than the others.

It is interesting to note that the sensitivity model with a MAR assumption ($\eta = 0$) is also included in the plausible set as shown in Figure 3. Although the MAR model is not the model having the shortest distance to the true model, there is no evidence to reject it. Or in other words, the non-ignorability of the missingness is not severe. This is a by-product of using the SSA method.



(a) $\hat{\mu}$

(b) average distance

Figure 3: Children weight data: estimation of mean weight μ and the average KNN distance for a set of sensitivity parameter; 'MAR' stands for the model with MAR assumption.

4 Discussion

Missing data can usually be modelled by a sensitivity model $f(D, R; \theta, \eta)$, but the estimate of θ, the parameter of interest, is often biased unless the true value of sensitivity parameter η is given. Unfortunately η is often inestimable based on the observed data D_{obs} particularly for non-ignorable missing data. A sensitivity analysis investigates how sensitively the estimate of θ depends on the choice of η. To get some meaningful results, we need to use some statistical quantities which have clear physical interpretation. The simulation-based sensitivity analysis method we proposed is based on the KNN distance. It can be applied in a general fashion and can help in gaining new and further insights into the problem at hand. The method can be applied to a great variety of missing data and model misspecification problems as we have shown in four different examples throughout this manuscript. The numerical studies show that meaningful results can be obtained for each of those different problems.

One of the key steps when conducting a simulation-based sensitivity analysis is to be able to simulate missing data when η is given. Although this may be straightforward for some problems such as the meta-analysis with publication bias discussed in Section 3.1, in many other cases it will not be a trivial step; an example of this was shown in Section 3.2 which required of a somewhat elaborate process to tackle the generation of the missing observations. Other models can be treated similarly although a different array of techniques may be needed as shown in the extended examples provided in Appendix.

Whereas the most plausible value of η will provide a final result, in certain occasions it may be more advisable to work with a range of solutions. As we discussed for the meta-analysis example in Section 3.1, there may be a set of η for which the corresponding KNN distances are all very close to the shortest distance. In those cases, we recommend using a set of estimates within a small range rather than a single value.

Acknowledgment

The authors would like to thank the editor, associate editor and two reviewers for their critical and constructive comments and suggestions. The authors are grateful to the Centre for Longitudinal Studies, Institute of Education for the use of the data and to the UK Data Archive and Economic and Social Data Service for making them available; they however bear no responsibility for the analysis and interpretation of the data presented in this paper. The authors also thank Dr Javier Serradilla of Newcastle University, UK for helpful suggestions to improve the presentation of the paper.

References

Chamberlain, R. (1973). *1970 British Cohort Study: 42-Month Subsample, 1973.* [computer file]. Colchester, Essex: UK Data Archive [distributor], August 2008. SN: 2690, http://dx.doi.org/10.5255/UKDA-SN-2690-1, 2nd edition.

Chamberlain, R., Chamberlain, G., and of London. Institute of Education. Centre for

Longitudinal Studies., U. (1970). *1970 British Cohort Study: Birth and 22-Month Subsample, 1970-1972*. [computer file]. Colchester, Essex: UK Data Archive [distributor], May 2013. SN: 2666, http://dx.doi.org/10.5255/UKDA-SN-2666-2, 3rd edition.

Copas, J. (1999). what works?: selectivity models and meta-analysis. *J.R.Statist.Soc.A.*, 162:95–109.

Copas, J. (2013). A likelihood-based sensitivity analysis for publication bias in meta-analysis. *J.R.Statist.Soc.C.*, 62:47–66.

Copas, J. and Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *J.R.Statist.Soc.B.*, 63(4):871–895.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias. *J.R.Statist.Soc.B.*, 67:459–513.

Copas, J. and Li, H. (1997). Inference for non-random samples. *J.R.Statist.Soc.B.*, 59:55–95.

Copas, J. and Shi, J. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, 1:247–262.

Copas, J. and Shi, J. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statist. Meth. in Med. Res.*, 10:251–265.

Cornfield, J., Haenszel, W., Hammond, W., Lilienfeld, A. M., Shimkin, M., and Wynder, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natn. Cancer Inst.*, 22:173–203.

Daniels, M. and Hogan, J. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman & Hall/CRC.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J.R.Statist.Soc.B.*, 57(1):45–97.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. CRC Press, Boca Ration, FL, 1st edition.

