JMLR: Workshop and Conference Proceedings vol 52, 449-461, 2016 PGM 2016

Estimating Mutual Information in Under-Reported Variables

Konstantinos Sechidis KONSTANTINOS.SECHIDIS@MANCHESTER.AC.UK
School of Computer Science
University of Manchester (UK)

Matthew Sperrin MATTHEW.SPERRIN@MANCHESTER.AC.UK
Centre for Health Informatics, Institute of Population Health
University of Manchester (UK)

Emily Petherick E.PETHERICK@LBORO.AC.UK
School of Sport, Exercise & Health Sciences
Loughborough University (UK)

Gavin Brown GAVIN.BROWN@MANCHESTER.AC.UK
School of Computer Science
University of Manchester (UK)

Abstract

Under-reporting occurs in survey data when there is a reason to systematically misreport the response to a question. For example, in studies dealing with low birth weight infants, the smoking habits of the mother are very likely to be misreported. This creates problems for calculating effect sizes, such as bias, but these problems are commonly ignored due to lack of generally accepted solutions. We reinterpret this as a problem of learning from missing data, and particularly learning from positive and unlabelled data. By this formalisation we provide a simple method to incorporate prior knowledge of the misreporting and we present how we can use this knowledge to derive corrected point and interval estimates of the mutual information. Then we show how our corrected estimators outperform more complex approaches and we present applications of our theoretical results in real world problems and machine learning tasks.

Keywords: Under-reported; misclassification bias; missing data; mutual information.

1. Introduction

Smoking during pregnancy is a key risk factor for adverse outcomes, including preterm birth and low birth weight (LBW). Like many health behaviours, accurate measurement of smoking habits can be difficult and expensive during pregnancy. For that reason, many studies use *self-reported* data, e.g. Wright et al. (2013). Given that most smokers know their habit to be harmful both to themselves and their unborn child, there are strong motivations for women to *under-report* or deny their smoking status (Dietz et al., 2011). As such, the frequency of smokers in a sample is expected to be less than the true frequency. Any statistical measure of association between this *under-reported* (UR) variable and another will be biased in a manner that is specific to the nature of the association measure itself, and the degree/pattern of under-reporting. Thus, any policy decisions made on the basis of such a biased measure will be questionable.

The only guaranteed method to completely correct for this is to manually identify individuals that are likely to have misreported, and ignore/correct their testimony, raising issues of data privacy. As an alternative to this, authors in medical statistics treat it as a problem of *misclassification bias*, and combine data with a prior belief of the pattern of misclassification, to derive corrected estimators

for the log-odds ratio (Chu et al., 2006; Edwards et al., 2013), and the relative-risk (Rahardja and Young, 2011). These solutions suffer from a number of weaknesses, which are addressed in this paper. Firstly, they naturally only apply to binary data – arbitrary categorical data is handled via a one-vs-one or one-vs-all strategy. Secondly, conditional estimation with respect to other categorical variables are problematic for similar reasons. Due to these weaknesses, ranking of variables in relation to a target – a common need in feature selection and other machine learning tasks – is not straightforward. One way to overcome this limitation is to derive corrected estimators for the *mutual information* (MI), a measure of effect size widely used in machine learning applications with several nice properties (Brillinger, 2004).

To derive this estimator we reinterpret the challenge not as dealing with misclassification and biased data, but as a problem of *learning from missing data*. We present solutions based on mutual information and a graphical representation called *missingness graphs* (Mohan et al., 2013) – this naturally handles categorical data, and incorporates a prior belief of the misreporting at the population (or appropriate sub-demographic) level. This is made possible by examining independence properties observable via the *m*-graph representation.

In this paper, we present the following novel contributions: (1) Consistent and efficient estimators of the mutual information between an UR variable and an arbitrary categorical variable, for both conditional and unconditional MI, including interval estimates; (2) a case study using 13,776 births in the north of England, demonstrating some significant false conclusions that might be drawn when ranking variables without correcting for UR; (3) an application using our estimators for feature selection when training/test distributions differ.

2. Background Material

To the best of our knowledge, our work is the first that tackles the problem of estimating MI in under-reporting scenarios. In classic statistics there are some works that estimate other types of effect sizes (i.e. odds/risk ratios, obviously limited to binary data) and we review them in Section 2.1. Section 2.2 shows how the under-reported can be phrased as missing data problem. Finally, Section 2.3 gives the background on estimating MI.

2.1 Under-reporting as Misclassification Bias Problem

We assume that we have two random variables X and Y, representing a scenario where X is likely to be UR. In this case, we cannot observe the true value of X, but instead receive observations from a proxy variable $\widetilde{X}$. In the notation below we use lower case letters $(y, x, \widetilde{x})$ to denote a realisation from these variables. In our example of smoking during pregnancy, $y \in \{0, 1\}$, is an indicator[1] of LBW, $x \in \{0, 1\}$ is whether the mother smoked during pregnancy (1 for smoking and 0 for not smoking), and $\widetilde{x} \in \{0, 1\}$ is whether the mother *reported* that she smoked in pregnancy (1 for reported smoking and 0 for not smoking).

