Adjusting misclassification using a second classifier with an external validation sample

Jonas F. Schenkel[1] and Li-Chun Zhang[2,3,1]

[1] *University of Oslo, Norway (email: jonasfsc@math.uio.no)*
[2] *University of Southampton, UK*
[3] *Statistics Norway, Norway*

Abstract

Administrative data may suffer from delays or mistakes in reporting. To adjust for the resulting measurement errors, it is often necessary to combine data from related sources, such as sample survey, administrative or 'big' data. However, the additional measure variable usually has a different definition and errors of its own, and the available joint data set may not have a completely known sampling distribution. We develop a modelling approach which capitalises on one's knowledge and experience with the data source where they exist, and apply it to register and survey based Employed status. Comparisons are made to adjustments by hidden Markov models. Our approach is applicable to similar situations involving big data sources.

Keywords: Misclassification, discriminant, calibration probability, matrix method, maximum likelihood estimation

1 Introduction

Making greater use of data originated from administrative sources for statistical purposes has become an increasingly important topic for many National Statistical Offices (NSO). Administrative data are not perfect, and combination of data from multiple sources is often needed to overcome various known deficiencies, such as when capture-recapture methods are applied to multiple registers to adjust for their combined under-coverage, or when latent class analysis is applied to adjust for the discrepancies among similarly defined variables from different sources. See Zhang (2012), di Zio et al. (2017) and Hand (2018) for broad discussions of the relevant statistical topics, methods and challenges.

In our motivating application later, we have a register-based Employed status, denoted by X, which is known for everyone in the population, based on an administrative source that delivers to Statistics Norway on a monthly basis. Due to delays and misreporting, X may be erroneous for the true register status, denoted by Y, for a considerable period after the reference month, although the two will eventually reach agreement over time.

A concurrent Employed status is also available from the continuous Labour Force Survey (LFS), denoted by Z, which follows a different definition of employment, so that Z and Y can differ regardless if Z is a true measure of its own definition. The setting is generically summarised in Table 1, where $R = 1$ if a unit is in the joint sample and 0 otherwise.

Table 1: Presence (✓) of two fallible classifiers

Classifier	Joint sample ($R = 1$)	Rest of population ($R = 0$)
X	✓	✓
Z	✓	—

We shall develop and apply a modelling approach for adjusting the misclassification of X by making use of the second classifier Z that is also subject to misclassification. We assume that, based on the knowledge and experience about the source of X, it is possible to define the part of the population where one is confident that X is either correct or nearly so, denoted by $B = 1$, and the errors of X are only much more likely in the rest population, denoted by $B = 2$. We call B a *discriminant*. As Table 2 shows, the discriminant creates a validation sample of (Z, Y) in the subpopulation of $B = 1$ so that $\Pr(Z|Y, B = 1)$ can be estimated. Our model allows one to apply the estimated $\Pr(Z|Y, B = 1)$ to the subpopulation of $B = 2$, in order to estimate $\Pr(Y|X, B = 2)$ and then $\Pr(Y|B = 2)$.

Table 2: Presence (✓) of two fallible classifiers given discriminant B

	Joint sample ($R = 1$)		Rest of population ($R = 0$)	
	$X = Y$	$\Pr(X \neq Y) > 0$	$X = Y$	$\Pr(X \neq Y) > 0$
Classifier	$B = 1$	$B = 2$	$B = 1$	$B = 2$
X	✓	✓	✓	✓
Z	✓	✓	—	—

The discriminant provides a previously unexplored possibility to make use of one's knowledge and experience of the source. This is often possible with administrative data. For the aforementioned register Employed status X, many indicators in the administrative sources can be used for the construction of the discriminant, such as the type of job or position, the length of employment history, the level of income, and so on. Register-based highest level of education is another example, where e.g. native borns with a completed education history in the relevant registers can be reliably classified. As a third example, registered address of a person (or family) may be mistaken due to lack of updating or misreporting. Provided it is possible to obtain various 'signs of life' addresses via licence registration, social or health services, utility bills, etc. one may be able to distinguish which registered addresses are most likely to be correct from the rest. As a final example here, dwelling addresses registered for primary residence or recreation uses may be mistaken or not-in-use in reality. Provided it is possible to obtain relevant activity data of electricity smart meter, mobile phone signal or airborne laser scanning, one may be able to identify those most likely to have correctly classified uses.

In all such or similar situations, the discriminant creates a validation sample for a second fallible classifier Z, which is external to the subpopulation where adjustment of X is needed. As will be demonstrated later, in practice, X only needs to be nearly always correct for the tacit assumption of validation sample to lead to useful adjustments for the rest population, even if the assumption is not completely true. The remaining task of modelling is to enable one to usefully apply $\Pr(Z|Y, B = 1)$ estimated in the validation sample.

There exists a large body of literature on categorical data analysis in the presence of misclassification. See the excellent reviews of Kuha and Skinner (1997) and Kuha et al. (1998), who note in particular a strong tradition in medical studies. Bross (1954) shows how conclusions drawn from 2×2 tables can be affected by misclassification. Tenenbein (1970) introduces the double sampling methodology for binary classifiers. It is assumed that a (simple) random sample is classified by a cheap, but fallible classifier. A subsample is taken, from which a more costly *true* classifier is obtained. The subsample, from which we can learn the misclassification mechanism is a validation sample. It is shown that making use of the fallible classifier observed outside the validation sample is more efficient than using only the true classifier in the validation sample.

The basic double sampling method can be extended to allow for multinomial variables (Tenenbein, 1971, 1972). Hochberg (1977) consider hypothesis testing for multidimensional jointly observed data. Hochbeg and Tenenbein (1983) and Chen et al. (1984) extend double sampling to triple sampling, where only the true classifier is observed in one sample, only the fallible in another and both jointly in a third sample. See also Chen (1989) for a review of the related methods. Swensen (1988) considers the setting, where register-based measure variables are fallible and survey variables are true. Haitovsky and Rapp (1992) study efficient sampling design of the validation sample beyond simple random sampling.

Chen (1979) introduces the framework of log-linear models to the double sampling methodology, where one specifies a log-linear model of the misclassification probability matrix, as well as another log-linear model of the true classifications. The log-linear model framework facilitates maximum likelihood estimation (MLE) using the EM-algorithm (Dempster et al., 1977). See e.g. Chen (1979, 1989) and Espeland and Odoroff (1985) for applications of the EM algorithm to misclassification problems.

For situations involving two fallible classifiers without a true classifier, identifiability of model parameters requires additional assumptions. Hui and Walter (1980) assume the misclassification mechanisms of two diagnostic tests are the *same* for all the units in the population. A partition of the population is introduced, where the case prevalence varies across the subpopulations and the number of subpopulations is such that there is enough degrees of freedom to allow for parameter identification. Lie et al. (1994) study binary variables from two different health registers, both of which are subject to misclassification errors, under the assumption that the positive cases missed by either classifier will *all* be correctly classified by the other. Qiu et al. (2018) propose two models for confidence interval procedures of the population proportion. Under both the models, they assume there are no false positives for *both* classifiers.

Meanwhile, multiple fallible classifiers have been studied in situations involving repeated

measures. For instance, for estimating gross labour flows, misclassification models of reinterview data have been considered by Abowd and Zellner (1985), Poterba and Summers (1986), Chua and Fuller (1987) and Singh and Rao (1995). Zhang (2005) proposes a special sparse misclassification model, which does not require reinterviews.

There is a huge body of literature on latent class analysis or structural equation models of multiple fallible classifiers, provided there are enough degrees of freedom in the observed data, which usually requires a longitudinal setting. Hidden Markov models (HMMs) are often applied for misclassification adjustment (e.g. Van de Pol and Langeheine, 1997; Biemer and Bushery, 2000; Vermunt et al., 2008; Magidson et al., 2009; Vermunt, 2010). See e.g. Yoon (2009) for a review for biological sequence analysis. In particular, Pavlopoulos and Vermunt (2015) apply an extended HMM to survey and administrative register data on temporary employment. We will apply the HMM in an off-the-shelf manner to provide a comparison to the method developed in this paper.

The rest of the paper is organised as follows. The model we propose is developed in Section 2, together with the estimation methods. The HMM model for comparison is also briefly described. The application is presented in Section 3. Finally, Section 4 contains a summary and an outline of some topics for future research.

2 Models

Denote by $U = \{1, ..., N\}$ the population that the variables (X, Y) are associated with, denoted by (X_i, Y_i) for $i \in U$. Let X and Y both take values $1, ..., K$. Under the setting of Table 1, we observe two fallible classifiers X and Z (of the true Y) jointly in a sample s, but only X outside s. Let R be the binary observation indicator, where $R_i = 1$ if $i \in s$ and 0 if $i \in U \setminus s$. The selection mechanism of R may be unknown generally.

To focus on the central issues, we assume $\{X_i : i \in U\}$ to be known for the whole population, which is the case in our application later. But the modelling approach developed below is also applicable in the case of double-sampling, where the joint sample s is a subset of a larger probability sample of X from U.