Fisher, R. A. (1971). *The Design of Experiiments.* New York: Hafner Publishing, 8th edition.

Fix, E. and Hodges, J. (1951). Discriminatory analysis- nonparametric discrimination: Consistency properties. Technical report, University of California, Berkeley.

Greenland, S. (2005). Multiple bias modelling for analysis of observational data. *J.R.Statist.Soc.A.*, 168:267–306.

Hackshaw, A., Law, M., and Wald, N. (1997). The accumulated evidence on lung cancer and environmental tobacoo smoke. *British Medical Journal*, 315:980–988.

Hall, P., Park, B., and Samworth, R. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, 36(5):2135–2152.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman & Hall/CRC., 1st edition.

Hechman, J. (1979). Sample selection bias as a specification error. *Econometrics*, 47:153–161.

Kim, J. and Yu, C. (2011). A semiparametric estimation of mean functionals with non-ignorable missing data. *Journal of the American Statistical Association.*, 106(493).

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data.* Wiley, New York, second edition.

McCandless, L., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, 26:2331–2347.

Molenberghs, G., Kenward, M., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the slovenian plebiscite case. *Appl. Statist*, 50(1):15–29.

Nigsch, F., Bender, A., Buuren, B., Tissen, J., Nigsch, E., and Mitchell, J. (2006). Melting point prediction employing k-nearest neighbor algorithm and genetic parameter optimization. *J. Chem. Inf. Model.*, 46:2142–2422.

Resseguier, N. (2011). Multivariate imputation by chained equations (iteration step for sensitivity analysis). *Epidemiology*, 22(2):282.

Shi, J. and Choi, T. (2013). *Gaussian Process Regression Analysis for Functional Data.* Chapman & Hall/CRC., 1st edition.

Sneath, P. (1957). The application of computers to taxonomy. *Journal of General Mocrobiology*, 17:201–226.

Sokal, R. and Michener, C. (1958). A statistical method for evaulating systematic relationships. *University of Kansas Science Bulletin.*, 38:1409–1438.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5:1–34.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Vemuri, V., Dracup, J., Erdmann, R., and Vemuri, N. (1969). Sensitivity analysis method of system identification and its potential in hydrologic research. *Water Resources Research*, 5:341–349.

Weisberg, S. (2005). *Applied linear regression.* John Wiley & Sons, INC., 3rd edition.

Wong, M. and Lane, T. (1983). A kth nearest neighbour clustering procedure. *Journal of the American Statistical Association.*, 45:362–368.

Appendix

A.1 K nearest-neighbour (KNN) distance

Two observations x_i and x_j are defined to be neighbours if (see Definition 1, page 364 Wong and Lane, 1983):

$$d(x_i, x_j) \leq d_k(x_i) \,\mathrm{or}\, d_k(x_j),$$

where d is the Euclidean metric and $d_k(x_i)$ is the kth nearest-neighbour distance to point x_i. Prior to defining the distance between two clusters using the KNN method, the distance between an individual point and a cluster needs to be formalized first. Given a cluster $D = \{x_i, i = 1, \ldots, n\}$, an individual observation x is said to be neighbour of cluster D if there exists at least one point x_i in cluster D that

$$d(x_i, x) \leq d_k(x_i).$$

where d is the Euclidean metric and $d_k(x_i)$ is the distance of kth nearest-neighbour within the cluster to the point x_i.

Now we can define the similarity and the distance between two clusters. Let $D^* = \{x_j^*, j = 1, \ldots, m\}$ be another set; we expect that most of observations in D^* are in the nearest-neighbour of D if the two clusters are quite similar or in 'close proximity'. A measure of how close both cluster are is given by the proportion of observations in D^*. This is formally defined by the average $\mathrm{E}(I_1)$ where its j-th element has the form:

$$I_1^{(j)} = \begin{cases} 1, & \sum_{x_i \in D} \{I(d(x_i, x_j^*) < d_k(x_i)\} > 0; \\ 0, & \text{otherwise.} \end{cases} \quad \text{for } j = 1, \ldots, m. \tag{13}$$

Here I is indicator function which takes 1 if the condition is satisfied and 0 otherwise, and thus $\sum_{x_i \in D} \{I(d(x_i, x_j^*) < d_k(x_i)\}$ takes an integer in the set $\{0, 1, 2, ..n\}$. Only when it equals to 0, the observation x_j is not the nearest-neighbour of cluster D.