A classical solution to the under-reporting problem is to consider it as *misclassification bias* (Greenland, 2014). Following their terminology, for an under-reported variable, the *specificity* is $p(\widetilde{x} = 0|x = 0) = 1$, while the *sensitivity* is $p(\widetilde{x} = 1|x = 1) < 1$. Here, the specificity is the probability that a non-smoker would tell the truth (equal to 1 in this setting) and the sensitivity is the probability that a smoker would tell the truth (in our setting strictly < 1). As presented, this is

1. The techniques presented in this work are also applied to categorical data with more than two levels $|\mathcal{Y}| > 2$.

the simplest scenario – referred to as *non-differential* — that is, the probabilities do not vary with respect to Y. The more complex case is when the sensitivity depends on Y, that is $p(\widetilde{x} = 1|x = 1, y)$, known as *differential misclassification* (Greenland, 2014). In this work, we will focus on the non-differential UR scenario, and leave the differential as a future work, outlined in Section 8.

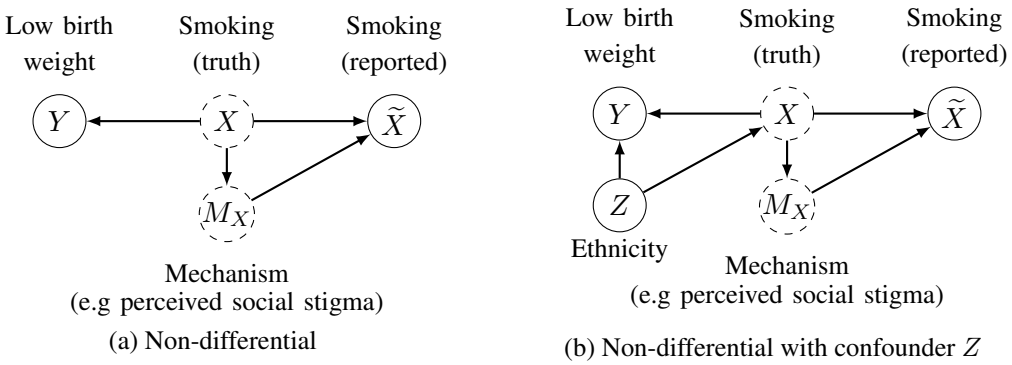
Estimating the strength of association between variables, using this misclassification approach, is a well explored challenge in epidemiology. For example, Chu et al. (2006) derive corrected estimators for the log-odds ratio, while Rahardja and Young (2011) for the relative-risk. To derive these corrections, knowledge over the specificities/sensitivities, or in other words knowledge of the misclassification rates, is needed. This knowledge can be derived in different ways, such as *validation studies* or domain *prior knowledge*. A different way of estimating these effect sizes is to use a model to impute the values of the possibly misclassified examples, for example Edwards et al. (2013) present a way of using multiple imputations to estimate log-odds ratios. With our work we derive corrections for the mutual information, by incorporating simple forms of prior knowledge.

2.2 Under-reporting as Missing Data Problem

A different way to phrase the under-reporting problem is by connecting it with the equivalent problem from the missing data literature. The first step is to consider the under-reporting bias as a *positive and unlabelled* (PU) problem (Elkan and Noto, 2008). That is, a semi-supervised binary classification problem where we have a set of positive examples and a separate set of unlabelled examples, which can be either positive or negative. The positive examples can be seen as the reported "smoking" cases ($\widetilde{x} = 1$), while the unlabelled can be seen as the reported "non-smoking" cases ($\widetilde{x} = 0$).

Furthermore, from the missing data literature we borrow a graphical representation which will help us to make apparent the assumptions behind the under-reporting mechanism. Mohan et al. (2013) introduced a formalism for graphical modelling in the presence of missing data – known as *missingness graphs* or m-graphs. While in the literature of misclassification bias there is a different graphical representation (Greenland, 2014), our modification of the m-graphs provides more useful information, by capturing both the data generation model and the causal mechanisms responsible for the misclassification process.

Figure 1a shows the simplest case of *non-differential* under-reporting. A solid node indicates a fully observed variable, whilst dashed nodes represent unobserved variables. Associated with every unobservable variable X there are two additional nodes. Firstly M_X, which controls whether a value from X is correctly reported ($m_x = 1$) or not ($m_x = 0$). And secondly, the proxy variable $\widetilde{X}$ which is fully observed. The major difference between missingness graphs used by Mohan et al. (2013) and those here is that the mechanism M_X is not observable, and for that reason we must incorporate prior knowledge over the sensitivity $p(m_x = 1|x = 1)$ and specificity $p(m_x = 1|x = 0)$. Figure 1b shows the more realistic situation where *ethnicity* has an effect on both the smoking status and the LBW. The current paper shows (in its simplest case) how to recover the value $I(X;Y)$ from $I(\widetilde{X};Y)$ by deriving a correction based on prior belief over the mechanism M_X. The m-graph representation allows us to read off independence properties such as: $Y \perp\!\!\!\perp M_X|x = 1$ — which corresponds to the *selected completely-at-random* assumption in the positive and unlabelled literature (Elkan and Noto, 2008). Sechidis et al. (2014) suggested informed ways of testing independence in PU data, while our work focuses on the estimation of MI.