2.1 Modelling given discriminant

We introduce our model for the setting of Table 2 in two steps. First, we introduce the discriminant B and the simplest assumptions of Z and R, so that $\Pr(Y|X, B = 2)$ is identifiable in the joint subsample where $(B, R) = (2, 1)$, and $\Pr(Y|B = 2)$ can be estimated. This leads to two simple models, which contain some strong assumptions of the sample observation and misclassification mechanisms. Next, additional covariates are introduced to relax these assumptions, yielding the model that is more generally applicable.

For the classifier X, we assume there exists a known binary discriminant, denoted by

$B = 1$ or 2, such that we have, in the population,

$$\Pr(Y = X \mid B = 1) = 1 \tag{1a}$$

$$\Pr(Y = y \mid X = x, B = 2) = \eta_{yx} \, . \tag{1b}$$

The idea is simple. Given that $X = Y$ conditional on $B = 1$, the joint subsample of (Z, X) is a validation sample of (Z, Y) where $(B, R) = (1, 1)$. Provided suitable assumptions of R, so that $\Pr(Z|Y)$ is transportable from those with $B = 1$ to the others with $B = 2$, one would be able to disentangle the conditional distribution $\Pr(Y|X, B = 2)$ from the joint distribution $\Pr(Z, X|B = 2)$ in the subsample where $(B, R) = (2, 1)$.

Figure 1: Independence graphs of model M_0 (left) and model M_B (right): true variable Y, fallible classifiers (X, Z), joint sample observation indicator R, discriminant B.

Figure 1 gives the independence graphs (Edwards, 2012) of two models of (Y, X, Z, B, R), where two groups of variables are independent of each other if they are unconnected in the graph, and two (groups of) variables are conditionally independent given the variables that separate them in the graph. In the terminology of Rubin (1976), R is missing completely at random (MCAR) under the first model M_0, and it is missing at random (MAR) given (X, B) under the second model M_B. Under either model, Z is independent of (X, B, R) conditional on Y, denoted by

$$\Pr(Z = z \mid Y = y, R = 1, B, X) = \Pr(Z = z \mid Y = y) \equiv \lambda_{zy} \, . \tag{2}$$

In the terminology of Kuha and Skinner (1997), misclassification by Z is *nondifferential* with respect to (X, B, R) under (2).

Moreover, under either model, (Z, Y) are conditionally independent of R given (X, B), i.e. $\Pr(Z, Y \mid X, B, R = 1) = \Pr(Z, Y \mid X, B)$, such that we obtain

$$\psi_{zx} \equiv \Pr(Z = z \mid X = x, B = 2, R = 1) = \sum_{y=1}^{K} \lambda_{zy} \eta_{yx} \tag{3}$$

by integrating out Y from $\Pr(Z, Y \mid X, B)$. The misclassification probabilities λ_{zy} are said to be transportable (Kuha and Skinner, 1997) from $B = 1$ to $B = 2$ by virtue of (2). The probabilities $\Pr(Y|X, B = 2)$ are referred to as the *calibration probabilities*, denoted by

$$\eta_{yx} \equiv \Pr(Y = y \mid X = x, B = 2) \, .$$

The conditional probabilities $\Pr(Z|X, B, R = 1)$ by (2) or (3) for the two subsamples given $B = 1$ or 2 are summarised in Table 3. Neither of the two classifiers is necessarily more accurate than the other in the subpopulation of $B = 2$.

Table 3: Subsample conditional probabilities $\Pr(Z \mid X, B, R = 1)$.

	$B = 1$					$B = 2$				
Z	$X = 1$	$\dots$	$X = x$	$\dots$	$X = K$	$X = 1$	$\dots$	$X = x$	$\dots$	$X = K$
1	λ_{11}	$\dots$	λ_{1x}	$\dots$	λ_{1K}	ψ_{11}	$\dots$	ψ_{1x}	$\dots$	ψ_{1K}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
z	λ_{z1}	$\dots$	λ_{zx}	$\dots$	λ_{zK}	ψ_{z1}	$\dots$	$\sum_{y=1}^{K} \lambda_{zy}\eta_{yx}$	$\dots$	ψ_{zK}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
K	λ_{K1}	$\dots$	λ_{Kx}	$\dots$	λ_{KK}	ψ_{K1}	$\dots$	ψ_{Kx}	$\dots$	ψ_{KK}

MCAR for R is a strong assumption that may be unrealistic in many applications. When it comes to MAR for R in Figure 1, allowing B in addition to X is unlikely to be a useful relaxation of using only X to control for R, since the discriminant B is defined with respect to misclassification by X. Whereas the fact that X is a known 'proxy' of Y is usually favourable to the MAR assumption. In the extreme case, if $X = Y$, then R must be independent of Y given X. Or, heuristically speaking, whatever the effect Y has on R, it would be largely controlled for given X if X contains much information about Y. Still, a reasonable approach is to introduce additional covariates, as it is common in the literature of modelling survey nonresponse or nonprobability sample selection in the absence of misclassification, not least because this would also enable one to relax the assumptions that $\Pr(Z|Y)$ is the same for everyone in the population and $\Pr(Y|X)$ is the same for everyone in the subpopulation of $B = 2$.

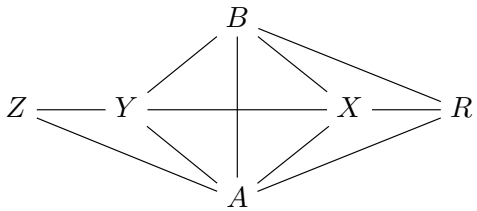

Figure 2: Independence graph of model M_{AB}: true variable Y, fallible classifiers (X, Z), joint sample observation indicator R, discriminant B, other covariates A.

First, for the population calibration probabilities $\Pr(Y|X)$, we modify the discriminant assumption (1) to include the additional known covariates, denoted by A,

$$\Pr(Y = X \mid B = 1, A) = 1 \tag{4a}$$

$$\Pr(Y = y \mid X = x, B = 2, A = a) = \eta_{yx|a}\,. \tag{4b}$$

Next, denote by M_{AB} the model whose independence graph is given in Figure 2, where we allow A to be connected to all the other variables in the graph. Under M_{AB}, misclassification

by Z is nondifferential with respect to (X, B, R) conditional on A, i.e.

$$\Pr(Z = z \mid Y = y, A = a, X, B, R) = \Pr(Z = z \mid Y = y, A = a) \equiv \lambda_{zy|a}\,. \tag{5}$$

Note that (5) is similar to (2), albeit with conditioning on A in addition. Moreover, similarly as for ψ_{zx} by (3) under M_B, (Z, Y) are conditionally independent of R given (X, A, B), i.e. $\Pr(Z, Y \mid X, A, B, R = 1) = \Pr(Z, Y \mid X, A, B)$, such that

$$\psi_{zx|a} \equiv \Pr(Z = z \mid X = x, A = a, B = 2, R = 1) = \sum_{y=1}^{Y} \lambda_{zy|a}\eta_{yx|a}\,. \tag{6}$$

The model M_{AB} defined by (4), (5) and (6) thus encompasses the model M_B defined by (1), (2) and (3). Of course, these assumptions of M_{AB} may still not hold completely in applications. The sensitivity of the resulting estimator of the target proportions will be investigated later in the application as well as by a simulation study.

2.2 Estimation

Provided the sample size accommodates it, one may let A be a population stratification variable based on the relevant covariates. A so-called matrix method (Kuha and Skinner, 1997) follows immediately. For $A = a$, let $\boldsymbol{\Lambda}_a$ be the matrix of probabilities $\lambda_{zy|a}$ and $\boldsymbol{\Psi}_a$ that of $\psi_{zx|a}$ and $\boldsymbol{H}_a$ that of $\eta_{yx|a}$. Given $(B, R) = (2, 1)$, we have

$$\boldsymbol{\Psi}_a = \boldsymbol{\Lambda}_a \boldsymbol{H}_a$$

under the model M_{AB}. Provided the inverse matrix exists, an estimator of $\boldsymbol{H}_a$ is given by

$$\hat{\boldsymbol{H}}_a = \hat{\boldsymbol{\Lambda}}_a^{-1} \hat{\boldsymbol{\Phi}}_a\,,$$

where $\hat{\boldsymbol{\Lambda}}_a$ is estimated from the subsample of $(Z, X) = (Z, Y)$ given $(A, B) = (a, 1)$, and $\hat{\boldsymbol{\Psi}}_a$ from the subsample of (Z, X) given $(A, B) = (a, 2)$.

Next, let $\boldsymbol{\mu}_{X|ab}$ be the vector of subpopulation proportions of X given $(A, B) = (a, b)$. Similarly for $\boldsymbol{\mu}_{Y|ab}$, where $\boldsymbol{\mu}_{Y|a1} = \boldsymbol{\mu}_{X|a1}$ by (4a), and an estimator of $\boldsymbol{\mu}_{Y|a2}$ is given by

$$\hat{\boldsymbol{\mu}}_{Y|a2} = \hat{\boldsymbol{H}}_a\, \boldsymbol{\mu}_{X|a2}\,.$$

The estimator $\hat{\boldsymbol{\mu}}_{Y|a2}$ is easily consistent, as all the stratum sample sizes tend to infinity asymptotically. An estimator of the overall proportions is then given by

$$\hat{\boldsymbol{\mu}}_Y = \frac{1}{N} \sum_a (N_{a1} \boldsymbol{\mu}_{X|a1} + N_{a2} \hat{\boldsymbol{\mu}}_{Y|a2})$$

where N_{ab} is the stratum subpopulation size with $(A, B) = (a, b)$.

