How to test the Missing Data mechanism in a Hidden Markov Model

Malika Chassan, Didier Concordet

Journal Pre-proof

How to test the Missing Data mechanism in a Hidden Markov Model

Malika Chassan and Didier Concordet

PII: S0167-9473(23)00034-8
DOI: https://doi.org/10.1016/j.csda.2023.107723
Reference: COMSTA 107723

To appear in: *Computational Statistics and Data Analysis*

Received date: 22 February 2021
Revised date: 5 February 2023
Accepted date: 6 February 2023

Please cite this article as: M. Chassan and D. Concordet, How to test the Missing Data mechanism in a Hidden Markov Model, *Computational Statistics and Data Analysis*, 107723, doi: https://doi.org/10.1016/j.csda.2023.107723.

How to test the Missing Data mechanism in a Hidden Markov Model

Malika Chassan, Didier Concordet [1]

[a]*INTHERES, Université de Toulouse, INRAE, ENVT, 23 chemin des Capelles, 31076 Toulouse Cedex 3, France*

Abstract

A Hidden Markov Model with missing data in the outcome variable is considered. The initial and transition probabilities of the Markov chain and the emission probability of the HMM are allowed to depend on fully observed covariables. Tests for the ignorable and for the MCAR mechanisms are proposed. These tests do not require grouping the individuals by their missing pattern, making them easier to apply in practice. They are based on the estimates of the conditional probabilities of emitting a missing data given the latent state of the Markov chain and some observed covariables. When the ignorable mechanism holds, the conditional probabilities of emitting a missing value are the same for a given value of the observed variables. On the contrary, when the MCAR mechanism holds, these probabilities are all the same. A practical implementation of these tests based on simulations is proposed, along with a presentation of their performances. A real example from piglet farming illustrates their use.

Keywords: Hidden Markov Model, Hypothesis test, Missing data, Missing data mechanisms

1. Introduction

Missing data (MD) arises quite often in statistics. This situation should be analysed carefully since many classical statistical analyses cannot be directly applied when this occurs.

[1]Corresponding author E-mail address: didier.concordet@envt.fr

For example, an elementary way to handle MD is suppression. Referred as "deletion", this can be performed listwise (all the observations with at least one non-observed variable are excluded) or pairwise (only the observations where the variable of interest is missing are deleted, but the missing values of the unused variables are kept). However, these methods have two major drawbacks. First, they always decrease the size of the dataset. Second, if for a given variable, the probability of occurrence of an MD is correlated to the value that should have been observed (for example, in a survey about incomes, people with highest incomes are more likely to refuse to answer), deletion may generate a bias in the estimation.

Besides deletion, there are two classes of methods to handle missing values in a dataset: include them in the analysis using dedicated methods, such as, for example, the well-known EM algorithm of Dempster et al. (1977), see also Little and Rubin (2014) for an extended review of methods; or impute missing values (i.e., replace them with plausible data). The imputation problem will not be discussed here but the reader may refer to the book of Van Buuren (2018).

When one wants to include MD, the way to do it depends on the type of missing values mechanism involved in the dataset. In his seminal work, Rubin (1976) presented three types of MD mechanisms: Missing Completely at Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). In the case of MCAR mechanism, the missing values of a given variable are independent of all the other variables. In other words, the set of observations with a missing value is a uniform random subset of the complete set of observations.

In the case of MAR mechanism, the MD is allowed to depend on some observed variables but not on the MD themselves. The MD mechanism is then completely random, conditionally on some observed variables. When neither the MCAR nor MAR mechanisms hold, the missing values mechanism is MNAR. In this case, the missing observation depends on the missing variable itself; in other words, on the value that should have been observed, as for

the income survey example.

Many estimation methods (e.g., MLE) can be adapted for data with MAR or MCAR MD. MCAR and MAR mechanisms are often grouped under the name "ignorable MD", contrary to the MNAR mechanism, which is called "non-ignorable". Indeed, for MCAR or MAR MD, either the missing values do not cause a loss of information for the variable (except the decrease in the number of observations) or the information about the missing values can be found elsewhere in the dataset. In contrast, for the MNAR case, at least a part of the missing information is not available.