(a) Non-differential

(b) Non-differential with confounder Z

Figure 1: A graphical representation for under-reporting. (a) Non-differential, where low birth weight Y is assumed to be associated with smoking X, so we want to know the strength of association $I(X;Y)$ on this arc. However, X is *under-reported*, so the true value is unobservable, and instead we have a proxy $\widetilde{X}$, determined by X and the misclassification mechanism M_X. (b) Non-differential with a confounding variable Z, in this case we are also interested in the strength of the conditional (adjusted) association $I(X;Y|Z)$.

2.3 Estimating Mutual Information

In practical applications we want to explore relationships between random variables. Just giving a yes/no answer through a hypothesis test may not be of much interest, and estimating the size of the effect gives more useful information. For example, how strongly smoking is correlated with low birth weight. In machine learning one of the main ways of measuring the strength of this association is by estimating Shannon's mutual information (MI) (Brillinger, 2004). The maximum likelihood (ML) estimate of the MI is:

$$\widehat{I}(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \widehat{p}(x,y) \ln \frac{\widehat{p}(x,y)}{\widehat{p}(x)\widehat{p}(y)}.$$

Asymptotic distribution theory has a set of tools to derive the sample distribution of the ML-MI estimator and the following theorem presents this known result (Brillinger, 2004).

Theorem 1 (ML-MI estimator, asymptotic distribution)
For the estimator $\widehat{I}(X;Y)$ it holds that: $\sqrt{n}\left(\widehat{I}(X;Y) - I(X;Y)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{MI}\right)$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. The standard error of the estimator is:

$$SE\left[\widehat{I}(X;Y)\right] = \frac{\sigma_{MI}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \left(\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} \right)^2 - I(X;Y)^2 \right)^{\frac{1}{2}} \quad (1)$$

Proof sketch: This result can be proved by using delta methods (Agresti, 2013).

While the asymptotic variance here depends on the population values $p(x,y)$, in practice for interval estimation we replace them by their sample values $\widehat{p}(x,y)$. This standard procedure (Agresti, 2013, Section 3.1.7) is followed for all the sampling distributions we present in this work. The distribution of the conditional mutual information: $\widehat{I}(X;Y|Z) = \sum_{z\in\mathcal{Z}} \widehat{p}(z)\widehat{I}(X;Y|z)$ can be derived in a similar manner, using the same methodology as we did for the unconditional.

3. Motivating the Problem of Estimating MI in Under-Reported Scenarios

The *ideal* method to completely correct UR is to spend resources to identify the individuals that have misreported, and correct their testimony. For example, it could be done by performing cotinine blood tests to all women that reported non-smoking ($\widetilde{x} = 0$). This approach is expensive, and it also raises issues of data privacy. On the other hand, the simplest way to estimate mutual information in under-reported scenarios is to follow a *naive* approach and just use the observed data. Unfortunately, this estimator, $\widehat{I}(\widetilde{X}; Y)$, is not consistent for estimating $I(X; Y)$. This can be easily proved, since under the model of Figure 1a the following strict inequality holds[2]: $I(\widetilde{X}; Y) < I(X; Y)$.

Another way to estimate mutual information is by trying to "predict" the real values of the misclassified examples using some prediction model. Then, impute new values for this examples, and finally, estimate MI using the imputed data. This is similar to solving the missing data problem by *imputation* (Allison, 2001). In our running example this means imputing the actual values of the women who reported not smoking ($\widetilde{x} = 0$). To do so we need to build a model to derive the Bayesian posterior distribution $p(x = 1|y, \widetilde{x} = 0)$, and using this model to impute the values for the examples with $\widetilde{x} = 0$. Then, we can use these imputed values to derive point and interval estimates of the MI using the expressions presented in Section 2.3. One limitation of single-imputation is that estimating standard error using conventional methods –such as eq. (1)– does not take into account the fact that some of the data were imputed (Rubin, 2004). One solution to this problem is to perform multiple-imputations and use improved ways of estimating the standard errors, such as Rubin's rule presented in (Allison, 2001, Chapter 5). But also multiple-imputation has some limitations; for example, it is computationally expensive, while, in the case of estimating MI, there are no guarantees that the confidence intervals derived by Rubin's rule will have the coverage defined by the nominal (user specified) level. For more details on the strengths and weaknesses of multiple-imputation we refer to Rubin (2004).

Our contribution: In the next section we present a corrected estimator for the mutual information that takes into account the under-reporting and overcomes the above limitations: (1) it is consistent, unlike the naive approach, (2) it produces valid interval estimates, unlike the simple-imputation, and (3) it is computational-efficient/imputation-free, unlike the multiple-imputations.

4. Correcting Mutual Information for Under-Reporting

To estimate mutual information in the under-reported scenario, we need to come up with a way to estimate marginal and joint/conditional probabilities, despite the restrictions of the problem. While we can estimate the marginal $p(y)$ from all data, the conditionals are more challenging. For example, the conditional $p(y|x = 1)$ is inaccessible, as we do not have access to the full set of the examples with $x = 1$, i.e. we do not know the identities of all smokers, but only those that self-reported it ($\widetilde{x} = 1$). Because of the event based independence assumption $Y \perp\!\!\!\perp M_X|x = 1$ it holds that $p(y|x = 1) = p(y|x = 1, m_x = 1) \Leftrightarrow$

$$p(y|x = 1) = p(y|\widetilde{x} = 1). \tag{2}$$

To find the other conditional $p(y|x = 0)$ we use a simple trick first introduced by Denis et al. (2003) in the context of positive and unlabelled data. By using (2) we can write the marginal as

2. We can prove this result by using Jensen's inequality, and the fact that in non-differential under-reporting the following strict inequality holds: $p(\widetilde{x} = 1) < p(x = 1)$.