We adopt model-based inference in this paper, where the possibly complex sampling design of s is ignorable conditional on the model covariates. Under the model M_{AB}, we

treat the realised stratum sample sizes $n_{x|ab}$ as fixed, which is the number of units with $(X, A, B, R) = (x, a, b, 1)$, for which we treat the associated Z as random given $B = 1$ and the associated (Z, Y) as random given $B = 2$. This is justifiable, because R is conditionally independent of (Y, Z) given (X, A, B) and X is known for the population. From each subsample $(x, a, b, 1)$, we draw $n_{x|ab}$ units with the associated Z randomly and with replacement, to obtain a corresponding bootstrap replicate subsample. Repeating this separately for each combination of (x, a, b) yields then an entire bootstrap replicate sample, based on which we obtain a corresponding bootstrap replicate estimate of $\boldsymbol{\mu}_{Y|+2}$ for the subpopulation with $B = 2$. The bootstrap variance estimator of $\hat{\boldsymbol{\mu}}_{Y|+2}$ can be obtained based on a sufficient number of repetitions of the procedure.

One can also consider MLE, using the stratum-specific misclassification probabilities $\boldsymbol{\Lambda}_a$ given $A = a$ and $\boldsymbol{\Theta}_a$ of $\Pr(X|Y, B = 2)$ given $A = a$. The stratum likelihood is

$$L_a(\boldsymbol{\Lambda}_a, \boldsymbol{\Theta}_a, \boldsymbol{\mu}_{Y|a2}) \propto \Big[\prod_{x=1}^{K} \prod_{z=1}^{K} \lambda_{zx|a}^{n_{zx|a1}} \Big] \Big[\prod_{x=1}^{K} \prod_{z=1}^{K} \big(\sum_{y=1}^{K} \theta_{xy|a} \lambda_{zy|a} \mu_{y|a2} \big)^{n_{zx|a2}} \Big] \Big[\prod_{x}^{K} \big(\sum_{y=1}^{K} \theta_{xy|a} \mu_{y|a2} \big)^{m_{x|a2}} \Big],$$

where $n_{zx|ab}$ is the sample cell count given $(A, B) = (a, b)$, and $m_{x|a2}$ is the corresponding out-of-sample total of $X = x$. As shown in Appendix A, the MLE of $\mu_{y|a2}$ is the same as the matrix method estimator above. However, the MLE is also applicable given more parsimonious specifications of $\lambda_{zy|a}$ and $\theta_{xy|a}$ in terms of the covariates A when A is not a stratification variable, say, $\boldsymbol{\Lambda}_a = \boldsymbol{\Lambda}(a, \boldsymbol{\lambda})$ and $\boldsymbol{\Theta}_a = \boldsymbol{\Theta}(a, \boldsymbol{\theta})$ with respective parameter vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$. The likelihood is then given by

$$L(\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\mu}_{Y|a2}) = \prod_{a} L_a(\boldsymbol{\Lambda}_a, \boldsymbol{\Theta}_a, \boldsymbol{\mu}_{Y|a2}) \,.$$

2.3 Estimation using hidden Markov models

Here we briefly outline the HMM considered by Pavlopoulos and Vermunt (2015) and Pankowska et al. (2018), which we will later apply in an off-the-shelf manner to provide a comparison to the model M_{AB} developed above.

The discrete time variable sequence $\boldsymbol{Y} = (Y_1, ..., Y_T)$ is a Markov chain, where T denotes the current time point of interest. This true sequence is unobserved, or hidden. At each time point one observes one or more fallible classifiers (or measure variables) that are dependent on the true variable. It is common to assume that the true and measure variables are independent for different units.

For the setting of two fallible classifiers in Table 1, we observe X_t for everyone in the population at time t, as well as Z_t for anyone in the sample at time t. Let $R_t = 1$ if Z_t is observed or 0 otherwise. We assume MAR for R_t given the observed data, and the indicators $\boldsymbol{R} = (R_1, ..., R_T)$ can be omitted in the HMM path diagrams.

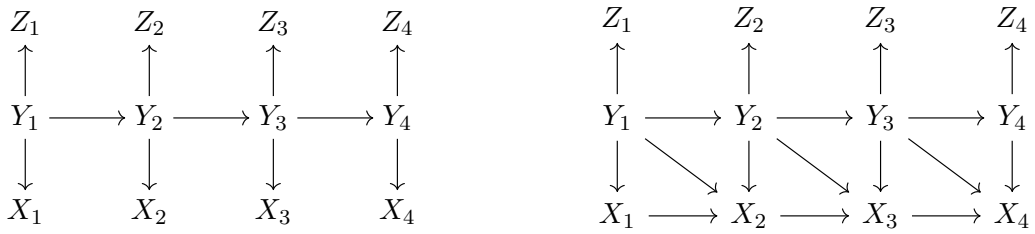

Figure 3: Path diagrams for HMM_0 (left) and HMM (right) conditional on A

Consider the path diagram of HMM_0 in Figure 3, where X_t is conditionally independent of all the other variables given (Y_t, A), and similarly for Z_t. Notice that A is omitted in Figure 3 to avoid cluttering the diagrams, as it would be connected to all the variables. In the literature this is often referred to as the assumption of independent classification errors (ICE). The joint probability of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ given A and $\boldsymbol{R}$ can be written as

$$P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y} \mid A, \boldsymbol{R} = \boldsymbol{r}) = \Pr(Y_1 = y_1 \mid A) \prod_{t=2}^{T} \Pr(Y_t = y_t \mid Y_{t-1} = y_{t-1}, A)$$
$$\prod_{t=1}^{T} \Pr(Z_t = z_t \mid Y_t = y_t, A)^{r_t} \Pr(X_t = x_t \mid Y_t = y_t, A)$$

Notice that, given the time points $t = 1, ..., T$, the term involving $\boldsymbol{Z}$ varies according to $\boldsymbol{r}$, i.e. the times anyone is actually in the sample over time.

However, the ICE assumption is most likely too simplistic for our application later. In particular, delay or mistake of reporting is the cause of misclassification by X, so that whether an error has already occurred at $t-1$ is likely to affect the misclassification probability of X_t, whether the error is repeated or corrected at t. Consider instead the path diagram of HMM in Figure 3, where X_t is conditionally independent of the other variables given (Y_{t-1}, X_{t-1}) in addition to (Y_t, A). This type of HMM has been considered by Pavlopoulos and Vermunt (2015) and Pankowska et al. (2018). Here we use a simple specification for X_t given (Y_t, Y_{t-1}, X_{t-1}), where the misclassification of X_t differs according to whether $X_{t-1} = Y_{t-1}$ or not, indicated by $\mathbb{I}(X_{t-1} = Y_{t-1})$. The joint probability of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ given A and $\boldsymbol{R}$ can then be written as

$$\begin{aligned} P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y} \mid A, \boldsymbol{R} = \boldsymbol{r}) &= \Pr(Y_1 = y_1 \mid A) \Pr(X_1 = x_1 \mid Y_1 = y_1, A) \\ &\prod_{t=2}^{T} \Pr(Y_t \mid Y_{t-1}, A) \Pr\left(X_t \mid Y_t = y_t, \mathbb{I}(X_{t-1} = Y_{t-1}), A\right) \\ &\prod_{t=1}^{T} \Pr(Z_t = z_t \mid Y_t = y_t, A)^{r_t} \end{aligned} \tag{7}$$

For estimation, Pavlopoulos and Vermunt (2015) apply pseudo MLE which incorporates the sampling design weights. We adopt adopt model-based inference in this paper. For

parameter estimation under the HMM, we use the Baum-Welch algorithm (Baum et al., 1970; Vermunt et al., 2008); see Appendix B. We can incorporate the discriminant B where, at each iteration of the Baum-Welch algorithm, we simply set $\Pr(X_T = Y_T|B = 1, A) = 1$ in the subpopulation of $B = 1$, in which case the model is denoted by HMM_B.

For variance estimation, we use basically the same bootstrap procedure described in Section 2.2. The only difference is that for each group of respondents in the last quarter with given (x_T, a, b), there is also a group of nonrespondents with the same (x_T, a, b), who have responded previously and contribute to the likelihood. Separate bootstrap resampling is applied to these last-quarter nonrespondents.

3 Application

3.1 Data and setting

Statistics Norway publishes monthly register-based employment statistics, the chief source of which is an administrative service coordinated by the Norwegian Labour and Welfare Administration, the Norwegian Tax Administration and Statistics Norway. Since its introduction in 2015, all employers are legally obliged to report all the contractual employer-employee relationships and various related payments every month. However, each month a certain amount of reports are actually corrections of reports in earlier months, which may be due to delays or mistakes. Consequently, an erroneous employment relationship in the data, say, at time t may be removed later at some time point, whereas a missing employment relationship at time t may appear at some time point later.