Similarly, we can define the proportion $\mathrm{E}(I_2)$ of the points in D with

$$I_2^{(i)} = \begin{cases} 1, & \sum_{x_j^* \in D^*} \{I(d(x_i, x_j^*) < d_k(x_j^*)\} > 0; \\ 0, & \text{otherwise.} \end{cases} \quad \text{for } i = 1, \ldots, n. \tag{14}$$

The average of $E(I_1)$ and $E(I_2)$

$$s(D, D^*) = \frac{1}{2}(E(I_1) + E(I_2))$$

is considered as a similarity measure between D and D^* and $(1 - s(D, D^*))$ is used as the 'KNN distance'.

Other measures such as Mahalanobis metric may also be used.

A.2 Proof of Lemma 1

Using Bayes Theorem, we have

$$f(T|D_{obs}, R = 0) = \frac{P(R = 0|D)f(T|D_{obs})}{P(R = 0|D_{obs})}.$$

Similarly, we have

$$f(T|D_{obs}, R = 1) = \frac{P(R = 1|D)f(T|D_{obs})}{P(R = 1|D_{obs})}.$$

This leads to the following equation

$$f(T|D_{obs}, R = 0) = f(T|D_{obs}, R = 1)\frac{P(R = 0|D)}{P(R = 1|D)}\frac{P(R = 1|D_{obs})}{P(R = 0|D_{obs})}.$$

Let $Q(D)$ be the one defined in (9), then we have

$$\begin{aligned} E(Q(D)|D_{obs}, R = 1) &= \int \frac{P(R = 0|D_{obs}, T)}{P(R = 1|D_{obs}, T)} f(T|D_0, R = 1)dT \\ &= \int \frac{f(T|D_0, R = 1)}{P(R = 1|D_{obs}, T)} P(R = 0|D_{obs}, T)dT \\ &= \int \frac{f(T|D_{obs})}{P(R = 1|D_{obs})} P(R = 0|D_{obs}, T)dT \\ &= \int f(T|D_{obs}, R = 0)\frac{P(R = 0|D_{obs})}{P(R = 1|D_{obs})}dT \\ &= \frac{P(R = 0|D_{obs})}{P(R = 1|D_{obs})}. \end{aligned}$$

This proves the Lemma.

A.3 Additional examples

A.3.1 Mean estimation with non-ignorable missing data

Assume that a continuous variable X has mean μ and variance σ^2. The population mean μ is of interest. The complete dataset is $D = (X_{obs}, X_{mis})$ where the MDM depends on the missing value, and thus it is a non-ignorable missing data problem.

Let R be missing data indicator and further assume that the MDM is modeled by a logistic model

$$P(R=1|X=x) = \text{expit}\{\eta(x+\lambda)\} \tag{15}$$

where $\text{expit}(x) = \exp(x)/(1+\exp(x))$ and λ is assumed to be known (it can be estimated if the proportion of the missing data is known). The choice of $\eta = 0$ specifies that R and X are independent (i.e., MAR).

The parameter of interest is $\mu = E(X)$, the mean of the complete data, can be expressed by

$$\mu = E(X|R=1)P(R=1) + E(X|R=0)P(R=0) = \pi\mu_1 + (1-\pi)\mu_2,$$

where $\pi = P(R=1)$ is the proportion of the observed data; μ_1 and μ_2 are the means of the observed and the missing data respectively. So the evaluation of μ_2 is the main task. Using Bayes theorem, we have

$$\begin{aligned} f(x|R=0) &= \frac{P(R=0|x)f(x)}{P(R=0)} \\ &= f(x|R=1)\frac{P(R=1)}{P(R=0)}\frac{P(R=0|x)}{P(R=1|x)}. \end{aligned} \tag{16}$$

Denote that $\pi_x = P(R=0|x)$, then

$$\frac{P(R=0|x)}{P(R=1|x)} = \frac{1-\pi_x}{\pi_x} = \frac{1}{\exp(\eta(x+\lambda))}$$

is the odds of missing when $X = x$. The second equation comes from MDM model (15). The mean of the missing data can therefore be expressed by

$$\begin{aligned} \mu_2 &= \int x f(x|R=0)dx \\ &= \int x f(x|R=1)\frac{\pi}{1-\pi}\frac{1}{\exp(\eta(x+\lambda))}dx \\ &= \frac{\pi}{1-\pi}E_{X|R=1}\left[\frac{X}{\exp(\eta(X+\lambda))}\right]. \end{aligned}$$

In this example, η is the sensitivity parameter. It is clear that we are unable to estimate this parameter from the observed data alone as it depends on the missing data as well.

