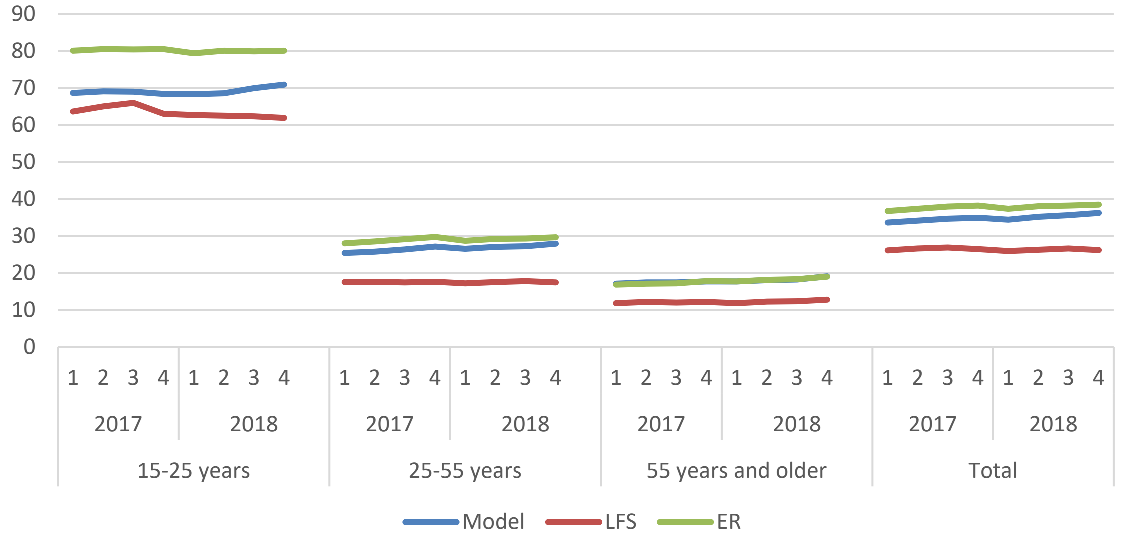
Using Hidden Markov and macro-integration models for combining data from different sources

Nino Mushkudiani, Jeroen Pannekoek and Sander Scholtus
(Statistics Netherlands)

UNECE Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna

Motivating example

Proportion of flexible contracts among employees in the Netherlands



(LFS = Labour Force Survey; ER = Employment Register)

Motivating example

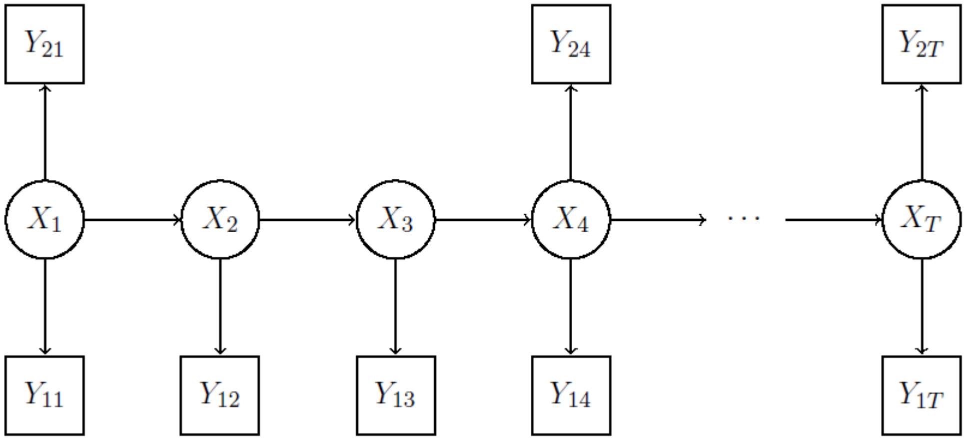
Two approaches:

- Hidden Markov Model
  (Pavlopoulos & Vermunt, 2015; Bakker et al., 2021)
- Macro-integration
  (Mushkudiani & Pannekoek, 2019)

Hidden Markov Model

Assumptions:

- Markov property
- Measurement errors described by probabilities $\pi_{y|x}^{Y_{lt}|X_t} = \mathrm{P}(Y_{lt} = y | X_t = x)$ independent between units and between Y_{1t} and Y_{2t}



Parameters of interest here:

- Marginal probabilities $\pi_x^{X_t} = \mathrm{P}(X_t = x)$
- Transition probabilities $\pi_{x_t|x_{t-3}}^{X_t|X_{t-3}} = \mathrm{P}(X_t = x_t | X_{t-3} = x_{t-3})$

Macro-integration

Mushkudiani & Pannekoek (2019): integration problem on aggregated data

information from source 1

$t-3$	$t-2$	$t-1$	t	proportion
A	A	A	A	p_{AAAA}
A	A	B	A	p_{AABA}
A	A	C	A	p_{AACA}
A	B	A	A	p_{ABAA}
A	B	B	A	p_{ABBA}
A	B	C	A	p_{ABCA}
A	C	A	A	p_{ACAA}
A	C	B	A	p_{ACBA}
A	C	C	A	p_{ACCA}
…	…	…	…	…
C	C	C	C	p_{CCCC}

information from source 2

$t-3$	t	proportion
A	A	p_{AA}
A	B	p_{AB}
A	C	p_{AC}
B	A	p_{BA}
B	B	p_{BB}
B	C	p_{BC}
C	A	p_{CA}
C	B	p_{CB}
C	C	p_{CC}

Comparison

Both approaches applied to the same data (ER and LFS, 2009 data):

- HMM results taken from Pankowska et al. (2018)
- MI results taken from Mushkudiani & Pannekoek (2019)

Marginal distribution of contract type:

	Permanent	Flexible	Other
ER	58.5%	15.1%	26.4%
LFS	65.3%	11.0%	23.7%
MI	62.5%	12.9%	24.6%
HMM	61.1%	12.8%	26.1%

Three-month transition probability:

	Flexible -> Permanent
ER	7.3%
LFS	5.8%
MI	6.2%
HMM	1.7%

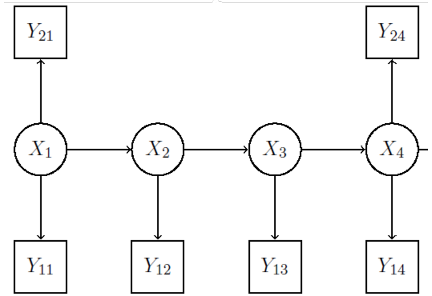
This study

To explain differences:

- Theoretical analysis of the simplest scenario ($T = 4$)
- Simulation study on more complicated scenarios

This presentation: only the first part

The case *T* = 4



Hidden Markov Model: Minimize

$$\mathcal{D}_{\text{HMM}} = -n \sum_{y_{11}=1}^{K} \sum_{y_{12}=1}^{K} \sum_{y_{13}=1}^{K} \sum_{y_{14}=1}^{K} \sum_{y_{21}=1}^{K} \sum_{y_{24}=1}^{K} p_{y_{11},y_{12},y_{13},y_{14},y_{21},y_{24}}^{Y_{11},Y_{12},Y_{13},Y_{14},Y_{21},Y_{24}} \times$$
$$\log\left(\sum_{x_1=1}^{K}\sum_{x_2=1}^{K}\sum_{x_3=1}^{K}\sum_{x_4=1}^{K} \pi_{x_1}^{X_1} \pi_{x_2|x_1}^{X_2|X_1} \pi_{x_3|x_2}^{X_3|X_2} \pi_{x_4|x_3}^{X_4|X_3} \pi_{y_{11}|x_1}^{Y_{11}|X_1} \pi_{y_{12}|x_2}^{Y_{12}|X_2} \pi_{y_{13}|x_3}^{Y_{13}|X_3} \pi_{y_{14}|x_4}^{Y_{14}|X_4} \pi_{y_{21}|x_1}^{Y_{21}|X_1} \pi_{y_{24}|x_4}^{Y_{24}|X_4}\right)$$