Table 4: Register Employed status for November 2018, two weeks after November (X) or six weeks after November ($\tilde{Y}$). LFS Employed status Z, September - November 2018

	All			$R = 1$		
	$\tilde{Y} = 0$	$\tilde{Y} = 1$	Total	$Z = 0$	$Z = 1$	Total
$X = 0$	1429974	5226	1435200	5577	1178	6755
$X = 1$	38191	2583254	2621445	554	13318	13872
Total	1468165	2588480	4056645	6131	14496	20627

Take the binary variable of Employed or not, denoted by 1 or 0. The left half of Table 4 illustrates the measurement errors in the register Employed status. The reference time point is November 2018, where X is the Employed status based on the data that are available two weeks after, and $\tilde{Y}$ is that after six weeks and the basis of monthly publication. The proportion of Employed is changed from 0.646 by X to 0.638 by $\tilde{Y}$, given the corrections during the month between them. To provide a context, the difference is about 4 times the standard error of the Norwegian LFS estimator of the proportion of Employed.

Note that *progressive* sources as such is the case for many administrative data on tax, benefits, migration, etc. The difference from one situation to another is merely the extent

of the resulting measurement errors rather than their existence. For instance, Zhang and Fosen (2012) examine the administrative sources for employment statistics, which existed before the current service was introduced in 2015, where progressive measurement errors are noticeable even years after the reference time point.

A binary Employed status Z is available from the LFS. The Norwegian LFS sample is a quarterly rotating panel consisting of 8 rotation groups, where close to 20000 persons are surveyed every quarter. The design at the time of these data is geographically stratified single-stage cluster sampling, where the clusters are families in the Central Population Register. The LFS Employed status follows the ILO definition, which differs from the register Employed status based on contractual employer-employee relationships.

The register and the LFS sample can be linked at the individual level using the unique person identification number, which exists in many register-rich countries including Norway. The discrepancies between Z and X can be seen in the right half of Table 4. Clearly, discrepancies are also unavoidable between Z and the true register Employed status Y. Notice that we do not require Z to measure the same construct as X (and Y) in our approach, but simply model the misclassification error of Z statistically where misclassification error is the case if $Z \neq Y$. However, because Z has a different definition to X (and Y), assuming the same $\Pr(Z|Y)$ for the whole population is unlikely to be realistic, which is why additional covariates A are needed as remarked earlier in Section 2.

Since X and Z are available at about the same time, the question arises whether one can adjust the errors of X given the additional information provided by Z, even though Z is also subject to misclassification. Notice that, since Z is treated as fallible, the fact that those collected in the LFS September or October are not entirely concurrent with X for November is not a principle obstacle, compared to the more important variance reduction by using quarterly instead of monthly LFS sample in this context.

Finally, the Norwegian LFS does suffer from survey nonresponse (e.g. Thomsen and Villund, 2011; Hamre and Heldal, 2013). Previous studies by Zhang (1999) and Zhang et al. (2013) suggest that nonresponse in the LFS is not MCAR, e.g. the proportion of $Z = 1$ is most likely to be lower among the nonrespondents. This makes it necessary to model the selection mechanism of R, in addition to the misclassification mechanisms.

Below we shall first introduce (B, A) and then apply the model M_{AB} to these data, to estimate the true proportion of register Employed Y. Provided this is possible, one may e.g. consider producing monthly flash estimates at an earlier time point than the current practice, whereas the completely register-based quarterly or yearly statistics can be published at a later time point, allowing more time for the progressive source to settle.

To apply the HMM for comparison, we use 2 successive quarterly LFS samples. This is the option requiring the least amount of extra data compared to applying the model M_{AB}. Let $T = 6$ be the month of interest, such as November 2018 in Table 4. Instead of a separate misclassification mechanism for X_1, we simply set $X_0 \equiv Y_0$, which allows one to use the same model for X_t given Y_t and $\mathbb{I}(X_{t-1} = Y_{t-1})$, for all $t = 1, ..., T$, under a model with fewer parameters. Pankowska et al. (2018) use the same approach. Finally, we shall let A be the same population stratification variable as for the model M_{AB}.

3.2 Choice of (B, A)

For this study we have access to the $(X, \tilde{Y}, Z)$ for June - November 2018. To define the discriminant based on the available data, we let $B = 1$ if an individual's register Employed status has no change at all in terms of $\tilde{Y}$ for July - October and X for November, and $B = 2$ otherwise, where X for November and $\tilde{Y}$ for October both become available two weeks after November. The intuition is that the true status Y is less likely to change in November, for someone with a stable status leading to November, in which case the observed status X for November is also less likely to be erroneous.

Table 5: Population $\Pr(X|\tilde{Y}, B)$ by discriminant B

	$B = 1$		$B = 2$	
	$\tilde{Y} = 0$	$\tilde{Y} = 1$	$\tilde{Y} = 0$	$\tilde{Y} = 1$
$X = 0$	0.9856	0.0005	0.9067	0.0181
$X = 1$	0.0144	0.9995	0.0933	0.9819
Total	1252700	2359329	215465	229151

The population $\Pr(X|\tilde{Y}, B)$ for $B = 1$ or $B = 2$ are shown in Table 5. It can be seen that the agreement between X and $\tilde{Y}$ is much better given $B = 1$ than $B = 2$. Since the variable $\tilde{Y}$ is naturally closer to the true Y, we take this to indicate that the misclassification errors of X are indeed much lower given $B = 1$ than $B = 2$.

It is unnecessary to be overly concerned with the chosen length of register history when defining B above, since making greater use of the relevant administrative data such as those mentioned in Section 1 is likely more effective for further reducing the errors of X given $B = 1$. However, such data are not available to this study. Moreover, should Statistics Norway decide to produce monthly flash estimates, it would surely involve many other details that we will not be able to cover here. Thus, we consider the definition of B here to be acceptable for demonstrating the potentials of the proposed methodology.

When it comes to the choice of additional covariates A with respect to noresponse, Nguyen and Zhang (2020) evaluate empirically reweighting methods for nonresponse adjustment in the Norwegian LFS. The register Employed status is the most effective covariate in this respect. In addition, age, gender, level of education, county, income and nationality are found to be among the most relevant ones. Many of these variables are commonly mentioned in household survey nonresponse studies. After some experimentation with the available variables, we find including age (in addition to X) to be nearly as effective as other more elaborated choices. Notice that age is also an important aspect of the definitional difference between Z and Y. A parsimonious stratification by age is chosen as A, where $A = 1$ for age 15 to 24, $A = 2$ for age 25 to 49, and $A = 3$ for age 50 to 74.

The left half of Table 6 gives the conditional distribution of X given $\tilde{Y}$ in the population according to the chosen A and B. The agreement between X and Y remains much better given $B = 1$ in each stratum defined by A. The right half of Table 6 shows the conditional

Table 6: Conditional $\Pr(X|\tilde{Y})$ and $\Pr(Z|\ \tilde{Y})$ given A and B

			Population			Sample		
			$X=0$	$X=1$	Total	$Z=0$	$Z=1$	Total
$A=1$	$B=1$	$\tilde{Y}=0$	0.9692	0.0308	216043	0.9167	0.0833	1045
		$\tilde{Y}=1$	0.0010	0.9990	250213	0.0696	0.9304	1408
	$B=2$	$\tilde{Y}=0$	0.9164	0.0836	98441	0.8042	0.1958	521
		$\tilde{Y}=1$	0.0121	0.9879	77649	0.2864	0.7136	419
$A=2$	$B=1$	$\tilde{Y}=0$	0.9759	0.0241	270405	0.6877	0.3123	999
		$\tilde{Y}=1$	0.0004	0.9996	1056768	0.0100	0.9900	5377
	$B=2$	$\tilde{Y}=0$	0.8993	0.1007	61148	0.5348	0.4652	230
		$\tilde{Y}=1$	0.0197	0.9803	93857	0.1050	0.8950	419
$A=3$	$B=1$	$\tilde{Y}=0$	0.9936	0.0064	766252	0.8595	0.1405	3836
		$\tilde{Y}=1$	0.0004	0.9996	1052348	0.0143	0.9857	5811
	$B=2$	$\tilde{Y}=0$	0.8975	0.1025	55876	0.6231	0.3769	268
		$\tilde{Y}=1$	0.0237	0.9763	57645	0.2755	0.7245	294

distribution of Z given $\tilde{Y}$ in the sample. Under the assumption (5), $\Pr(Z|Y, A=a)$ should be the same in each stratum by A whether $B=1$ or $B=2$. However, in light of Table 6, this assumption is unlikely to hold in reality. Indeed, comparing $(\tilde{Y}, X)$ to $(\tilde{Y}, Z)$ within each stratum of A, one can notice that (i) the observed probabilities of Z given $\tilde{Y}$ differ more between $B=1$ and $B=2$ compared to those of X given $\tilde{Y}$, and (ii) misclassification errors by Z are greater than by X.

Thus, it is intriguing to see whether useful adjustments, for the misclassification errors of X given $B=2$, can nevertheless be achieved by incorporating a second classifier Z that is less accurate (or weaker) than X itself, when the assumption that makes the misclassification mechanism of Z fully transportable from $B=1$ to $B=2$ is not quite true.

3.3 Results

Table 7 gives the data for the application of the model M_{AB}, where the indicator $R=1$ associated with the sample is suppressed to save space. Outside of the sample, where $R=0$, we have the population counts of X given (A, B). In addition, to apply the HMM and HMM$_B$, we need one quarterly LFS sample June - August 2018, and the corresponding monthly X for June - October 2018.