$p(y) = p(y|\widetilde{x}=1)p(x=1) + p(y|x=0)p(x=0)$ and solving for $p(y|x=0)$:

$$p(y|x=0) = \frac{p(y) - p(y|\widetilde{x}=1)p(x=1)}{1-p(x=1)}. \tag{3}$$

Finally, since we do not have access to the marginal distribution $p(x=1)$, and since it cannot be estimated without modelling assumptions, we incorporate prior knowledge[3] as a parameter γ, provided by a user's belief over the true prevalence $p(x=1)$. Incorporating prior knowledge over the true prevalence is a widely used approach in the positive and unlabelled literature (Sechidis et al., 2014).

By assuming perfect knowledge over the prevalence $\gamma = p(x=1)$ and using only the observed variables Y and $\widetilde{X}$ we can estimate $I(X;Y)$ using the following corrected estimator.

Definition 2 (Corrected ML-MI estimator)
The corrected for under-reporting estimator of the mutual information is:

$$\widehat{I}_\gamma(\widetilde{X};Y) = \sum_{y\in\mathcal{Y}} \left(\gamma\, \widehat{p}(y|\widetilde{x}=1) \ln \frac{\widehat{p}(y|\widetilde{x}=1)}{\widehat{p}(y)} + (\widehat{p}(y) - \gamma\widehat{p}(y|\widetilde{x}=1)) \ln \frac{\widehat{p}(y) - \gamma\widehat{p}(y|\widetilde{x}=1)}{\widehat{p}(y)\,(1-\gamma)} \right).$$

To prove that the estimator is consistent is straightforward, since when $\gamma = p(x=1)$, by using (2), (3) it holds that $I_\gamma(\widetilde{X};Y) = I(X;Y)$. Proving only the consistency of the corrected estimator is not so useful, and we need to capture also the variance that it has in the finite sample size. We do so by the following theorem.

Theorem 3 (Corrected ML-MI estimator, asymptotic distribution)
For the estimator $\widehat{I}_\gamma(\widetilde{X};Y)$ it holds that: $\sqrt{n}\left(\widehat{I}_\gamma(\widetilde{X};Y) - I(X;Y)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{MI_\gamma}\right)$, when we have perfect prior knowledge $\gamma = p(x=1)$. The standard error is:

$$SE\left[\widehat{I}_\gamma(\widetilde{X};Y)\right] = \frac{\sigma_{MI_\gamma}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \left(\sum_{\widetilde{x}\in\widetilde{\mathcal{X}}, y\in\mathcal{Y}} \left(p(\widetilde{x},y)\phi^2_{\widetilde{x},y} \right) - \left(\sum_{\widetilde{x}\in\widetilde{\mathcal{X}}, y\in\mathcal{Y}} \left(p(\widetilde{x},y)\phi_{\widetilde{x},y} \right) \right)^2 \right)^{\frac{1}{2}}, \tag{4}$$

$\phi_{\widetilde{x}=0,y} = \ln \frac{p(y)-\gamma p(y|\widetilde{x}=1)}{p(y)}$, $\phi_{\widetilde{x}=1,y} = \phi_{\widetilde{x}=0,y} + \frac{\gamma}{p(\widetilde{x}=1)} \sum_{y'\in\mathcal{Y}} \left(p(y'|\widetilde{x}=1) - \delta_{yy'} \right) \ln \frac{p(y')-\gamma p(y'|\widetilde{x}=1)}{\gamma p(y'|\widetilde{x}=1)}$.

Proof sketch: This result can be proved by using delta methods.

Furthermore, we can use our corrected estimator to estimate conditional effects under non-differential UR (Figure 1b). For example we can consistently estimate $I(X;Y|Z)$ by using the estimator: $\sum_{z\in\mathcal{Z}} \widehat{p}(z)\widehat{I}_{\gamma'}(X;Y|z)$, where $\gamma' = p(x=1|z)$. To estimate γ' we need only knowledge of $\gamma = p(x=1)$, since under the model of Figure 1b it holds that:

$$\gamma' = p(x=1|z) = \frac{p(z|x=1)p(x=1)}{p(z)} = \frac{p(z|\widetilde{x}=1)}{p(z)}\gamma$$

and both $p(z|\widetilde{x}=1)$ and $p(z)$ can be consistently estimated by the observed data. The conditional sampling distribution remains normal with similar parameters to the one described in Theorem 3, but also converges more slowly because γ' is estimated.

3. Using prior knowledge over $p(x=1)$ is equivalent to using prior knowledge over the sensitivity, an approach that is followed to correct the misclassification bias of epidemiological effect sizes (Greenland, 2014). This can be shown by the fact that in the non-differential under-reporting it holds that: Sensitivity $= p(m_x=1|x=1) = \frac{p(m_x=1,x=1)}{p(x=1)} = \frac{p(\widetilde{x}=1)}{p(x=1)}$, and the $p(\widetilde{x}=1)$ can be estimated by the observed data.