Consequently, it is important to distinguish between the ignorable and non-ignorable cases. Enders (2010) stated that "MCAR is the only mechanism yielding to testable propositions" and that if one wants to test the hypotheses $\mathcal{H}_0$ =MAR or $\mathcal{H}_0$ =MNAR, additional assumptions should be made. For example, Breunig (2019) proposed a test of the MAR mechanism that uses an instrumental variable. As announced by Enders, the application of Breunig's test assumes the availability of such an instrumental variable and that this variable does not have missing values.

Many tests for the MCAR mechanism have been proposed. The first tests were Student univariate tests from Dixon and Brown (1983). The individuals are gathered into two groups: for a given variable with MD, individuals with an MD and those without and the equality of means between the two groups for all other variables is tested. Then Little (1988) proposed a test where the individuals are gathered by their pattern of MD (a pattern corresponds to the set of variables whose values are missing for an individual). The main idea of the test relies on the fact that in the MCAR case, all these groups come from the same population, with the same joint distribution of variables. The equality of joint distributions implying the equality of moments, the test statistic measures the homogeneity of means and/or covariances across the groups of individuals with the same missingness pattern. Little proposed different versions of his test. First, a test for homogeneity of means when the covariance matrix

is supposed to be known. Second, a test for homogeneity of means when the covariance matrix is unknown but supposed to be the same for all groups and estimated by MLE. And third, a test for homogeneity of means and covariance matrices when the covariance matrices are estimated by MLE for each group separately. For the three versions, the test statistic is asymptotically chi-squared distributed (with adequate number of degrees of freedom) under $\mathcal{H}_0$ =MCAR and allows rejecting the null hypothesis of MCAR MD or considering it as compatible with the data. Finally, there exist several tests based on Little's work Jamshidian and Jalal (2010); Jamshidian et al. (2014); Kim and Bentler (2002). The main idea is still to test for the homogeneity of means and covariances. They aim at improving the performance of Little's test when working with smaller samples. They also allow overcoming a restriction of Little's work: Little's test with covariance matrices estimated by group requires the number of individuals in a group to be larger than the number of observed variables in this group. Otherwise, the empirical covariance matrix is singular.

Anyway, none of these tests apply when each individual has a distinct pattern of MD, a common situation when working with longitudinal data. The dataset is composed of several explanatory variables (quantitative and categorical), and one categorical outcome variable. This is the most comprehensible framework to present the test methodology, possible extensions to other types of variables are discussed in Section 7 . Only the outcome variable has MD. For an individual, the pattern of MD is the vector of measurement times when the outcome variable is missing. If the number of measurement times is large compared to the number of individuals, it is likely to have numerous groups of patterns with very few (or one) individuals, causing the tests presented above to perform badly. For example, Park and Davis (1993) proposed a test for the MD mechanism in the case of repeated categorical data. Once again, the idea of the test is to group individuals according to the pattern of their MD and to compare the moments of the distribution of these groups. Park proposed to fit a selected model to each group and then to test the homogeneity of the model pa-

rameter estimates across groups. When the number of individuals strongly varies between groups, or when the number of groups is high, the test has a low power. As a solution, Park suggested grouping individuals in only two groups: individuals without missing values and individuals with at least one missing value. But the authors admit that this approach could be misleading (Section 2.2 of Park and Davis (1993)).

As already mentioned, this article deals with longitudinal data with missing values in the outcome variable. A Hidden Markov Model (HMM) is used to model the observed variable. In an HMM, the outcome variable is assumed to be the consequence of an underlying and unobserved variable, whose evolution over time has a Markov structure. This unobserved variable is often referred to as a hidden or latent variable, while the outcome variable is referred to as the emitted or observed variable. An HMM is defined by the latent Markov chain (state space, initial and transition probabilities), and by the emission probability (the conditional distribution of the observed variable according to the latent variable).

Considering a categorical emission variable, the MD can be affected to an additional observed value. Then the missingness mechanism can be investigated throught the emission probabilities. A test for the MD mechanism in an HMM framework is proposed. To our knowledge, this has not already been proposed. More precisely, the latent Markov variable is defined such that it corresponds to the values of the emitted variable that should have been observed is there were no MD. For example, if the emitted variable $Y \in \{1, 2, \text{NA}\}$, then the latent variable U can take two values, 1 and 2. Then the probabilities of emitting $Y = \text{NA}$, conditionally on U can be estimated and compared.