The sub-population with $B=2$ constitutes about 11% of the population for employment statistics. For simplicity, denote by μ_2 the true proportion of register Employed ($Y=1$) given $B=2$, and by $\hat{\mu}_2^{\text{method}}$ its estimate using a given method. The estimates by $\tilde{Y}$, X, model M_{AB}, HMM and HMM$_B$ are shown in the last row of Table 8, where $(A, B)=(+, 2)$

Table 7: Data for adjustments by model M_{AB}, November 2018

			$X=0$	$X=1$				$X=0$	$X=1$
$A=1$	$B=1$	$Z=0$	944	112	$A=2$	$B=1$	$Z=0$	681	60
		$Z=1$	76	1321			$Z=1$	289	5346
		$R=0$	208622	255181			$R=0$	263327	1057470
	$B=2$	$Z=0$	393	146		$B=2$	$Z=0$	117	50
		$Z=1$	90	311			$Z=1$	104	378
		$R=0$	90672	84478			$R=0$	56623	97733
$A=3$	$B=1$	$Z=0$	3293	87	All	$B=1$	$Z=0$	4918	259
		$Z=1$	518	5749			$Z=1$	883	12416
		$R=0$	757937	1051016			$R=0$	1229886	2363667
	$B=2$	$Z=0$	149	99		$B=2$	$Z=0$	659	295
		$Z=1$	101	213			$Z=1$	295	902
		$R=0$	51264	61695			$R=0$	198559	243906

Table 8: Estimated proportion of $Y=1$ by $\tilde{Y}$, X, model M_{AB}, HMM and HMM$_B$, given $(A,B)=(a,b)$, with population size N_{ab}. Estimated standard errors in parentheses.

(A,B)	N_{a2}	$\tilde{Y}$	X	M_{AB}	HMM	HMM$_B$
$(1,2)$	176090	0.441	0.482	0.413 (0.017)	0.265 (0.003)	0.468 (0.002)
$(2,2)$	155005	0.606	0.633	0.628 (0.023)	0.608 (0.002)	0.655 (0.001)
$(3,2)$	113521	0.508	0.546	0.495 (0.023)	0.533 (0.002)	0.517 (0.002)
$(+,2)$	444616	0.515	0.551	0.509 (0.013)	0.453 (0.002)	0.546 (0.001)
$(+,+)$	4056645	0.638	0.646	0.642 (0.001)	0.635 (0.001)	0.646 (0.000)

refers to the subpopulation of $B=2$ and $(A,B)=(+,+)$ refers to the whole population. Applying the model of Hui and Walter (1980) with the 3 subpopulations by A given $B=2$ does not yield plausible adjustments of $\hat{\mu}_2^X$, which are omitted here.

The difference of $\hat{\mu}_2^X$ and $\hat{\mu}_2^{\tilde{Y}}$ is 3.6%. It is reduced to 0.6% between the M_{AB}-estimate $\hat{\mu}_2^{M_{AB}}$ and $\hat{\mu}_2^{\tilde{Y}}$. The M_{AB}-estimate is closer to $\tilde{Y}$ than X in all the strata. If we treat the estimate $\hat{\mu}_2^{\tilde{Y}}$ as the target parameter μ_2, then $(\hat{\mu}_2-\hat{\mu}_2^{\tilde{Y}})^2$ for a given estimator $\hat{\mu}_2$ is by definition an unbiased estimator of MSE($\hat{\mu}_2$). Thus, if $|\hat{\mu}_2^{M_{AB}}-\hat{\mu}_2^{\tilde{Y}}|<|\hat{\mu}_2^{X}-\hat{\mu}_2^{\tilde{Y}}|$, then the estimate of MSE($\hat{\mu}_2^{M_{AB}}$) is smaller than that of MSE($\hat{\mu}_2^{X}$).

The difference $\hat{\mu}_2^{M_{AB}}-\hat{\mu}_2^{\tilde{Y}}=0.6\%$ given $B=2$ is comparable to $\hat{\mu}_1^{X}-\hat{\mu}_1^{\tilde{Y}}=0.5\%$ given $B=1$ which can be obtained from Table 5, and the estimate of the overall proportion μ using the model M_{AB} (Table 8) is about as precise as $\hat{\mu}_1^X$ for μ_1 in the subpopulation of $B=1$. Since the discriminant and transportability assumptions are not fully satisfied in

this application, as noted before, these results suggest that the exact assumptions (4), (5) and (6) can potentially be replaced by the approximate conditions

$$\Pr(X \neq Y \mid B = 1, A) \ll \Pr(X \neq Y \mid B = 2, A) \tag{8a}$$

$$\Pr(Z \mid Y, A, B = 1) \approx \Pr(Z \mid Y, A, B = 2) \tag{8b}$$

$$\Pr(Z, Y \mid X, A, B, R = 1) \approx \Pr(Z, Y \mid X, A, B)\ . \tag{8c}$$

Given (8) one can apply the model M_{AB} with a classifier Z that could be weaker than X, and obtain useful adjustments where X is worst according to the discriminant B.

The results of HMM differ quite much from those of HMM_B, overall as well as within each stratum by A, where HMM_B incorporating the discriminant B does seem to be an improvement over HMM. Although it seems not straightforward to obtain good results here using these models, it does not mean that it is impossible to achieve better adjustments using other HMMs. Notice that the HMMs tend to place relatively strong assumptions on the 'time homogeneity' of the latent Markov transition and both the misclassification mechanisms. Another concern, common to all latent class analysis, is that the model itself cannot tell which of the two estimated latent classes corresponds to $Y = 1$. We simply assume that, within each stratum, the total of misclassifications must be lower than that of correct classifications, in order to assigned one latent class to Employed.

Finally, the estimated standard errors shown in the parentheses (Table 8) are computed from 200 bootstrap replicate samples. It is clear that the estimated standard errors of either HMM are dominated by their respective biases, so that the associated uncertainty is underestimated, without taking into account the bias. Meanwhile, the estimated standard errors under the model M_{AB} are much larger, because it uses a much smaller amount of data compared to the HMMs. Although the estimators under the model M_{AB} here cannot be unbiased in truth, the confidence intervals based on the estimated standard errors are not unreasonable. For instance, the nominal 95% interval $(0.509 \pm 1.96 \cdot 0.013) = (0.483, 0.535)$ seems quite likely to cover the true μ_2.

3.4 A simulation study

We include here a small simulation study with set-ups that are close to the application above. The aim is to explore the sensitivity of the M_{AB}-estimator against departures from the discriminant and transportability assumptions, as well as to check the performance of the associated bootstrap variance estimator.

Let the population and sample be those in stratum $A = 1$, as shown in Table 7, with the target proportions $(\mu_1, \mu_2) = (0.537, 0.441)$ given $B = 1$ or 2. Based on the observed sample of $(Z, \tilde{Y})$ and the subsamples given $B = 1$ or 2, we obtain

$$(\lambda_{11}, \lambda_{00}) = (0.879, 0.881),\ (\lambda_{11}^{B=1}, \lambda_{00}^{B=1}) = (0.930, 0.917) \text{ and } (\lambda_{11}^{B=2}, \lambda_{00}^{B=2}) = (0.804, 0.714).$$

We have $(\theta_{11}^{B=1}, \theta_{00}^{B=1}) = (0.999, 0.969)$ for the actual subpopulation $\Pr(X = \tilde{Y} | \tilde{Y}, B = 1)$. In terms of these values, let three simulation set-ups be as given below.

a. Set $X = Y$ in the subpopulation of $B = 1$. Fix the subsample of Z given $B = 1$ as observed. Simulate Z in the subsample given $B = 2$ using the probabilities $(\lambda_{11|2}, \lambda_{00|2})$, where $\lambda_{11|2} = 0.7, 0.75, 0.8, 0.85, 0.9, 0.95$ and $\lambda_{00|2} = 0.75, 0.8, 0.85, 0.9, 0.95$.

b. Simulate X in the entire subpopulation of $B = 1$ using the probabilities $(\theta_{11}, \theta_{00})$, where $\theta_{11}, \theta_{00} = 0.95, 0.975, 0.99, 1$. Simulate the sample of Z using the probabilities $(\lambda_{11}, \lambda_{00})$ given above regardless $B = 1$ or 2.

c. Simulate X in the entire subpopulation of $B = 1$ using $(\theta_{11}, \theta_{00})$, where $\theta_{11} = 0.95$, 0.98, 0.995 and $\theta_{00} = 0.95, 0.97, 0.98, 0.99$. Simulate Z in the subsample given $B = 2$ using $(\lambda_{11|2}, \lambda_{00|2})$, where $\lambda_{11|2}, \lambda_{00|2} = 0.7, 0.75, 0.8, 0.85, 0.9, 0.95$.