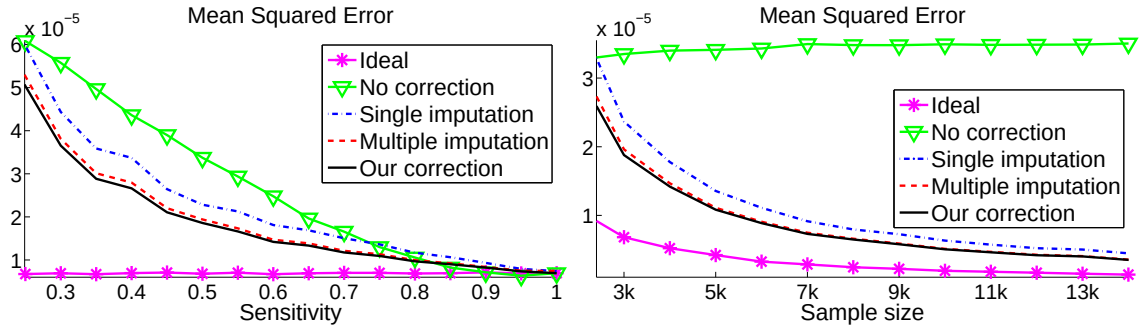
5. Experiments with Synthetic Data: Perfect Prior Knowledge

As a "sanity check" for our theoretical results we generated synthetic random variables X and Y with different degrees of dependency. To create the data, firstly we generate the values of X, by taking N samples from a Bernoulli distribution with parameter $p(x = 1)$. Then, we randomly choose the parameters $p(y|x)$ that guarantee the desired degree of dependency, expressed in terms of $I(X;Y)$, and we use these parameters to sample the values of Y. To create the under-reported variable $\widetilde{X}$ we sample with Sensitivity$= p(\widetilde{x} = 1|x = 1)$ the examples with $x = 1$. We estimate mutual information using five different methods:

- **Ideal:** using the unobservable estimator $\widehat{I}(X;Y)$ and eq. (1) for standard error,
- **No correction:** using the under-reported estimator $\widehat{I}(\widetilde{X};Y)$ and eq. (1) for standard error,
- **Single imputation:** using a model to impute possible misclassified data and then estimate MI and standard error by eq. (1),
- **Multiple imputations:** using a model to impute multiple[4] times and then average MI across the imputed datasets and using Rubin's rule (Allison, 2001) for standard error.
- **Our correction:** using our corrected estimator $\widehat{I}_\gamma(\widetilde{X};Y)$ and eq. (4) for standard error.

To get a fair comparison between the last three methods[5], we used the same modelling assumptions and γ is assumed to known and equal to $p(x = 1)$.

Figure 2 compares the five methods in terms of their mean squared error. The three methods that take into account the under-reporting (single/multiple imputation and our corrected estimator) outperform the naive estimator, which is not consistent. As the sample size/sensitivity increases, all of these three approaches tend to behave the same in a similar way to the ideal estimator. Our corrected estimator outperforms the imputation-based approaches, especially in small sample sizes and small levels of sensitivity – which are the most challenging situations. Interestingly, our method clearly outperforms methods with the same complexity (no correction and simple imputation).



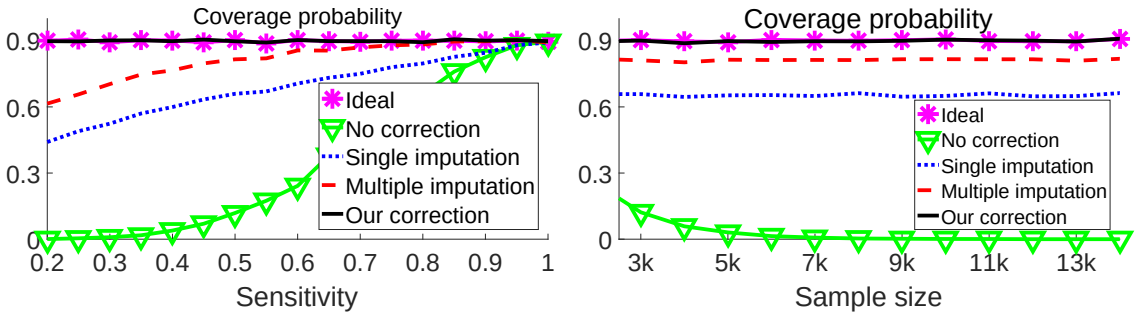
(a) Different levels of UR with $N = 3$k (b) Different sample sizes with Sensitivity = 0.50

Figure 2: Comparison in terms of mean (over $5{,}000$ repetitions) squared error. In each repetition we set $I(X;Y)$=0.01 and we randomly choose: $|\mathcal{Y}| \in \{2-5\}$ and $p(x = 1) \in \{0.1-0.5\}$.

4. To decide this number, we used the White et al. (2011) guideline that the number of imputations should be approximately 100 times the fraction of missing information. In under-reporting this can be phrased as using $100\times$(1-Sensitivity) imputations.