Macro-integration (using Kullback-Leibler divergence): Minimize

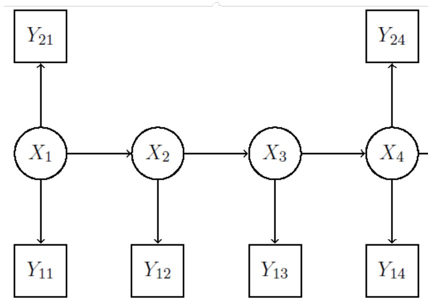
$$\mathcal{D}_{\text{KL}} = \sum_{x_1=1}^{K}\sum_{x_2=1}^{K}\sum_{x_3=1}^{K}\sum_{x_4=1}^{K} \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4}\left(\log \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4} - \log p_{x_1,x_2,x_3,x_4}^{Y_{11},Y_{12},Y_{13},Y_{14}} - 1\right)$$
$$+ \sum_{x_1=1}^{K}\sum_{x_4=1}^{K} \pi_{x_1,x_4}^{X_1,X_4}\left(\log \pi_{x_1,x_4}^{X_1,X_4} - \log p_{x_1,x_4}^{Y_{21},Y_{24}} - 1\right),$$

under restrictions

$$1 = \sum_{x_1=1}^{K}\sum_{x_2=1}^{K}\sum_{x_3=1}^{K}\sum_{x_4=1}^{K} \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4};$$

$$\pi_{x_1,x_4}^{X_1,X_4} = \sum_{x_2=1}^{K}\sum_{x_3=1}^{K} \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4}.$$

The case $T = 4$

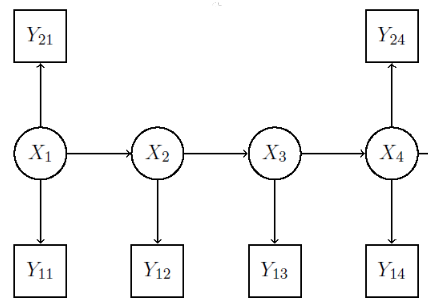


Three differences between HMM and MI:

1. HMM uses joint distribution of $(Y_{11}, Y_{12}, Y_{13}, Y_{14}, Y_{21}, Y_{22})$ so requires **linked microdata**.
   MI uses separate distributions of $(Y_{11}, Y_{12}, Y_{13}, Y_{14})$ and (Y_{21}, Y_{22}) so requires **no linked microdata**.
2. HMM contains an explicit **measurement error model**.
   MI does not, but does assume that
   $$E\left(p_{a,b}^{Y_{11},Y_{14}}\right) \approx E\left(p_{a,b}^{Y_{21},Y_{24}}\right) \approx \pi_{a,b}^{X_1,X_4}.$$
3. HMM uses the **Markov assumption**
   $$\pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4} = \pi_{x_1}^{X_1} \pi_{x_2|x_1}^{X_2|X_1} \pi_{x_3|x_2}^{X_3|X_2} \pi_{x_4|x_3}^{X_4|X_3}.$$
   MI does not impose any structure on $\pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4}$.

The case $T = 4$

Small simulation study inspired by LFS-ER application



Marginal distribution of contract type:

	Permanent	Flexible	Other
Source 1	49.0%	21.2%	29.8%
Source 2	53.5%	17.2%	29.4%
MI	51.3%	19.1%	29.6%
HMM	50.3%	20.0%	29.7%
HMM*	50.2%	20.0%	29.7%
HMM**	51.2%	19.2%	29.6%
HM**	51.2%	19.2%	29.6%

Three-month transition probability:

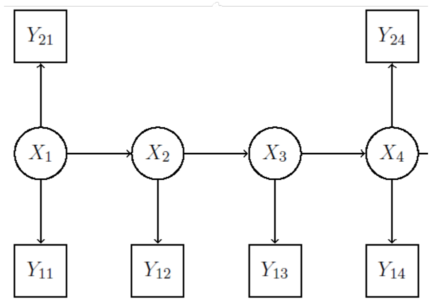
	Flexible -> Permanent
Source 1	17.5%
Source 2	20.2%
MI	18.9%
HMM	5.2%
HMM*	5.0%
HMM**	29.8%
HM**	18.7%

no linked microdata

no error model

no Markov property

The case $T = 4$



HM**: Minimize

$$\mathcal{D}_{\text{HM}}^{**} = -n \left\{ \sum_{x_1=1}^{K} \sum_{x_2=1}^{K} \sum_{x_3=1}^{K} \sum_{x_4=1}^{K} p_{x_1,x_2,x_3,x_4}^{Y_{11},Y_{12},Y_{13},Y_{14}} \log \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4} + \sum_{x_1=1}^{K} \sum_{x_4=1}^{K} p_{x_1,x_4}^{Y_{21},Y_{24}} \log \pi_{x_1,x_4}^{X_1,X_4} \right\}$$