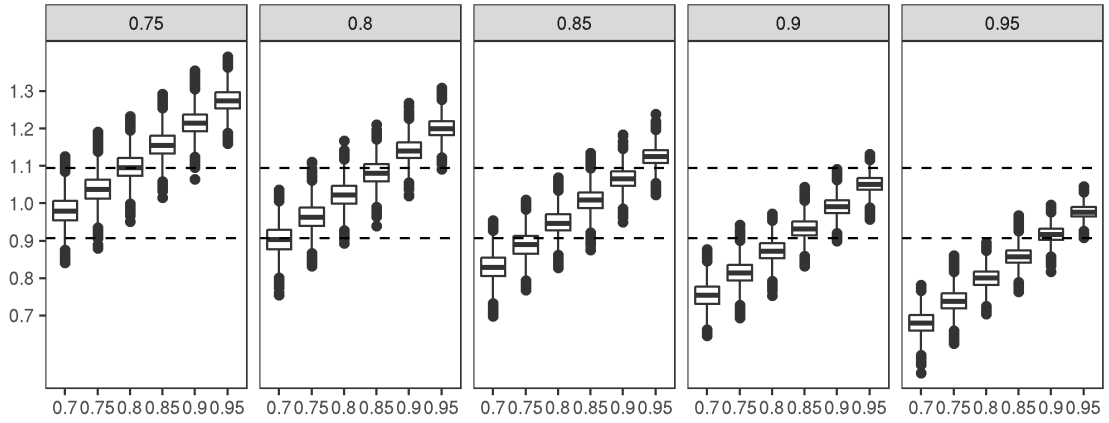

Figure 4: Set-up (a), box plots of $\hat{\mu}_2^{M_{AB}}/\mu_2$ for different combinations of $\lambda_{11|2}$ (x-axis of each panel) and $\lambda_{00|2}$ (top of each panel), 10000 simulations for each combination

The set-up (a) explores departures from the transportability assumption. The results are given in Figure 4, as box plots of the ratio $\hat{\mu}_2^{M_{AB}}/\mu_2$ for different combinations of $(\lambda_{11|2}, \lambda_{00|2})$ based on 10000 simulations for each combination. The horizontal dashed lines mark the region where an estimate is closer to μ_2 than $\hat{\mu}_2^X$ is. We notice the following.

- The combination $(\lambda_{11|2}, \lambda_{00|2}) = (\lambda_{11}^{B=2}, \lambda_{00}^{B=2})$ is sandwiched between the first two box plots in the second panel from the left, according to which it is more likely than not that the model M_{AB} would yield an improvement to $\hat{\mu}_2^X$ in the application, provided the discriminant assumption holds exactly. Despite a small departure from the discriminant assumption where $(\theta_{11}^{B=1}, \theta_{00}^{B=1}) = (0.999, 0.969)$ instead of $(1, 1)$, the actual $\hat{\mu}_2^{M_{AB}}$ is most likely closer to $\hat{\mu}_2$ than $\hat{\mu}_2^X$ to $\hat{\mu}_2$, as discussed in Section 3.3.

- The combination $(\lambda_{11|2}, \lambda_{00|2}) = (\lambda_{11}^{B=1}, \lambda_{00}^{B=1})$ is sandwiched between the last two box plots in the second panel from the right, according to which it is most likely that the model M_{AB} would yield an improvement over $\hat{\mu}_2^X$, under a true model M_{AB} that is

close to the assumed one in the application. For instance, the results obtained for $(\lambda_{11|2}, \lambda_{00|2}) = (0.9, 0.9)$ suggest $\text{MSE}(\hat{\mu}_2^{M_{AB}}) < \text{MSE}(\hat{\mu}_2^X)$ under this model M_{AB}; and similarly for $(\lambda_{11|2}, \lambda_{00|2}) = (0.85, 0.85)$ or $(0.95, 0.95)$.

- The estimator $\hat{\mu}_2^{M_{AB}}$ performs better than $\hat{\mu}_2^X$ when $\lambda_{11|2} \approx \lambda_{00|2}$ in all the panels. The likely reason is that $\lambda_{11}^{B=1} \approx \lambda_{00}^{B=1}$ in this simulation study. This suggests that the model M_{AB} is likely more robust against the departure from the transportability assumption, as long as $\lambda_{11|2}/\lambda_{11|1} \approx \lambda_{00|2}/\lambda_{00|1}$.

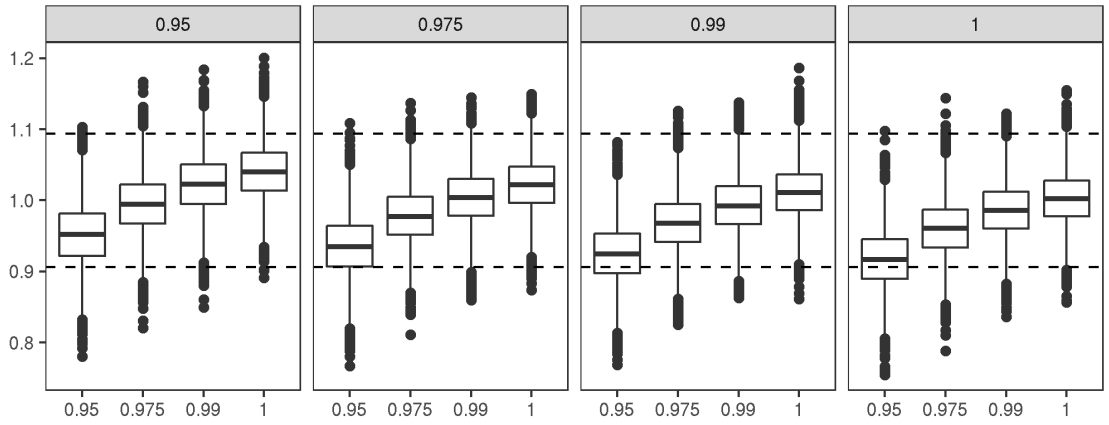

Figure 5: Set-up (b), box plots of $\hat{\mu}_2^{M_{AB}}/\mu_2$ for different combinations of θ_{11} (x-axis of each panel) and θ_{00} (top of each panel), 10000 simulations for each combination

The set-up (b) explores departures from the discriminant assumption. The results from 10000 simulations are given in Figure 5, as box plots of $\hat{\mu}_2^{M_{AB}}/\mu_2$ for different combinations of $(\theta_{11}, \theta_{00})$ in the subpopulation of $B = 1$, where the horizontal dashed lines mark the region where an estimate is closer to μ_2 than $\hat{\mu}_2^X$ is. Clearly, provided the transportability assumption, the improvements of the estimator $\hat{\mu}_2^{M_{AB}}$ over $\hat{\mu}_2^X$, both in terms of bias and MSE, are quite robust against small departures from the discriminant assumption.

Indeed, one only needs to be concerned with the results where $|\hat{\mu}_1^X - \mu_1|$ is sufficiently small. For instance, the deteriorating results for $\theta_{11} = \theta_{00} = 0.95$ in the leftmost panel is not a worrisome issue in practice, because it implies $\Pr(X \neq Y|B = 1) = 0.05$ which is unlikely to be acceptable for a definition of the discriminant. Recall that in the application earlier, we have $|\hat{\mu}_1^X - \mu_1| = 0.005$, which is a property of the discriminant B that can be tracked and verified retropectively over time.

The set-up (c) explores departures from both the discriminant and the transportability assumptions at the same time. For each combination of $(\theta_{11}, \theta_{00}, \lambda_{11|2}, \lambda_{00|2})$, the proportion over 10000 simulations where $|\hat{\mu}_2^{M_{AB}} - \mu_2| < |\hat{\mu}_2^X - \mu_2|$ is indicated in Figure 6. The separate conclusions above remain largely the same under both types of departure at the same time. In particular, for the panels in the bottom-right corner, where the violation

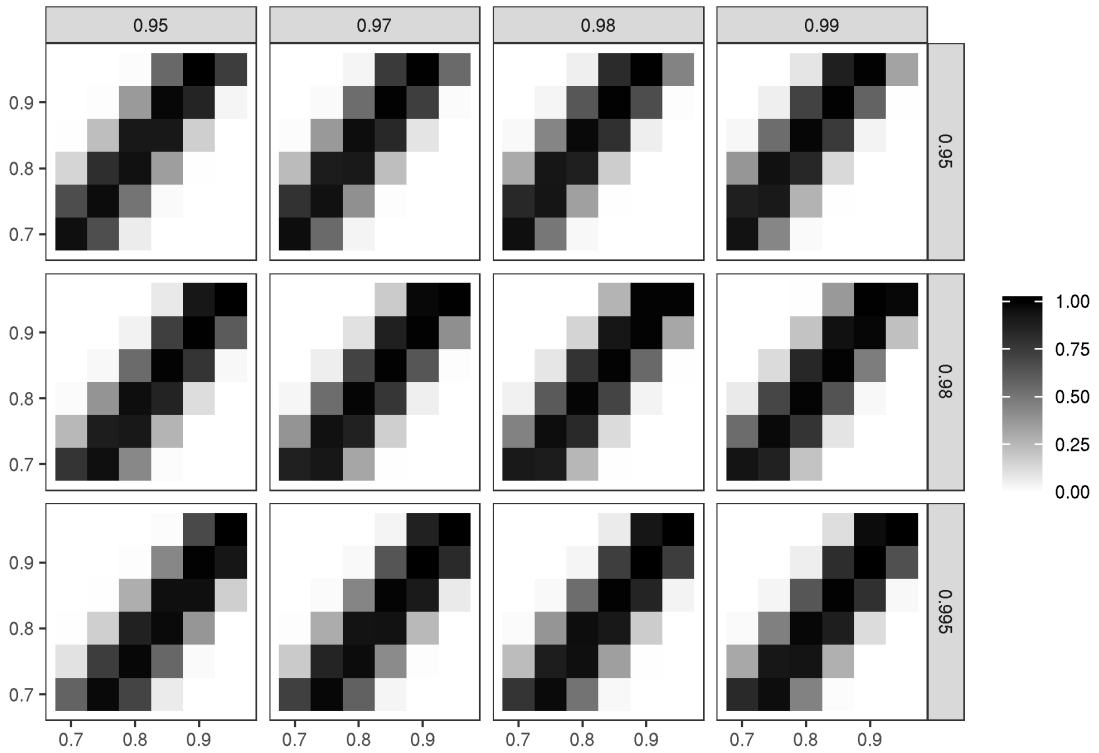

Figure 6: Set-up (c), proportions of results where $|\hat{\mu}_2^{M_{AB}} - \mu_2| < |\hat{\mu}_2^X - \mu_2|$ for different combinations of θ_{11} (right of each panel), θ_{00} (top of each panel), $\lambda_{11|2}$ (x-axis of each panel) and $\lambda_{00|2}$ (y-axis of each panel), 10000 simulations for each combination