5. For the imputation-based approaches, we imputed the potentially misclassified examples by the following posterior, which can be naturally derived by the model of Figure 1a: $p(x = 1|y, \widetilde{x} = 0) = \frac{p(y,\widetilde{x}=0|x=1)\gamma}{p(y,\widetilde{x}=0)} = \frac{(p(y|x=1)-p(y,\widetilde{x}=1|x=1))\gamma}{p(y,\widetilde{x}=0)} = \frac{p(y|\widetilde{x}=1)(\gamma-p(\widetilde{x}=1))}{p(y,\widetilde{x}=0)}$. As we mentioned, we use perfect prior knowledge over γ, while the rest of the parameters are estimated through ML from the observed data.

Figure 3 verifies that the suggested standard error in Theorem 3 is correct, and that our method is a valid way to derive interval estimates, similar to those derived using the ideal estimator. In this figure we estimate the proportion of times that the 90% confidence intervals, derived by using different standard errors for the different methods, contain the true value of the mutual information $I(X;Y)$. Since the estimated coverage probability for the ideal and our proposed method are at the nominal (user specified) level of 90% we can conclude that only these methods produce accurate interval estimates.



(a) Different levels of UR with $N = 5$k

(b) Different sample sizes with Sensitivity = 0.50

Figure 3: Comparing in terms of coverage. We set the nominal level to be 0.90 (90% confidence intervals) and we observe the proportion (over $5{,}000$ repetitions) of the times that suggested intervals contain the true value of the mutual information.

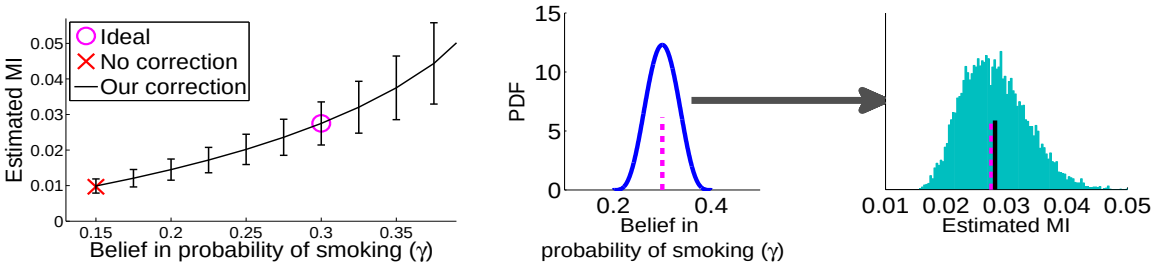
6. Experiments with Synthetic Data: Uncertain Prior Knowledge

Perfect prior knowledge, i.e. $\gamma = p(x = 1)$, will not always be available. Therefore it is important to explore ways to deal with uncertain prior knowledge and examine the behaviour with incorrect priors – results are presented below for an artificial scenario where we can exert control over the "quality" of prior knowledge.

Let us assume that birth weight of non-smoker births are drawn from a normal distribution with $\mu = 3500$g and $\sigma = 500$g, while birth weight of smoker births are drawn from a normal distribution with $\mu = 3000$g and $\sigma = 500$g. Birthweight was considered to be "low", $p(x = 1)$, if the weight was < 2500g (Wright et al., 2013). We assume that in a cohort of $N = 5000$ pregnant mothers, 30% are smokers, so $p(x = 1) = 0.3$. However, only half of the mothers on average would admit to this, so $p(\widetilde{x} = 1) = 0.15$. In a typical simulated draw from this simulation, the mutual information is estimated with an under-reported variable. However, after using our corrected estimator and by incorporating the prior knowledge that the X variable is non-differential under-reported, the estimated mutual information increases by a factor of three (Figure 4a).

One way to handle uncertain prior knowledge is by performing a *sensitivity analysis* as Figure 4a shows. To do so we plot the interval estimates for the corrected MI, calculated by eq. (4), for different values of our belief over the probability of smoking (γ). Interestingly, the 80% confidence interval estimate for $\gamma = p(x = 1) = 0.30$ (perfect knowledge) contains the true (ideal) value of the MI. A different way to handle uncertainty is through a *simulation based analysis*, where we represent uncertainty over γ as a probability distribution, sample from this distribution many times, and estimate the corrected MI for each value. For example, in Figure 4b we model γ as a generalised Beta distribution (bounded between a minimum and a maximum value) and we explore the resultant

uncertainty in the point estimate of the corrected mutual information. As we observe, the true value of the mutual information is very close to the average over the simulations.



(a) Sensitivity analysis. (b) Simulation based analysis.

Figure 4: Different ways to handle uncertain prior knowledge. (a) Sensitivity analysis. (b) Simulation based analysis: [LEFT] the user's prior belief over γ, [RIGHT] the resultant uncertainty in the estimated mutual information through our correction. The dashed line shows the true (but unknown) value, while the solid line the average over the simulations.

7. Applications in Real-world Datasets

In this section we present two applications of our results — ranking the risks factors that may lead to adverse birth outcomes derived from a large real-world dataset, and feature selection when the training/test distributions differ.