If the emission probability is allowed to depend on an observed covariable (such as, for example, the gender of the respondent), estimates for the conditional emission probabilities of having an MD, given each value of the latent variable and the covariable can be computed. If the MD mechanism is MCAR, these probabilities should not depend on the values of the latent variable or covariable. The test of the MCAR null hypothesis consists in a test for the

homogeneity of the probability estimates for the emission law across the latent variable and covariable values. If the MD mechanism is ignorable (MAR or MCAR), these probabilities can depend on the covariable but not on the latent variable. Once again, the null hypothesis can be tested by a homogeneity test for the probability estimates, across the values of the latent variable for a given value of the covariable. Based on these properties, two tests are proposed. Test number 1 is a test of the ignorable null hypothesis (against MNAR) and test number 2 is a test of the MCAR null hypothesis (against not MCAR). These two tests allow distinguishing between the ignorable/non-ignorable cases and between the MCAR/non-MCAR cases. They do not allow distinguishing between MAR and MCAR, unless they are performed sequentially. Then, a third (procedure of) test is proposed. It first tests the ignorable null hypothesis and, if it is not rejected, it tests the MCAR null hypothesis.

The remainder of this paper is organised as follows. In Section 2, the HMM with covariables is described. In Section 3, the link between the MD mechanism and the HMM is developed. The test procedure is explained. Section 5 evaluates the tests' performance on simulated data and presents a practical implementation of the tests. Section 6 is dedicated to an application to real data. Lastly, Section 7 contains a discussion and conclusion.

2. Hidden Markov Model with covariables

An HMM is a statistical model where the studied phenomenon is assumed to be an unobserved Markov chain. An HMM with covariables allows the initial and transition probabilities of the hidden Markov chain and the emission probability to be functions of the covariables. The covariables are denoted by X_t when involved in the initial and transition probabilities of the Markov chain and O_t when involved in the emission probability.

Let $(U_t)_{t=1...T}$ be a Markov chain, with $U_t \in \mathcal{U}$, a discrete finite set. The initial distribution of U_1 is assumed to be a function of the covariable X_1 with parameter β. For $t = 2, ..., T$, the transition probability from time $t-1$ to t is assumed to be a function of X_{t-1} with

parameter γ. The observation at time t is denoted by Y_t. Then the conditional distribution of Y_t given $U_t = u$ and $O_t = p$ is assumed to be a function of u, p and a parameter α. All the parameters are gathered in $\theta = (\alpha, \beta, \gamma)$. Figure 1 gives a schematic representation of the model with its three components: the initialization of the Markov chain, its transitions and the emission process.

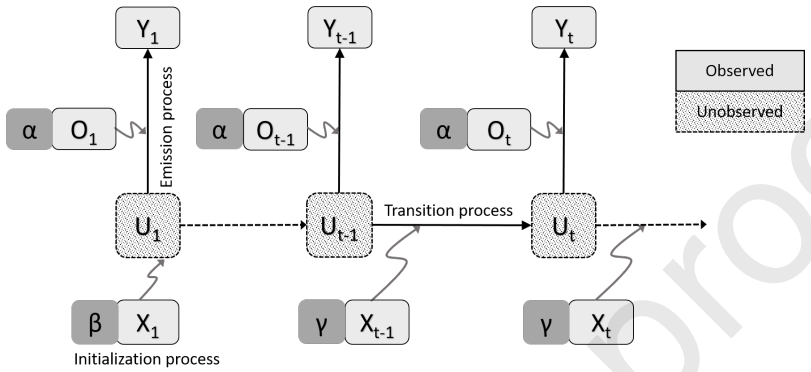

Figure 1: Representation of the HMM : the non-observed states of the Markov chain $(U_t)_t$ are the variable of interest. The U_t's depend on some observed covariables $(X_t)_t$. The observations $(Y_t)_t$ should give information on the U_t's,. However, some Y_t are missing. The Y_t's may depend on some observed covariables $(O_t)_t$.

As already mentioned, an HMM can handle missing values in an observation by considering them as an additional outcome of the emission process. If NA denotes the "missing value", then assumes that for all $t = 1, ..., T$, Y_t belongs to $\mathcal{Y} = \{1, ..., y_{max}, \text{NA}\}$. Using this notation, the parametric form of the emission process can be written as follows. For $u \in \mathcal{U}$, $p \in \mathcal{O}$ and $y \in \mathcal{Y}$, and if the NA state is taken as a reference,

$$\text{for } k \in 1, ..., y_{max} \quad \alpha_{k,u,p} = \log \frac{\mathbb{P}_\alpha(Y_t = k | U_t = u, O_t = p)}{\mathbb{P}_\alpha(Y_t = NA | U_t = u, O_t = p)}. \tag{1}$$

The $\alpha_{k,u,p}$ are gathered into the vector α according to a fixed order. To simplify, the emission probability is assumed to be constant over time (the α's do not depend on t). This assumption could be relaxed with some additional assumptions on the dependence of the emission probability on time.