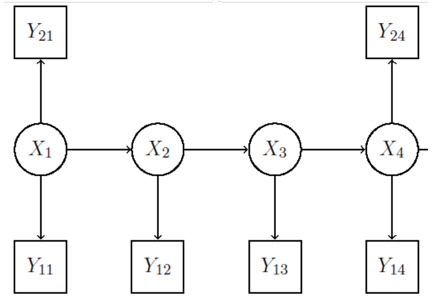
MI (using Kullback-Leibler divergence): Minimize

$$\mathcal{D}_{\text{KL}} = \sum_{x_1=1}^{K} \sum_{x_2=1}^{K} \sum_{x_3=1}^{K} \sum_{x_4=1}^{K} \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4} \left(\log \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4} - \log p_{x_1,x_2,x_3,x_4}^{Y_{11},Y_{12},Y_{13},Y_{14}} - 1\right)$$
$$+ \sum_{x_1=1}^{K} \sum_{x_4=1}^{K} \pi_{x_1,x_4}^{X_1,X_4} \left(\log \pi_{x_1,x_4}^{X_1,X_4} - \log p_{x_1,x_4}^{Y_{21},Y_{24}} - 1\right),$$

Both approaches use the same restrictions

$$1 = \sum_{x_1=1}^{K} \sum_{x_2=1}^{K} \sum_{x_3=1}^{K} \sum_{x_4=1}^{K} \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4};$$
$$\pi_{x_1,x_4}^{X_1,X_4} = \sum_{x_2=1}^{K} \sum_{x_3=1}^{K} \pi_{x_1,x_2,x_3,x_4}^{X_1,X_2,X_3,X_4}.$$

The case *T* = 4



Both problems have a closed-form solution

- HM**:

$$\pi_{x_1,x_4}^{X_1,X_4} = \frac{p_{x_1,x_4}^{Y_{11},Y_{14}} + p_{x_1,x_4}^{Y_{21},Y_{24}}}{2}$$

(arithmetic mean of two distributions)

- MI (using Kullback-Leibler divergence):

$$\pi_{x_1,x_4}^{X_1,X_4} = \kappa\sqrt{p_{x_1,x_4}^{Y_{11},Y_{14}} p_{x_1,x_4}^{Y_{21},Y_{24}}}$$

(proportional to geometric mean of two distributions)

To be continued…

- Work in progress: simulation study to compare HMM and MI in other, more complicated scenarios
- Other possible applications of HMM
  - Evaluate accuracy (bias, variance) of statistical output due to random measurement errors
  - Use HMM as input for a selective editing approach

References

B.F.M. Bakker, G. Gringhuis, J. Hoogland, F. van der Linden, J. Michiels, J. Pannekoek, S. Scholtus & W. Smits (2021), Tijdelijke en Ae contracten. Verschillen tussen de schattingen uit de Polisadministratie en de Enquête beroepsbevolking verklaard? (in Dutch). Discussion paper, Statistics Netherlands, The Hague / Heerlen.

N. Mushkudiani & J. Pannekoek (2019), Estimating a time series of temporary employment using a combination of survey and register data. Discussion paper, Statistics Netherlands, The Hague.

P. Pankowska, B. Bakker, D. Oberski & D. Pavlopoulos (2018), Reconciliation of inconsistent data sources by correction for measurement error: the feasibility of parameter re-use. *Statistical Journal of the IAOS* **34**, 317–329.

D. Pavlopoulos & J.K. Vermunt (2015), Measuring Bible employment. Do survey or register data tell the truth? *Survey Methodology* **41**, 197–214.