Estimating mutual information in under-reported variables: Supplementary Material

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)

In this document we provide the proofs of Theorems 1 and 2 of Sechidis et al. (2016).

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:*

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

Proof Before giving the proof, we will formally introduce the delta method (Agresti, 2013, Section 16.1.4).

Lemma 1 (Delta method)
Suppose that cell counts $\boldsymbol{n} = \{n_{x,y}\}$ *have a multinomial distribution with cell probabilities* $\boldsymbol{p} = \{p(x,y)\}, \forall\ x \in \mathcal{X},\ y \in \mathcal{Y}$. *Let* $N = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} n_{x,y}$, *and let* $\hat{\boldsymbol{p}}$ *denote the sample proportions:* $\hat{p}(x,y) = n_{x,y}/N$. *Let* $g(\boldsymbol{p}) \in \mathbb{R}$ *be a differentiable function, and let* $\phi_{x,y} = \frac{\partial g}{\partial p(x,y)}(\boldsymbol{p})$, $\forall\ x \in \mathcal{X},\ y \in \mathcal{Y}$. *Assume that at least one* $\phi_{x,y}$ *is nonzero and then the distribution* $\sqrt{N}\left[g(\hat{\boldsymbol{p}}) - g(\boldsymbol{p})\right]$ *converges to the normal distribution* $\mathcal{N}\left(0, \sigma^2\right)$ *when* $N \to \infty$, *where* $\sigma^2 = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\phi^2_{x,y} - \left(\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\phi_{x,y}\right)^2$.

In order to compute the partial derivatives $\frac{\partial g(\boldsymbol{p})}{\partial p(x,y)} = \frac{\partial I(X;Y)}{\partial p(x,y)}, \forall\ x \in \mathcal{X},\ y \in \mathcal{Y}$ we first need to calculate the following partial derivatives:

- $\frac{\partial p(x',y')}{\partial p(x,y)} = \delta_{xx'}\delta_{yy'}$, where $\delta_{xx'}\delta_{yy'}$ is the Kronecker delta: $\delta_{xx'}\delta_{yy'} = 1$ if $x = x'$ and $y = y'$, otherwise it is 0.

- $\frac{\partial p(x')}{\partial p(x,y)} = \delta_{xx'}$, where $\delta_{xx'}$ is the Kronecker delta: $\delta_{xx'} = 1$ if $x = x'$, otherwise it is 0.
- $\frac{\partial p(y')}{\partial p(x,y)} = \delta_{yy'}$
- $\frac{\partial (p(x')p(y'))}{\partial p(x,y)} = \delta_{xx'} p(y') + \delta_{yy'} p(x')$
- $\frac{\partial \ln \frac{p(x',y')}{p(x')p(y')}}{\partial p(x,y)} = \frac{1}{p(x',y')} \frac{\frac{\partial p(x',y')}{\partial p(x,y)} p(x')p(y') - p(x',y') \frac{\partial (p(x')p(y'))}{\partial p(x,y)}}{p(x')p(y')}$
$= \frac{1}{p(x',y')} \left(\frac{\partial p(x',y')}{\partial p(x,y)} - \frac{p(x',y')}{p(x')p(y')} \frac{\partial (p(x')p(y'))}{\partial p(x,y)} \right) =$
$= \frac{1}{p(x',y')} \left(\delta_{xx'}\delta_{yy'} - \frac{p(x',y')}{p(x')p(y')} \left(\delta_{xx'} p(y') + \delta_{yy'} p(x') \right) \right) =$
$= \frac{1}{p(x',y')} \left(\delta_{xx'}\delta_{yy'} - \frac{p(x,y')}{p(x)} - \frac{p(x',y)}{p(y)} \right).$

We are now ready to compute the partial derivatives of the mutual information:

$$
\begin{aligned}
\phi_{x,y} &= \frac{\partial g(\boldsymbol{p})}{\partial p(x,y)} = \frac{\partial I(X;Y)}{\partial p(x,y)} = \frac{\partial \sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} p(x',y') \ln \frac{p(x',y')}{p(x')p(y')}}{\partial p(x,y)} = \\
&= \sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} \left(\frac{\partial p(x',y')}{\partial p(x,y)} \ln \frac{p(x',y')}{p(x')p(y')} + p(x',y') \frac{\partial \ln \frac{p(x',y')}{p(x')p(y')}}{\partial p(x,y)} \right) = \\
&= \sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} \left(\delta_{xx'}\delta_{yy'} \ln \frac{p(x',y')}{p(x')p(y')} + \delta_{xx'}\delta_{yy'} - \frac{p(x,y')}{p(x)} - \frac{p(x',y)}{p(y)} \right) = \\
&= \ln \frac{p(x,y)}{p(x)p(y)} + 1 - \sum_{y' \in \mathcal{Y}} \frac{p(x,y')}{p(x)} - \sum_{x' \in \mathcal{X}} \frac{p(x',y)}{p(y)} = \\
&= \ln \frac{p(x,y)}{p(x)p(y)} + 1 - \frac{p(x)}{p(x)} - \frac{p(y)}{p(y)} = \ln \frac{p(x,y)}{p(x)p(y)} - 1.
\end{aligned}
$$

Thus the sample variance of the MLE of the mutual information becomes

$$
\begin{aligned}
\sigma^2_{MI} &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \phi^2_{x,y} - \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \phi_{x,y} \right)^2 \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} - 1 \right)^2 - \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} - 1 \right) \right)^2 \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \left(p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} \right)^2 + p(x,y) - 2p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \right) \\
&\quad - \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} - 1 \right)^2 \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} \right)^2 + 1 - 2 \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$

$$
- \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \right)^2 - 1 + 2 \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}
$$

$$
= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} \right)^2 - \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \right)^2
$$

$$
= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} \right)^2 - I(X;Y)^2.
$$

■

Theorem 2 (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:

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

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

Proof To derive the asymptotic distribution of $\widehat{I}_\gamma(\widetilde{X};Y)$ we will use the delta method presented in Theorem 1. The expression of the corrected estimator is:

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

Since in the expression of $\widehat{I}_\gamma(\widetilde{X};Y)$ we have the maximum likelihood estimates for the probabilities $p(y), p(y|\widetilde{x}=1)$ the first step is to calculate the partial derivatives of these quantities with respect to the parameters of this model $p(y,\widetilde{x}=1)$ and $p(y,\widetilde{x}=0)$:

$$
\frac{\partial p(y')}{\partial p(y,\widetilde{x}=1)} = \delta_{yy'}, \qquad \frac{\partial p(y')}{\partial p(y,\widetilde{x}=0)} = \delta_{yy'},
$$

$$
\frac{\partial p(y'|\widetilde{x}=1)}{\partial p(y,\widetilde{x}=1)} = \frac{\delta yy' - p(y'|\widetilde{x}=1)}{p(\widetilde{x}=1)}, \qquad \frac{\partial p(y'|\widetilde{x}=1)}{\partial p(y,\widetilde{x}=0)} = 0,
$$

where $\delta_{yy'}$ is the Kronecker delta, which takes the value of 1 if $y = y'$ and 0 otherwise. By using the above partial derivatives we have the following results:

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

$$
\phi_{y,\widetilde{x}=0} = \frac{\partial I_\gamma(\widetilde{X};Y)}{\partial p(y,\widetilde{x}=0)} = \ln \frac{p(y) - p(y|\widetilde{x}=1)\gamma}{p(y)}.
$$

So by using delta method the asymptotic variance of the estimator equals

$$\sigma^2_{MI_\gamma} = \sum_{y\in\mathcal{Y}} \Big(p(y,\widetilde{x}=1)\phi^2_{y,\widetilde{x}=1} + p(y,\widetilde{x}=0)\phi^2_{y,\widetilde{x}=0}\Big) - \left(\sum_{y\in\mathcal{Y}} \big(p(y,\widetilde{x}=1)\phi_{y,\widetilde{x}=1} + p(y,\widetilde{x}=0)\phi_{y,\widetilde{x}=0}\big)\right)^2,$$

where $\phi_{y,\widetilde{x}=1}$ and $\phi_{y,\widetilde{x}=0}$ are calculated earlier and are functions of γ. Furthermore, when $\gamma = p(x=1)$ it holds that $I_\gamma(\widetilde{X};Y) = I(X;Y)$ and so the estimator $\widehat{I}_\gamma(\widetilde{X};Y)$ is asymptotically normally distributed around $I(X;Y)$. ■

References

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

K. Sechidis, M. Sperrin, E. Petherick, and G. Brown. Estimating mutual information in under-reported variables. In *Eighth International Conference on Probabilistic Graphical Models (PGM) 2016*, volume 52, pages 449–461. Journal of Machine Learning Research (JMLR): Workshop and Conference Proceedings, 2016.