of the discriminant assumption is the least, the estimator $\hat{\mu}_2^{M_{AB}}$ outperforms $\hat{\mu}_2^X$ when the transportability assumption holds exactly, and the improvement is quite robust against departures from the transportability assumption as long as $\lambda_{11|2}/\lambda_{11|1} \approx \lambda_{00|2}/\lambda_{00|1}$.

Lastly, we use double bootstrap to investigate the performance of the proposed bootstrap variance estimator. At the outer level, the replicate sample of Z is simulated by parametric bootstrap; at the inner level, the bootstrap variance procedure described in Section 2.2 is applied to the simulated sample to yield an estimate of $V(\hat{\mu}_2^{M_{AB}})$, denoted by $\hat{\sigma}^2$ here. We consider two scenarios for the outer level. In scenario-I, all the model assumptions are satisfied, where Y has a Bernoulli probability μ_1 given $B = 1$ or μ_2 given $B = 2$, and Z is generated from the Bernoulli distributions with $(\lambda_{11}, \lambda_{00})$, and $X = Y$ if $B = 1$ and X is generated using $(\theta_{11}^{B=2}, \theta_{00}^{B=2})$ if $B = 2$. Scenario-II is created in an *ad hoc* manner, where the transportability assumption is not satisfied. We fix Y to be the observed $\tilde{Y}$ in the population. We set $X = Y$ if $B = 1$ and fix X as observed if $B = 2$. We generate Z from the Bernoulli distributions with $(\lambda_{11}, \lambda_{00})$ given $B = 1$, whereas given $B = 2$ we generate Z using $(\psi_{11}^{B=2}, \psi_{00}^{B=2})$ which are the observed conditional probabilities $\Pr(Z|X, B = 2)$.

Table 9: Simulation results of bootstrap variance estimator $\hat{\sigma}^2$ for $\hat{\mu}_2^{M_{AB}}$

Scenario	Min	Q_1	Q_2	Mean	Q_3	Max	σ^2
I	0.00020	0.00030	0.00032	0.00032	0.00035	0.00048	0.00033
II	0.00027	0.00039	0.00042	0.00042	0.00045	0.00062	0.00042

For each scenario we have 10000 repetitions at the outer level, whereas the variance estimate $\hat{\sigma}^2$ at the inner level is based on 200 resamples as in the application reported above. The results of this double bootstrap are given in Table 9, which summarise the distribution of $\hat{\sigma}^2$ compared to $\sigma^2 = V(\hat{\mu}_2^{M_{AB}})$. We note that $\hat{\sigma}^2 = 0.00029$ in the stratum of $A = 1$ in the application reported earlier. In scenario-I, the model assumptions are satisfied and σ^2 is the unconditional variance of $\hat{\mu}_2^{M_{AB}}$. The conditional bootstrap variance estimator is essentially unbiased in these simulations, where 9477/10000 of the intervals $\hat{\mu}_2^{M_{AB}} \pm 1.96\hat{\sigma}$ contain the target parameter value μ_2. Scenario-II violates the transportability assumption; neither is M_{AB} exactly the data generation model otherwise. Nevertheless, the bootstrap variance estimator remains essentially unbiased for the actual $V(\hat{\mu}_2^{M_{AB}})$. We conclude that the potential bias due to model misspecification is a more critical element of the proposed adjustment method than the bootstrap variance estimator.

4 Summary

In the above we have developed a modelling approach for adjusting two fallible classifiers jointly observed in a nonprobability sample. A key innovation is the introduction of the discriminant B, which allows one to separate out the part of the population where the first classifier X is much worse than in the rest population, where misclassification adjustments is most effective for improving the estimation of the true classification. The bias caused by misclassification of X can be removed, if a second classifier Z together with X satisfy the assumptions (4), (5) and (6) exactly. Admittedly, this may not be the case in reality, as is common with any treatment of non-sampling errors. The application demonstrates that useful adjustment can nevertheless be achieved, when these assumptions are relaxed to (8), such that the proposed approach may potentially be helpful in many situations.

To implement the approach to produce official flash estimates, which accounts for the errors of X arising from the progressive nature of the administrative source, the model M_{AB} applied in Section 3 may be refined in two respects. First, we believe it is possible to improve the definition of the discriminant, by incorporating more extensively the relevant information available in the statistical data infrastructure at the NSO. Next, a more thorough process may be implemented to select the covariates A, in order to improve the transportability of the estimated misclassification mechanism of the second classifier Z. At the same time, one may look for a more parsimonious model specification, which can improve the tradeoff between bias adjustment and associated variance.

Two issues are then worth attention in practice. First, to obtain a more truthful as-

sessment of the uncertainty of the adjusted flash estimator, it will be useful to examine retrospectively the errors given by $e_t = \hat{\mu}_{2t}^{M_{AB}} - \hat{\mu}_{2t}^{\dot{Y}}$, where $\dot{Y}$ is the Employed status based on a sufficiently mature version of the register data, say, 3 - 6 months later than t. Analysis of e_t over time may also suggest other possibilities for improving the flash estimator.

Next, while normally the Norwegian labour market is by no means volatile, shocks do occur from time to time due to global events such as financial crisis or pandemic. In particular, we plan to apply the model M_{AB} to the data in 2020, where the labour market is subjected to considerable dynamics due to Covid-19. It would be interesting to study both the level and change estimates given by the flash-estimation methodology, in comparison with the LFS-based employment statistics which are traditionally considered as the leading indicator for changes in the labour market.

References

J. M. Abowd and A. Zellner. Estimating gross labor-force flows. *Journal of Business & Economic Statistics*, 3(3):254–283, 1985.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

P. P. Biemer and J. M. Bushery. On the validity of markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, 26(2):139–152, 2000.

I. Bross. Misclassification in 2 x 2 tables. *Biometrics*, 10(4):478–486, 1954.

T. T. Chen. Log-linear models for categorical data with misclassification and double sampling. *Journal of the American Statistical Association*, 74(366a):481–488, 1979.

T. T. Chen. A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine*, 8(9):1095–1106, 1989.

T. T. Chen, Y. Hochberg, and A. Tenenbein. Analysis of multivariate categorical data with misclassification errors by triple sampling schemes. *Journal of statistical planning and inference*, 9(2):177–184, 1984.

T. C. Chua and W. A. Fuller. A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82(397):46–51, 1987.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

M. di Zio, L.-C. Zhang, and A. de Waal. Statistical methods for combining multiple sources of administrative and survey data. *The Survey Statistician*, 2017(76):17–26, 2017.

D. Edwards. *Introduction to graphical modelling*. Springer Science & Business Media, 2012.

M. A. Espeland and C. L. Odoroff. Log-linear models for doubly sampled categorical data fitted by the em algorithm. *Journal of the American Statistical Association*, 80(391): 663–670, 1985.

Y. Haitovsky and J. Rapp. Conditional resampling for misclassified multinomial data with applications to sampling inspection. *Technometrics*, 34(4):473–483, 1992.

J. Hamre and J. Heldal. *Improved calculation and dissemination of coefficients of variation in the Norwegian LFS*. Documents 2013/46. Statistics Norway, 2013.

D. J. Hand. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3):555–605, 2018.

Y. Hochbeg and A. Tenenbein. On triple sampling scremes for estimating from binomial data with misclassification errors. *Communications in Statistics-Theory and Methods*, 12 (13):1523–1533, 1983.

Y. Hochberg. On the use of double sampling schemes in analyzing categorical data with misclassification errors. *Journal of the American Statistical Association*, 72(360a):914–921, 1977.

S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, pages 167–171, 1980.

J. Kuha and C. Skinner. Categorical data analysis and misclassification. In *Survey Measurement and Process Quality*, chapter 28, pages 633–670. John Wiley & Sons, Ltd, 1997. ISBN 9781118490013.

J. Kuha, C. Skinner, and J. Palmgren. Misclassification error. *Encyclopedia of Biostatistics*, 5, 1998.

R. T. Lie, I. Heuch, and L. M. Irgens. Maximum likelihood estimation of the proportion of congenital malformations using double registration systems. *Biometrics*, pages 433–444, 1994.