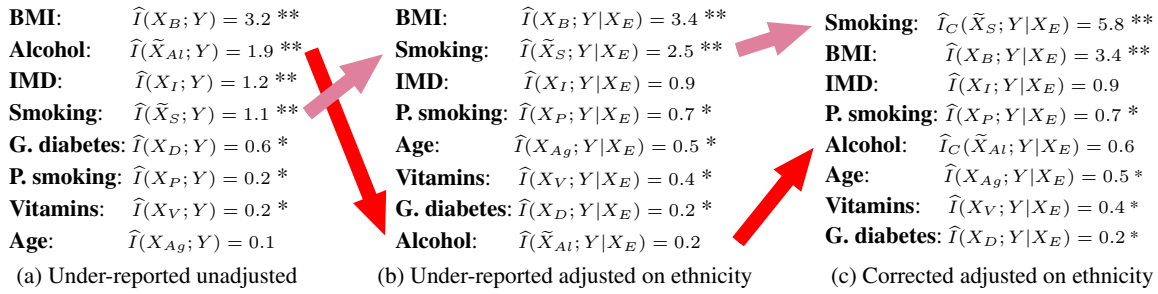
7.1 Risk Factors for Low Birth Weight Infants

To describe the usefulness of our theoretical findings we will use data from a prospective birth cohort, the Born in Bradford (BiB) study. BiB is a longitudinal multi-ethnic birth cohort study aiming to examine the impact of environmental, psychological and genetic factors on maternal and child health and well-being (Wright et al., 2013). Bradford is a city in northern England with high levels of socio-economic deprivation and ethnic diversity. The full BiB cohort recruited 12,453 women comprising 13,776 pregnancies between 2007 and 2010 and the cohort is broadly characteristic of the city's maternal population in terms of age, deprivation and ethnicity (Wright et al., 2013). Ethics approval for the study was granted by Bradford Research Ethics Committee (Ref 07/H1302/112).

In our experiments we will focus on ranking several factors according to their association with LBW. We focus on the following correctly-reported *categorical* variables: ethnicity X_E (3 levels), age X_{Ag} (3 levels), Body Mass Index (BMI) X_B (4 levels), index of multiple deprivation X_I (5 levels), gestational diabetes X_G (binary), taken vitamins X_V (binary), passive smoking X_P (binary) and the following binary UR variables: any smoking $\widetilde{X}_S$ and alcohol $\widetilde{X}_{Al}$ consumption during pregnancy. For the UR variables the observed priors are that 16% of the women smoked during pregnancy, while 31% drunk alcohol. Using domain knowledge we correct the priors to be 25% and 40% respectively, in other words we assume that 9% of the overall women UR these two habits, and we assume non-differential UR.

In Figure 5(a) we observe the ranking by using the unconditional MI of the observed covariates and the target variable (LBW). Due to the multi-ethnic characteristic of the sample, it is important to control (adjust) for ethnicity. In Figure 5(b) we see the derived ranking by using the conditional MI between the observed covariates and the target, conditioning over the ethnicity variable. Inter-

estingly, in this ranking *Alcohol* is the least significant factor, while in the overall population it was second. Finally, to correct the two UR variables, we use prior knowledge and our corrected conditional estimators presented in Sect. 4, and we derive the ranking of Figure 5(c). After the correction we see that the dominant factor is *Smoking*, even above *BMI*, while *Alcohol* moves back up.



(a) Under-reported unadjusted (b) Under-reported adjusted on ethnicity (c) Corrected adjusted on ethnicity

Figure 5: Variable ranking by their association with the LBW. (a) Ranked by MI, uncorrected. (b) Ranked by MI condition on ethnicity, uncorrected. (c) Ranked by MI condition on ethnicity, corrected. Units are milli-nats. The single star * means the null hypothesis (independence between the reported covariate and LBW) is rejected at $\alpha = 0.1$, while double stars ** at $\alpha = 0.01$. It should remembered that failure to reject the null does not imply insignificance as the test may not have sufficient power, which is likely the case in an UR test due to the power-loss (Greenland, 1988).

The differences between the three rankings illustrate the importance of having techniques that are able to produce estimates that are adjusted on some demographic characteristics and that also able to correct under-reporting. In the following section we present the merits of our analysis in a machine learning application using UCI datasets for which we have access to the ground truth.

7.2 Feature Selection with Event-level Covariate Shift

Covariate shift (Quionero-Candela et al., 2009) is when training/test distributions differ, in that $p_{test}(x) \neq p_{train}(x)$, but we still have the same posteriors $p_{test}(y|x) = p_{train}(y|x)$. The case of non-differential UR features in training data can be seen as an *event-level covariate shift*, since, because of eq. (2), we have $p_{test}(y|x = 1) = p_{train}(y|x = 1)$ but not for $x = 0$.

In the experiments of this section we perform mutual information based feature selection (Brown et al., 2012), where some of the features are under-reported during the feature selection step. Then, we build a model using the training data and estimate the classification error in the testing data. We used three categorical UCI datasets with different characteristics[6]. As a classification model we used a k-nearest neighbour ($k = 3$), this is chosen as it makes few assumptions about the data and it treats all features equally, a desirable property when we compare different feature selection methods. In Figure 6 we compare the three UR methods with the same complexity: no correction, simple imputation and our correction (for a fair comparison we used perfect prior knowledge for the last two approaches). Our suggested approach of selecting the features using the corrected estimator (Section 4) outperforms the other approaches in most of the settings and achieves similar performance as using the ideal (unobservable) estimator.

6. The data is available in UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Splice and Connect-4 are 3-class classification problems, while Chess is 2-class. Categorical attributes are expanded into several binary attributes.