A generalized linear model with a logit link function was used to model the influence of the covariables on the initial and transition probabilities of the Markov chain. Regarding the initial probability, let

$$\text{for} u \in \mathcal{U}, u \neq 1, \log \frac{\mathbb{P}_\beta(U_1 = u | X_1 = x_1)}{\mathbb{P}_\beta(U_1 = 1 | X_1 = x_1)} = {x_1}^\top \beta_u,$$

where x_1 is the vector containing the covariables for $t = 1$.

Regarding the transition probability, for $u, v \in \mathcal{U}$ and $t = 2, .., T$, let

$$\text{for } v \neq 1, \log \frac{\mathbb{P}_\gamma(U_t = v | U_{t-1} = u, X_{t-1} = x_{t-1})}{\mathbb{P}_\gamma(U_t = 1 | U_{t-1} = u, X_{t-1} = x_{t-1})} = {x_{t-1}}^\top \gamma_{u,v},$$

where the vector $\gamma_u = (\gamma_{u,\cdot})$ gathers the transition parameters from state u. As previously, x_{t-1} contains the covariables for time $t - 1$.

In the following section the three common types of MD mechanism and their consequences for the writing of the emission probability of the HMM are described. The estimation method is also presented, along with the two test statistics.

3. Missingness mechanism and HMM

The MD mechanism is linked to the parameter α alone. If the MD mechanism is MCAR, the probability of the event Y_t =NA neither depends on the value of Y_t that would have been observed nor on another variable. In the HMM framework, this is equivalent to saying that the probability of an MD does not depend on the state of the latent Markov chain U_t and does not depend on O_t either:

$$\begin{aligned} &\forall u, v \in \mathcal{U}, \forall p, q \in \mathcal{O}, \forall t = 1, ..., T, \\ &\mathbb{P}(Y_t = NA | U_t = u, O_t = p) = \mathbb{P}(Y_t = NA | U_t = v, O_t = q). \end{aligned} \tag{M1}$$

Equation (M1) can be written in more detail to make explicit the constraints on the parameter vector. Let $\tau(u,p) = 1 + \sum_{k \neq NA} e^{\alpha_{k,u,p}}$. Then $\mathbb{P}(Y_t = NA | U_t = u, O_t = p) = \frac{1}{\tau(u,p)}$ and Equation (M1) is equivalent to

$$\begin{aligned} &\forall u, v \in \mathcal{U}, \forall p, q \in \mathcal{O}, \\ &\tau(u,p) = \tau(v,q). \end{aligned} \tag{M1'}$$

If the MD mechanism is MAR, the probability of the event Y_t =NA does not depend on the value of Y_t that should have been observed but does depend on another observed variable, O_t here. This MAR hypothesis can be written as

$$\begin{aligned} &\text{for all } t = 1, ..., T, \text{the following conditions (a) and (b) are verified} \\ &(a)\ \forall p \in \mathcal{O}, \forall u, v \in \mathcal{U}, \\ &\qquad \mathbb{P}(Y_t = NA | U_t = v, O_t = p) = \mathbb{P}(Y_t = NA | U_t = u, O_t = p), \\ &(b)\ \exists p, q \in \mathcal{O}, p \neq q, \text{ such as } \forall u \in \mathcal{U}, \\ &\qquad \mathbb{P}(Y_t = NA | U_t = u, O_t = p) \neq \mathbb{P}(Y_t = NA | U_t = u, O_t = q). \end{aligned} \tag{M2}$$

The first line of (M2) says that the probabilities do not depend on U_t. The second line says that there are at least two values of the observed covariable leading to different probabilities of MD. This corresponds, for example, to probabilities that are equal for the same value of O_t but different when O_t changes. To test such a null hypothesis, something similar to what is often performed in a power analysis with post hoc inference should be done, that is, make assumptions about the values p and q considered and estimate the size of the effect: $\mathbb{P}(Y_t = NA | U_t = u, O_t = p) - \mathbb{P}(Y_t = NA | U_t = u, O_t = q)$ on the dataset. This approach is known to have a lot of problems Hoenig and Heisey (2001).