Equation (16) is the key to conduct a simulation-based sensitivity analysis as it is needed in step (iii) of the SSA procedure to simulate data.

We now conduct a simulation study to demonstrate the use of the SSA approach in order to address the non-ignorable missing data problem. The data is generated from a normal distribution $X \sim N(\mu, \sigma^2)$ with the true values selected as $\mu = 0, \sigma^2 = 1$ and $\eta = -1, \lambda = 0$ in model (15), indicating an average missing proportion of about 50%. Sample size of the complete data is 100. In this example λ is assumed to be fixed and η is treated as a sensitivity parameter. The SSA approach is designed as follows. We first select a series of η chosen from the interval of (-5,5). For each selected η, we evaluate the density $f(x|R = 0)$ by (16) and then use the density function to sample the missing data, denoting the imputed values as $x_{mis,\eta}$. Thus, $D_\eta = (x_{obs}, x_{mis,\eta})$ forms a simulated complete dataset. To compare the simulated dataset with the observed data, we further generate a set of x_η^{obs} from D_η using MDM (15) with the given value of η. The simulated set of x_η^{obs} is comparable with the observed dataset x_{obs}. Finally, the closeness of x_η^{obs} and x_{obs} is evaluated using the KNN distance.

Figure 4 shows the results with the MC size of 1000. The KNN distance takes the minimum at $\eta = -0.89$ when $K = 2$. The corresponding estimate is $\hat{\mu} = -0.064$ which is quite close to the true value. We also used the permutation test to find the plausible set of the sensitivity parameter. The dashed line in Figure 4 indicates the critical value at a significance level of 5%; all the ones below that line forms the set of plausible values resulting in estimates of μ in the range of (-0.82, 0.91) when K=2. We have also repeated the simulation for other values of K. As shown in the same figure, all of them give similar results although the values of the KNN distance is less sensitive to η for larger values of K. Further details regarding the selection of K can be found in Hall et al. (2008) and Nigsch et al. (2006).

Table 1 presents the simulation study results for 500 replications with different MC sizes and different values of K. The value of η is selected by the shortest KNN distance with $\hat{\mu}$ being the most plausible estimate. If we just use the sample mean of the observed data, the average of $\hat{\mu}$ is about -0.5 which is far away from the true value of 0. Table 1 shows that the SSA approach gives much better results. This is also supported by the values of coverage probability, which are also reported in Table 1.

The results also show that the estimates are quite consistent for different values of K even for small number of the MC sizes. Histograms of η for a range of MC sizes are provided in Figure 5. As it can be seen, the method with a MC size of 100 or above when $K = 2$ usually gives quite robust results.