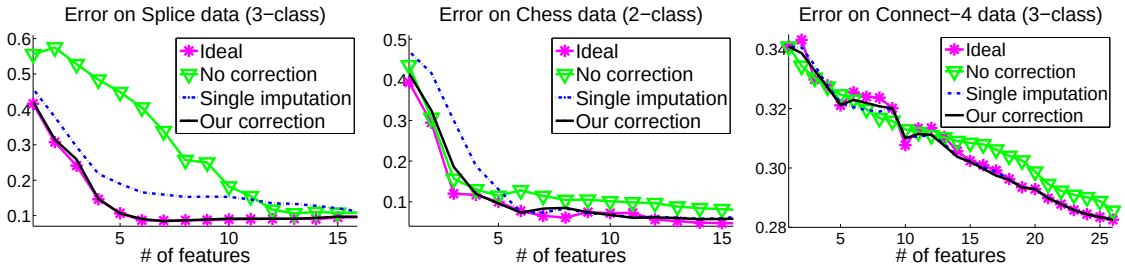


Figure 6: Average testing error over 10 random splits of the data into 50% training and 50% testing. For each training dataset we generate 10 UR datasets, by randomly non-differential under-reporting ten of the features with sensitivities chosen in the range $[0.25 - 0.75]$. For each under-reported method (no correction, simple imputation and our correction) we selected the most frequent features across the UR datasets.

8. Conclusions and Future Work

In this work we showed how to estimate mutual information in under-reporting scenarios. Initially, we presented how we can use the tool of missingness graphs to provide graphical representations of the different under-reported scenarios. Then, by connecting under-reporting with the problem of learning from positive and unlabelled data, we derived ways for estimating mutual information quantities by incorporating simple prior knowledge. Our theoretical results are supported through experiments with synthetic data. Finally, we showed how we can use our findings in a real-world health care application, ranking the risk factors that may lead to adverse birth outcome, and in a machine learning application, selecting important features when training/testing distributions differ.

Our future work is two-fold: theoretical extensions and empirical evaluations. Firstly, we will extend our theoretical results to derive informed ways for *testing independence* in under-reported scenarios, by controlling the two probabilities of errors: false positives and false negatives. Secondly, we will extend our findings to explore testing and estimation in *differential* under-reporting (i.e. when there is a direct arc between the missingness mechanism M_X and the variable Y in Figure 1a). Furthermore, it will give more insight to explore more applications and different simulation studies. For example, we can explore how robust to misspecified prior knowledge are the results of Section 7, or how we can select features that take into account both the relevancy and the redundancy.

We believe these results are highly applicable in a wide variety of machine learning applications, when we face the problem of under-reporting. Estimating mutual information, testing independence, ranking sets of features according to their relevancy/rendundancy, learning Bayesian network structures and sample size determination for experimental design are some –but not all– of the possible applications.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council, through a Centre for Doctoral Training [EP/I028099/1] and the Anyscale Apps project grant [EP/L000725/1].
Data access statement: All research data supporting this publication are directly available within this publication.

References

A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 3rd edition, 2013.

P. D. Allison. *Missing Data*. SAGE Publications, Inc., 2001.

D. R. Brillinger. Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics*, 18(6):163–183, 2004.

G. Brown, A. Pocock, M. Zhao, and M. Lujan. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *JMLR*, 13:27–66, 2012.

H. Chu, Z. Wang, S. R. Cole, and S. Greenland. Sensitivity analysis of misclassification: a graphical and a Bayesian approach. *Annals of Epidemiology*, 16(11):834–841, 2006.

F. Denis, A. Laurent, R. Gilleron, and M. Tommasi. Text classification and co-training from PU examples. In *ICML Workshop: The Continuum from Labeled to Unlabeled Data*, 2003.

P. Dietz, D. Homa, L. England, K. Burley, V. Tong, S. Dube, and J. Bernert. Estimates of Nondisclosure of Cigarette Smoking Among Pregnant and Nonpregnant Women of Reproductive Age in the US. *American Journal of Epidemiology*, 173(3):355–359, 2011.

J. Edwards, S. Cole, M. Troester, and D. Richardson. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *American Journal of Epidemiology*, 177(9):904–912, 2013.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *14th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, 2008.

S. Greenland. Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, 7(7):745–757, 1988.

S. Greenland. Sensitivity analysis and bias analysis. *Handb. of Epidemiology Ch. 19*, 2014.

K. Mohan, J. Pearl, and J. Tian. Graphical Models for Inference with Missing Data. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 1277–1285, 2013.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

D. Rahardja and D. M. Young. Confidence Intervals for the Risk Ratio Using Double Sampling with Misclassified Binomial Data. *Journal of Data Science*, 9(4):529–548, 2011.

D. B. Rubin. *Multiple imputation for nonresponse in surveys*. J. Wiley & Sons, 2004.

K. Sechidis, B. Calvo, and G. Brown. Statistical hypothesis testing in positive unlabelled data. In *ECML/PKDD*, pages 66–81. Springer Berlin Heidelberg, 2014.

I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

J. Wright, N. Small, P. Raynor, and et al. Cohort profile: The Born in Bradford multi-ethnic family cohort study. *International Journal of Epidemiology*, 42(4):978–991, 2013.