If only the condition (a) of (M2) is considered, the test of (M2 corresponds to the test of the ignorable hypothesis:

$$\begin{aligned} &\forall p \in \mathcal{O}, \forall u, v \in \mathcal{U}, \forall t, \\ &\mathbb{P}(Y_t = NA | U_t = v, O_t = p) = \mathbb{P}(Y_t = NA | U_t = u, O_t = p). \end{aligned} \tag{M3}$$

As for the MCAR case, Equation (M3) is equivalent to

$$\begin{aligned} &\forall p \in \mathcal{O}, \forall u, v \in \mathcal{U}, \\ &\tau(u, p) = \tau(v, p). \end{aligned} \tag{M3'}$$

When the ignorable hypothesis (M3) does not hold, the MD mechanism is MNAR, and the probability of the event Y_t =NA does depend on U_t. There is no precision about the dependency on an additional observed variable and a lot of mechanisms can correspond to this description. In the context of an HMM with covariables, the general MNAR hypothesis can be written

$$\begin{aligned} &\exists p \in \mathcal{O}, \exists u, v \in \mathcal{U}, u \neq v, \forall t, \\ &\mathbb{P}(Y_t = NA | U_t = v, O_t = p) \neq \mathbb{P}(Y_t = NA | U_t = u, O_t = p). \end{aligned} \tag{M4}$$

Two particular versions of the MNAR mechanism were investigated. In the so called MNAR1, the probability of the event Y_t = NA varies with U_t but is the same for all the values of O_t for a given value of U_t. In MNAR2 the probability of Y_t =NA depends on both U_t and O_t.

The build of a likelihood ratio test has failed. Actually, for both hypothesis tests, that for MCAR and that for an ignorable MD mechanism hypotheses, the constraints imposed by the null hypotheses cannot be easily written as explicit constraints on the parameter set. For example, if $\mathcal{Y} = \{1, 2, \text{NA}\}$, and according to the the multinomial logit parametrization

given in (1), the MCAR null hypothesis (M1') is equivalent to

$$\begin{aligned} &\forall u, v \in \mathcal{U}, \forall p, q \in \mathcal{O}, \\ &1 + e^{\alpha_{1,u,p}} + e^{\alpha_{2,u,p}} = 1 + e^{\alpha_{1,v,q}} + e^{\alpha_{2,v,q}}. \end{aligned} \tag{M1"}$$

The estimation of α with the constraints given by Equation (M1") is complicated. Thus, the estimation the parameters of the HMM under the MCAR or ignorable null hypotheses and the comparison of corresponding log-likelihoods is intricate.

To summarize, two sets of hypotheses to test can be tested. The first test, referred as test 1 is the test of the null hypothesis $\mathcal{H}_0$ = the MD mechanism is ignorable against $\mathcal{H}_1$ = the MD mechanism is MNAR. The second test, referred as test 2 is the test of the null hypothesis $\mathcal{H}_0$ = the MD mechanism is MCAR against $\mathcal{H}_1$ = the MD mechanism is not MCAR. The next section shows how to build these tests.

4. Building the tests

For both tests, the MLE for the parameters of the HMM was used to estimate the probabilities $\mathbb{P}(Y_t = NA | U_t = u, O_t = p)$. The general strategy for both tests was: (1) to estimate the model parameters without any constraint on α and (2) check whether or not the estimate $\hat{\alpha}$ is "close" to $\mathcal{H}_0$. Estimation for HMMs has been extensively documented and was not the purpose of the present paper. The reader could refer to Lagona and Picone (2013) and Bulla et al. (2012) and references within for exemples of the general estimation methods in HMM with ignorable missing data. θ was estimated by

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} \ell_\theta(y),$$

where $\ell_\theta(y)$ is the log-likelihood of the model, using the EM algorithm as described in Bartolucci and Farcomeni (2015). The method they proposed for the computation of the

Fisher information matrix of the model was also used. The EM algorithm presented in Bartolucci and Farcomeni (2015) was implemented in the package LMest Bartolucci et al. (2017). Nevertheless, this package does not handle covariables in both the emission and transition probabilities of the HMM. The estimation algorithm was then rewritten to include this capability.