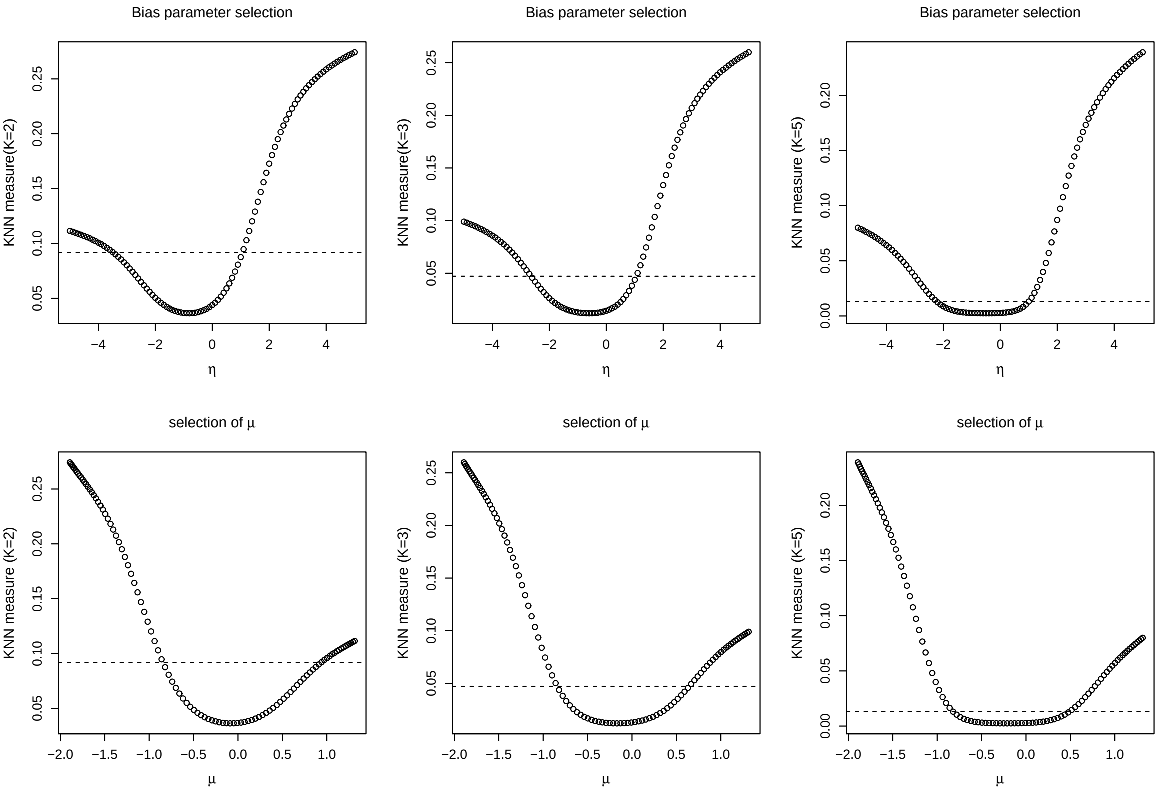


Figure 4: Mean estimation with non-ignorable missing data: Upper panel – KNN distances versus different values of η; Lower panel – KNN distance versus the corresponding estimate of μ for the given value of η. The dashed line indicates the critical values at 5% significance level.

Table 1: Mean estimation with non-ignorable missing data: simulation study

	average of $\hat{\mu}$ (sd)				Coverage Probability (%)			
MC size	10	20	100	1000	10	20	100	1000
K=2	-0.065 (0.16)	-0.069(0.16)	-0.064(0.16)	-0.064(0.16)	80.8	78.6	89.4	91.6
K=3	-0.133(0.16)	-0.138(0.16)	-0.158(0.16)	-0.154(0.16)	73.4	72.8	78.2	80.4
K=4	-0.143 (0.17)	-0.152 (0.16)	-0.190(0.16)	-0.191(0.16)	71.6	67.4	68.6	68.8
K=5	-0.169(0.17)	-0.187(0.17)	-0.211(0.16)	-0.204(0.16)	70.6	64.2	62.0	66.0

A.3.2 Regression analysis with missing covariates

We now consider a regression problem with missing confounders. Let $D = (D_{obs}, D_{mis})$, where D_{obs} denotes the data where both the response variable and the covariates are

Table 2: Mean estimation with non-ignorable missing data: selection of η

	average of selected η			
MC size	10	20	100	1000
K=2	-0.85	-0.87	-0.88	-0.89
K=3	-0.73	-0.71	-0.65	-0.67
K=4	-0.69	-0.67	-0.57	-0.57
K=5	-0.58	-0.59	-0.54	-0.54