J. Magidson, J. Vermunt, and B. Tran. Using a mixture latent markov model to analyze longitudinal us employment data involving measurement error. *New Trends in Psychometrics*, pages 235–242, 2009.

N. D. Nguyen and L.-C. Zhang. An appraisal of common reweighting methods for nonresponse in household surveys based on the norwegian labour force survey and the statistics on income and living conditions survey. *Journal of Official Statistics*, 36(1):151–172, 2020.

P. Pankowska, B. Bakker, D. L. Oberski, and D. Pavlopoulos. Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use. *Statistical Journal of the IAOS*, 34(3):317–329, 2018.

D. Pavlopoulos and J. Vermunt. Measuring temporary employment: Do survey or register data tell the truth? *Survey Methodology*, 41(1):197–214, 2015.

J. M. Poterba and L. H. Summers. Reporting errors and labor market dynamics. *Econometrica: Journal of the Econometric Society*, pages 1319–1338, 1986.

S.-F. Qiu, H. Lian, G. Zou, and X.-S. Zeng. Interval estimation for a proportion using a double-sampling scheme with two fallible classifiers. *Statistical Methods in Medical Research*, 27(8):2478–2503, 2018.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

A. C. Singh and J. Rao. On the adjustment of gross flow estimates for classification error with application to data from the canadian labour force survey. *Journal of the American Statistical Association*, 90(430):478–488, 1995.

A. R. Swensen. Estimating change in a proportion by combining measurements from a true and a fallible classifier. *Scandinavian Journal of Statistics*, pages 139–145, 1988.

A. Tenenbein. A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, 65(331):1350–1361, 1970.

A. Tenenbein. A double sampling scheme for estimating from binomial data with misclassifications: Sample size determination. *Biometrics*, pages 935–944, 1971.

A. Tenenbein. A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics*, 14(1):187–202, 1972.

I. Thomsen and O. Villund. Using register data to evaluate the effects of proxy interviews in the norwegian labour force survey. *Journal of Official Statistics*, 27(1):87–98, 2011.

F. Van de Pol and R. Langeheine. Separating change and measurement error in panel surveys with an ppplication to labor market data. *Survey Measurement and Process Quality*, pages 671–688, 1997.

J. K. Vermunt. Longitudinal research using mixture models. In *Longitudinal Research With Latent Variables*, pages 119–152. Springer, 2010.

J. K. Vermunt, B. Tran, and J. Magidson. Latent class models in longitudinal research. *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, pages 373–385, 2008.

B.-J. Yoon. Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415, 2009.

L.-C. Zhang. A note on post-stratification when analyzing binary survey data subject to nonresponse. *Journal of Official Statistics*, 15(2):329–334, 1999.

L.-C. Zhang. On the bias in gross labour flow estimates due to nonresponse and misclassification. *Journal of Official Statistics*, 21(4):591–604, 2005.

L.-C. Zhang. Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1):41–63, 2012.

L.-C. Zhang and J. Fosen. A modeling approach for uncertainty assessment of register-based small area statistics. *Journal of the Indian Society of Agricultural Statistics*, 66:91–104, 2012.

L.-C. Zhang, I. Thomsen, and Ø. Kleven. On the use of auxiliary and paradata for dealing with non-sampling errors in household surveys. *International Statistical Review*, 81(2): 270–288, 2013.

A MLE under M_{AB} with stratification variable A

Since A is a stratification variable, the likelihood can be maximised separately within each stratum by A, where we need to show that the MLE of $\mu_{y|2a}$ is given by the matrix method in Sec. 2.2. Thus, we can conveniently drop a in the notation, i.e. as if the population consisted of a single stratum. The likelihood can then be given as

$$L(\boldsymbol{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{\mu}_{Y|2}) \propto \prod_{x=1}^{K}\prod_{z=1}^{K} \lambda_{zx}^{n_{zx|1}} \cdot \prod_{x=1}^{K}\prod_{z=1}^{K} \big(\sum_{y=1}^{K} \theta_{xy}\lambda_{zy}\mu_{y|2}\big)^{n_{zx|2}} \cdot \prod_{x}^{K} \big(\sum_{y=1}^{K} \theta_{xy}\mu_{y|2}\big)^{m_{x|2}} .$$

Similarly to Tenenbein (1972), re-parameterisation and the invariance property of the MLE lead to the result. Since $\theta_{xy}\mu_{y|2} = \eta_{yx}\mu_{x|2}$, we have $\sum_{y=1}^{K} \theta_{xy}\mu_{y|2} = \big(\sum_{y=1}^{K} \eta_{yx}\big)\mu_{x|2} = \mu_{x|2}$ for the 3rd term above, and $\sum_{y=1}^{K} \lambda_{zy}\theta_{xy}\mu_{y|2} = \big(\sum_{y=1}^{K} \lambda_{zy}\eta_{yx}\big)\mu_{x|2} = \psi_{zx}\mu_{x|2}$ for the 2nd term. Since the first two terms of the likelihood refers to $B = 1$ and $B = 2$ separately, the MLE of the parameters $(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ are given by the corresponding subsample proportions of (Z, X), i.e. the matrix method estimator of $(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$. Next, by the invariance of the MLE, the matrix method estimator $\hat{\boldsymbol{H}} = \hat{\boldsymbol{\Lambda}}^{-1}\hat{\boldsymbol{\Psi}}$ is the MLE of the matrix of η, and $\hat{\boldsymbol{\mu}}_{Y|2} = \hat{\boldsymbol{H}}\boldsymbol{\mu}_{X|2}$ is the MLE of $\boldsymbol{\mu}_{Y|2}$. This completes the proof. □

B Baum-Welch algorithm for HMM

The Baum-Welch algorithm is a special case of the EM algorithm, which uses the forward-backward algorithm in the E-step. Below we outline the algorithm in terms of the sample units, the notation of which is simpler. In practice, the different units are grouped by distinct paths, which constitute the sufficient statistic. Given sample unit i at time t, let

$$\begin{aligned}
\alpha_{i,t}(k) &= \Pr(X_{i,1:t}, Z_{i,1:t}, Y_{i,t} \mid A_i) \\
\beta_{i,t}(k) &= \Pr(X_{i,t+1:T}, Z_{i,t+1:T} \mid Y_{i,t}, X_{i,t}, A_i)
\end{aligned}$$

be the parameters in the forward-backward algorithm, respectively. The forward sequence $\alpha_{i,1:T}$ is given by

$$\alpha_{i,1}(k) = \Pr(Y_{i,1} = k \mid A_i) \Pr(X_{i,1} = x_{i,1} \mid Y_{i,1} = k, A_i) \Pr(Z_{i,1} = z_{i,1} \mid Y_{i,1} = k, A_i)^{\delta_{i,1}}$$

for $t = 1$, and the recursive formula

$$\begin{aligned}\alpha_{i,t}(k) = \sum_{j=1}^{K} & \alpha_{i,t-1}(j) \Pr(Y_{i,t} = k \mid Y_{i,t-1} = j, A_i) \\ & \Pr(X_{i,t} = x_{i,t} \mid Y_{i,t} = k, Y_{i,t-1} = j, X_{i,t-1} = x_{i,t-1}, A_i) \\ & \Pr(Z_{i,t} = z_{i,t} \mid Y_{i,t} = k, A_i)^{\delta_{i,t}}\end{aligned}$$

for $2 \leq t \leq T$. The backward sequence is given by

$$\begin{aligned}\beta_{i,T}(k) &= 1 \\ \beta_{i,t}(k) = \sum_{l=1}^{K} & \beta_{i,t+1}(l) \Pr(Y_{i,t+1} = l \mid Y_{i,t} = k, A_i) \\ & \Pr(X_{i,t+1} = x_{t+1} \mid Y_{i,t+1} = l, Y_{i,t} = k, X_{i,t} = x_{i,t}, A_i) \\ & \Pr(Z_{i,t+1} = z_{i,t+1} \mid Y_{i,t+1} = l, A_i)\end{aligned}$$

For all $i \in U$, $t = 1, ..., T$ and $k, l = 1, ..., K$, one obtains the estimated probabilities

$$\begin{aligned}&\Pr(Y_{i,t} = k \mid A_i, X_{i,1:T}, Z_{i,1:T}) = \frac{\alpha_{i,t}(k)\beta_{i,t}(k)}{\Pr(X_{i,1:T}, Z_{i,1:T})} \\ &\Pr(Y_{i,t-1} = k, Y_{i,t} = l \mid X_{i,1:T}, Z_{i,1:T}, A_i) = \alpha_{i,t-1}(k) \Pr(Y_{i,t} = l \mid Y_{i,t-1} = l, A_i) \\ &\qquad \Pr(X_{i,t} \mid Y_{i,t} = l, Y_{i,t-1} = k, X_{i,t-1}, A_i) \\ &\qquad \Pr(Z_{i,t} \mid Y_{i,t} = l, A_i))^{\delta_{i,t}} \beta_{i,t}(l) \times \big(\Pr(X_{i,1:T}, Z_{i,1:T} \mid A_i)\big)^{-1}\end{aligned}$$