For individual $i = 1, ..., n$ and $t = 1, ..., T$, consider the longitudinal observations $\mathbf{y}_{i,\cdot} = (y_t^i)_{t=1,...,T}$, where $y_t^i \in \mathcal{Y} = \{1, ..., y_{max}, \text{NA}\}$. The individuals were assumed to be independent, so the log-likelihood of the model is

$$\ell_\theta(y) = \sum_{i=1}^{n} \ell_\theta(\mathbf{y}_{i,\cdot}),$$

with

$$\begin{aligned}\ell_\theta(\mathbf{y}_{i,\cdot}) &= \log \mathbb{P}_\theta(\mathbf{Y}_{i,\cdot} = \mathbf{y}_{i,\cdot}) \\ &= \log \left(\sum_{\mathbf{u}_{i,\cdot} \in \mathcal{U}^T} \mathbb{P}_\theta(\mathbf{Y}_{i,\cdot} = \mathbf{y}_{i,\cdot} | \mathbf{U}_{i,\cdot} = \mathbf{u}_{i,\cdot}) \mathbb{P}_\theta(\mathbf{U}_{i,\cdot} = \mathbf{u}_{i,\cdot}) \right),\end{aligned}$$

where $\mathbf{U}_{i,\cdot}$ is the random vector $(U_{i,t})_{t=1,...,T}$ and $\mathbf{u}_{i,\cdot}$ is one realization of this vector and $\mathcal{U} = \{1, ..., y_{max}\}$.

The complete log-likelihood was used in the EM algorithm: for given a value of the parameter θ_k, the complete likelihood is defined by the conditional expectation $\ell^*_{(\theta|\theta_k)}(y) = \sum_{i=1}^n \ell^*_{(\theta|\theta_k)}(\mathbf{y}_{i,\cdot})$ with

$$\ell^*_{(\theta|\theta_k)}(\mathbf{y}_{i,\cdot}) = \mathbb{E}_{\theta_k}\Big(\log \mathbb{P}_\theta(\mathbf{Y}_{i,\cdot} = \mathbf{y}_{i,\cdot}, \mathbf{U}_{i,\cdot} = \mathbf{u}_{i,\cdot}) \Big| \mathbf{Y}_{i,\cdot} = \mathbf{y}_{i,\cdot} \Big),$$

where the expectation is taken along $\mathbf{u}_{i,.} \in \mathcal{U}^T$, i.e.,

$$\ell^*_{(\theta|\theta_k)}(\mathbf{y}_{i,.}) = \sum_{\mathbf{u}_{i,.} \in \mathcal{U}^T} \log \mathbb{P}_\theta(\mathbf{Y}_{i,.} = \mathbf{y}_{i,.}, \mathbf{U}_{i,.} = \mathbf{u}_{i,.}) \times \mathbb{P}_{\theta_k}(\mathbf{U}_{i,.} = \mathbf{u}_{i,.} | \mathbf{Y}_{i,.} = \mathbf{y}_{i,.}).$$

Using the Markov property and the notation defined above, the complete log-likelihood in the HMM with covariate framework, can be written

$$\begin{aligned}
\ell^*_{(\theta|\theta_k)}(\mathbf{y}_{i,.}) \quad &= \textstyle\sum_{t=1}^T \sum_{u \in \mathcal{U}} \log \mathbb{P}_\alpha(Y_t^i = y_t^i | U_t^i = u, O_t^i = o_t^i) \mathbb{P}_{\theta_k}(U_t^i = u | \mathbf{Y}_{i,.} = \mathbf{y}_{i,.}) \\
&+ \textstyle\sum_{u \in \mathcal{U}} \log \mathbb{P}_\beta(U_1^i = u | X_1^i = x_1^i) \mathbb{P}_{\theta_k}(U_1^i = u | \mathbf{Y}_{i,.} = \mathbf{y}_{i,.}) \\
&+ \textstyle\sum_{t=2}^T \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{U}} \log \mathbb{P}_\gamma(U_t^i = v | U_{t-1}^i = u, X_{t-1}^i = x_{t-1}^i) \mathbb{P}_{\theta_k}(U_t^i = v, U_{t-1}^i = u | \mathbf{Y}_{i,.} = \mathbf{y}_{i,.}),
\end{aligned}$$