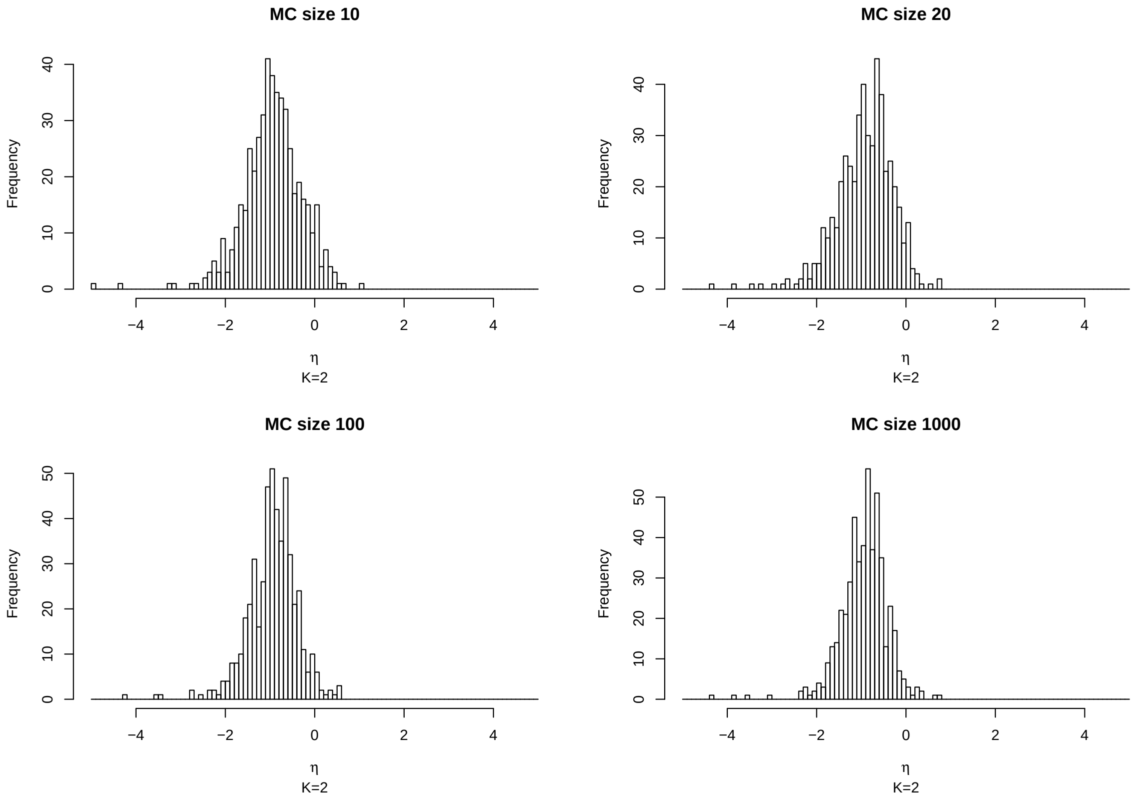


Figure 5: Mean estimation with non-ignorable missing data: histograms of the selected η with different MC sizes.

observed; while D_{mis} denotes the data where the response variable is observed but the covariates are only partly observed. We still use R as the missing data indicator and assume that the MDM depends on both D_{obs} and D_{mis}. The joint density of data D and the indicator R is expressed by $f(R, D)$, which can be factorized as $f(R|D)f(D)$.

We use the following semiparametric model (Kim and Yu, 2011) as the selection model for $f(R|D)$:

$$\text{logit}\{P(R=1|D_{obs}, D_{mis})\} = h(D_{obs}) + \eta D_{mis}, \tag{17}$$

where $h(\cdot)$ is a nonparametric function and η is a sensitivity parameter. This model takes a nonparametric model on the observed part D_{obs} and a simple linear form on the missing variable D_{mis}. Equation (17) can be rewritten as

$$\pi_D = P(R=1|D) = \text{expit}(h(D_{obs}) + \eta D_{mis}). \tag{18}$$

Using Lemma 1, we have

$$f(D_{mis}|D_{obs}, R=0) = f(D_{mis}|D_{obs}, R=1)\frac{\exp(-\eta D_{mis})}{E(\exp(-\eta D_{mis})|D_{obs}, R=1)}. \tag{19}$$

The sensitivity parameter η determines the amount of departure from the ignorability of the MDM. In formula (19), we need to compute the conditional distribution of missing data $f(D_{mis}|D_{obs}, R=1)$. A consistent estimate of $f(D_{mis}|D_{obs}, R=1)$ can be obtained parametrically or non-parametrically; an example is shown below.

To use the SSA approach we first choose a value of η from a predetermined set Γ. For each η we simulate missing data from either (8) or (19). Together, the (simulated) missing and the observed data form a dataset which is complete. From here, a subset D_η^{obs} can be re-sampled using the MDM specified in (17) and the KNN distance between D_η^{obs} and the true observed data D_{obs} can finally be calculated.

We now conduct a simulation study for non-ignorable missing covariate problem based on a real data example.

The US Federal Highway Administration published fuel consumption data over 50 states and the District of Columbia in 2001; see Weisberg (2005) (Chapter 1, page 15). The aim of the research is to understand the effect on fuel consumption (T) of several covariates namely, the number of federal-aid highway miles (X_1), the personal income (X_2), the number of licensed drivers (X_3) and the state gasoline tax rate (X_4).