where $o_t^i \in \mathcal{O} = \{1, ..., o_{max}\}$ is the vector of observed covariables for individual i and time t. The EM algorithm is iterative and starts at $k = 0$ from an initial point θ_0. Then the two steps E for Expectation and M for Maximization are alternated, as follows. The step E consists in computing, for a fixed θ_k, the probabilities $\mathbb{P}_{\theta_k}(U_t^i = u | \mathbf{Y}_{i,.} = \mathbf{y}_{i,.})$, $\mathbb{P}_{\theta_k}(U_1^i = u | \mathbf{Y}_{i,.} = \mathbf{y}_{i,.})$ and $\mathbb{P}_{\theta_k}(U_t^i = v, U_{t-1}^i = u | \mathbf{Y}_{i,.} = \mathbf{y}_{i,.})$. To do so, the Baum Welch recursions are used Baum et al. (1970). The step M consists in maximizing $\ell^*_{(\theta|\theta_k)}(\mathbf{y}_{i,.})$ with respect to θ. This step gives the estimate θ^{opt}. Then the algorithm returns to step E with $\theta_k = \theta^{opt}$.

Finally, the EM algorithm converges toward a local maximum assumed to be the MLE $\hat{\theta}_N$ of θ. To maximize the chances of finding the global maximum, the EM algorithm was run from several random starting points. Note that the computation time can be reduced using a short-run strategy Lagona and Picone (2013) or a stochastic EM Dedieu et al. (2014). $N = n \times T$ is the number of observations in the whole dataset (the number of individuals $\times$ the number of times). Let $\hat{\alpha}_N$ be the subvector of $\hat{\theta}_N$ containing the MLE of α. Let g be a function from $\mathbb{R}^{n_\alpha}$ to $\mathbb{R}^{n_u n_p}$ such that for α,

$$g(\alpha) = (g_i(\alpha))_{i=1..n_u n_p} = \left(\frac{1}{\tau(u,p)} \right)_{(u,p)},$$

where $n_u = \text{Card}(\mathcal{U})$ and $n_p = \text{Card}(\mathcal{O})$. It remains to check whether or not $\hat{\alpha}_N$ is close to $\mathcal{H}_0$. That is, whether the constraints imposed by $\mathcal{H}_0$ on $g(\hat{\alpha}_N)$ approximately hold. The MCAR mechanism hypothesis implies that all the $g_i(\alpha)$ are equal. The ignorable mechanism hypothesis implies that for each p, all the $g_i(\alpha)$ corresponding to the covariable $O_t = p$ are equal.

Let

$$T_N(g(\alpha)) = ||\Sigma_{\hat{\alpha}_N}^{-1/2}(g(\hat{\alpha}_N) - g(\alpha))||^2,$$

with $g(\hat{\alpha}_N)$ the function g evaluated at the MLE $\hat{\alpha}_N$, $g(\alpha)$ its expected value, and $\Sigma_{\hat{\alpha}_N}^{-1/2}$ its covariance matrix.

The idea of the test is to replace $g(\alpha)$ by what it should be if the null hypothesis was true and to measure whether the estimation is with this value.

If the MD mechanism is ignorable, this implies that $g(\alpha) = (\mu_1, \mu_2, ..., \mu_{n_p})^\top$, where each μ_j is repeated n_u times. In practice, the plug-in variable

$$\hat{\mu}_j = \frac{1}{n_u} \sum_{i=(u,p);p=j} g_i(\hat{\alpha}_N),$$

is used, and

$$T_N^1(g(\alpha)) = ||\Sigma_{\hat{\alpha}_N}^{-1/2}(g(\hat{\alpha}_N) - (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_{n_p})^\top)||^2.$$

Analogously, if the MCAR null hypothesis is true, $\mathcal{H}_0$ is equivalent to $\exists \mu \in \mathbb{R}; g(\alpha) = \mu \mathbb{1}_{n_u n_p}$, with $\mathbb{1}_{n_u n_p} = (1, 1, ..., 1, 1)^\top$. This time, the plug-in variable used is $\hat{\mu} = \frac{1}{n_u n_p} \sum_{i=1}^{n_u n_p} g_i(\hat{\alpha}_N)$, the empirical mean of $g(\hat{\alpha}_N)$ so that

$$T_N^2(g(\alpha)) = ||\Sigma_{\hat{\alpha}_N}^{-1/2}(g(\hat{\alpha}_N) - \hat{\mu}\mathbb{1}_{n_u n_p})||^2.$$