We consider the following linear regression model

$$t = \theta_0 + \sum_{j=1}^{4} \theta_j x_j + \epsilon. \tag{20}$$

By using the complete dataset, estimates for $(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ can be obtained and they are presented in the first row in Table 3 .

We now consider a missing data problem by letting income (X_2) to be partly missing with probability $P(R=0|D) = 1 - \text{expit}(1 + (x_1 - \bar{x}_1) - 0.5(x_2 - \bar{x}_2))$, where R is the missing data indicator and $\bar{x}$ is the sample mean. This is a non-ignorable missing data problem.

To use the SSA approach we resort to the semi-parametric selection model in (17), i.e.

$$P(R=1|t, x_1, x_2, x_3, x_4) = \text{expit}(h(t, x_1, x_3, x_4) + \eta x_2). \tag{21}$$

Following the discussion around (19), we simulate the missing data from:

$$f(x_2|t, x_1, x_3, x_4, R=0) = f(x_2|t, x_1, x_3, x_4, R=1)\frac{\exp(-\eta x_2)}{E(\exp(-\eta x_2)|t, x_1, x_3, x_4, R=1)},$$

where $f(x_2|t, x_1, x_3, x_4, R=1)$ is modelled by the following parametric model:

$$(x_2|t, x_1, x_3, x_4, r=1) \sim N(\gamma_0 + \gamma_1 t + \gamma_2 x_1 + \gamma_3 x_3 + \gamma_4 x_4, \tau^2),$$

where the unknown parameters $(\gamma_0, \gamma_1, \ldots, \gamma_4, \tau^2)$ can be estimated from the observed data and X_2 (personal income) is assumed to be normally distributed.

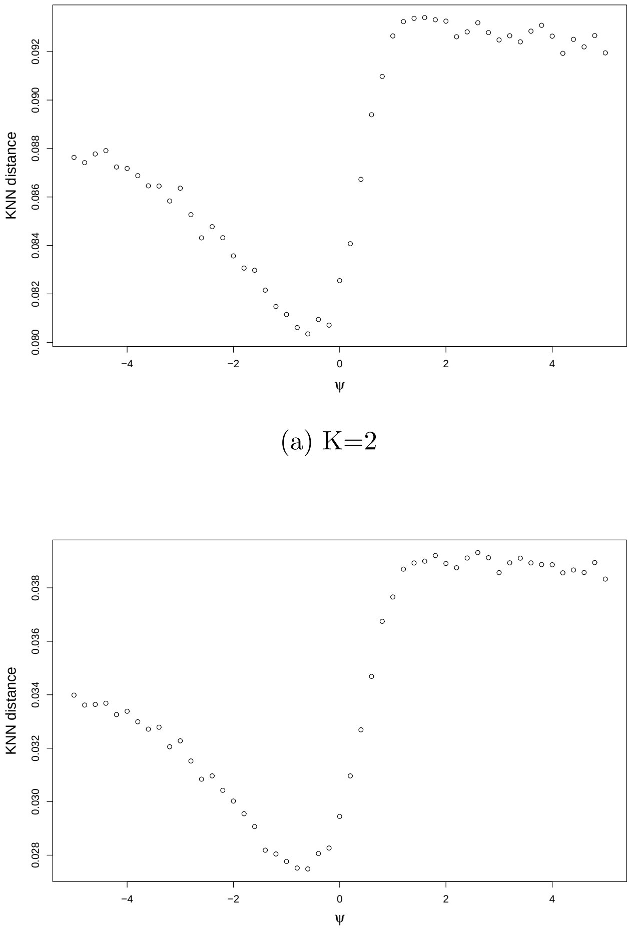
The simulated data for missing x_2 and the observed data can now be used to form a 'complete' dataset $D_\eta = (t, x_1, x_2^*, x_3, x_4)$ for each given η. Using D_η we can estimate the parameter $(\hat{\theta}_\eta, \hat{\sigma}_\eta^2)$ from the linear regression model (20). In order to make the SSA method more reliable numerically, we build on the idea of 'bootstrapping residuals' introduced in (Efron and Tibshirani, 1993). Instead of using D_η directly, we use $D_\eta^* = (t^*, x_1, x_2^*, x_3, x_4)$, where t^* is re-sampled conditionally on the following linear regression model with estimates $(\hat{\theta}_\eta, \hat{\sigma}^2)$ and imputed covariates $x^* = (x_1, x_2^*, x_3, x_4)^T$:

$$t^*|x_1, x_2^*, x_3, x_4 \sim N(\hat{\theta}_\eta^T x^*, \hat{\sigma}_\eta^2),$$

i.e. t^* is simulated by adding residuals on the predicted values, where the residuals are sampled from a normal distribution $N(0, \hat{\sigma}_\eta^2)$. And, finally, we then calculate the distance between $D_\eta^{obs*} = (t^*, x_1, x_3, x_4)$ and the observed dataset $D_{obs} = (t, x_1, x_3, x_4)$ by using an average distance a MC size of 100 (this will reduce the sampling error).

In this example η is selected from within the interval $(-5, 5)$ at equal-distance of 0.2 unit. Figure 6 shows the KNN distances with $K = 2$ against the values of η. As it can be seen, the 'most plausible' value or minimum distance is achieved at around

$\eta = -0.6$. It is also interesting to note that the corresponding estimates are very close to the ones obtained from the complete data. The ASL using the permutation test has also been calculated but its value is well above the critical value of 0.05 and hence in not showing in the plot.



(a) K=2

(b) K=3

Figure 6: KNN distance for the fuel consumption data.

The simulation study results with 100 replications are presented in Table 3. The row named as 'SSA' shows the average values for the 'most plausible estimate'. As seen, all the estimates are very close to the estimates calculated from the complete data except $\hat{\theta}_2$. This is further evidence that the most plausible model obtained via a SSA study needs to be chosen with caution. It may provide the 'best' results based on the observed data, but it may not be consistent.

As comparison, we also considered the complete case analysis, using multivariate imputation under MAR ('MICE' Package with Bayesian Linear regression method; see van Buuren and Groothuis-Oudshoorn (2011)) and multivariate imputation under MNAR (iteration step for sensitivity analysis, see Resseguier (2011)). Note the latter method modifies the imputation model under a MAR hypothesis to MNAR hypothesis by specifying a supplementary parameter (SupPar). For binary or categorical variables, this parameter is the odds ratio (i.e. the ration between the odds of the modality interest among subjects with missing values and subjects without missing value). Likewise, for continuous variables, this is the difference in expected values. As a comparison, we have also applied Monte Carlo sensitivity analysis (MCSA) proposed by Greenland (2005) and the Bayesian sensitivity analysis (BSA) proposed by McCandless et al. (2007). The prior distribution $f(\eta)$ is selected as U(-5,5) for both the MCSA and the BSA methods. As it can be seen in the simulation results with 100 replications presented in Table 3, overall the SSA method performs more favorably and robustly than the others.

Table 3: Simulation study for fuel consumption data

	$\hat{\theta}_0$ (se)	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
Complete data	154.19(9.08)	18.55(6.47)	-6.14(2.19)	0.47(0.13)	-4.23(2.03)
Complete case	155.06(14.44)	16.20(12.41)	-7.32(3.37)	0.24(0.18)	-7.29(2.24)
Multiple Imputation:					
MAR	130.15(13.11)	30.68(8.5)	-9.86(4.13)	0.50(0.15)	-5.17(2.17)
MNAR (SupPar=1)	137.70(12.03)	24.24(8.95)	-9.74(3.91)	0.43(0.14)	-5.42(2.37)
MNAR (SupPar=2)	141.26(11.27)	22.84(8.87)	-9.68(3.65)	0.42(0.14)	-5.33(2.34)
MNAR (SupPar=5)	152.11(9.69)	19.45(8.27)	-8.19(2.88)	0.42(0.14)	-5.01(2.24)
MNAR (SupPar=8)	159.49(9.36)	17.69(7.73)	-6.34(2.24)	0.43(0.14)	-4.75(2.17)
Sensitivity Analysis:					
SSA (K=2)	152.30(8.29)	18.24(5.59)	-3.99(2.18)	0.49(0.12)	-4.39(1.85)
SSA (K=3)	154.12(8.22)	17.01(5.57)	-4.05(2.08)	0.49(0.11)	-4.38(1.83)
MCSA	153.62(13.22)	22.87(6.94)	-3.44(1.87)	0.49(0.13)	-4.29(2.07)
BSA	131.51(10.58)	31.77(6.49)	-10.17(2.82)	0.45(0.12)	-4.75(2.00)