4.1. Asymptotic results

The asymptotic distributions of the two statistics $T_N^1(g(\alpha))$ and $T_N^2(g(\alpha))$ can be derived easily. The classical theory for MLE holds for HMM parameter estimates and so

$$I_N(\alpha)^{1/2}(\hat{\alpha}_N - \alpha) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbb{I}_{n_u n_p}) \text{ as } N \to +\infty,$$

where $I_N(\alpha)$ is the Fisher information matrix for α obtained with N observations, $\mathbb{I}_{n_u n_p}$ is the identity matrix of dimension $n_u n_p$ and $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

Because g is $\mathcal{C}^\infty$ on $\mathbb{R}^{n_u n_p}$, the delta method can be used to approximate the distribution of $g(\hat{\alpha}_N)$:

$$\Sigma_\alpha^{-1/2}(g(\hat{\alpha}_N) - g(\alpha)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbb{I}_{n_u n_p}), \text{ as } N \to +\infty, \tag{2}$$

where $\Sigma_\alpha = D_g(\alpha) I_N(\alpha)^{-1} D_g(\alpha)^\top$ and $D_g(\alpha)$ is the Jacobian matrix of g at α. By Slutsky's theorem, the convergence holds when replacing Σ_α by $\Sigma_{\hat{\alpha}_N} = D_g(\hat{\alpha}_N) I_N(\hat{\alpha}_N)^{-1} D_g(\hat{\alpha}_N)^\top$. Then, under $\mathcal{H}_0 =$ the MD mechanism is ignorable, respectively, $\mathcal{H}_0 =$ the MD mechanism is MCAR:

$$T_N^1(g(\alpha)) \xrightarrow{\mathcal{D}} T \text{ as } N \to +\infty, \text{ with } T \sim \chi^2((n_u - 1)n_p),$$

and respectively,

$$T_N^2(g(\alpha)) \xrightarrow{\mathcal{D}} T \text{ as } N \to +\infty, \text{ with } T \sim \chi^2(n_u n_p - 1),$$

where the number of degrees of freedom is, respectively, $(n_u - 1)n_p$ and $n_u(n_p - 1)$, because of the use of the plug-in estimators $\hat{\mu}_k$ and $\hat{\mu}$.

4.2. About the power of the tests

Regarding the statistic distribution under the alternative hypothesis: for the test of the MCAR mechanism, the alternative hypothesis is $\mathcal{H}_1 =$ the MD mechanism is not MCAR, i.e.,

$g(\alpha) \neq \mu \mathbb{1}_{n_u n_p}$, $\forall \mu \in \mathbb{R}$. Under $\mathcal{H}_1$, $T_N^2(g(\alpha))$ is still asymptotically chi-squared distributed, but this chi-square distribution is non-central with an unknown non-centrality parameter λ which depends on the distance between $g(\alpha)$ and the set of values verifying $\mathcal{H}_0$. Even if λ is unknown, the test is consistent. Indeed, if the MD mechanism is not MCAR, at least and without loss of generality, $g_1(\alpha) \neq g_2(\alpha)$. Then the random vector $g(\hat{\alpha}_N) - \hat{\mu}\mathbb{1}_{n_u n_p}$ is asymptotically normal with a non-zero mean. Moreover $\Sigma_{\hat{\alpha}_N}^{-1/2} = \sqrt{N}\Sigma_{\hat{\alpha}_1}^{-1/2}$ (the analogous matrix obtained using $I_1(\alpha)$). Thus, for any fixed quantile M, of the asymptotic distribution of $T_N^2(g(\alpha))$ under $\mathcal{H}_0$, the power of the test, $\mathbb{P}_{\mathcal{H}_1}\left(T_N^2(g(\alpha)) > M\right)$ tends to 1 as $N \to \infty$. An analogous argument gives the consistency of the test of the ignorable mechanism. The empirical powers of the two tests are given in Table 6. In the following section, the performances of these tests are evaluated and some modifications allowing a better control of the risks are proposed.

4.3. Empirical size

As illustrated in Section 5, the convergence of the test statistic empirical distribution toward the asymptotic one seems to be affected by the proportion of MD in the data set.

Although, it could be very interesting to investigate the theoretical convergence of the test statistic law toward its asymptotic distribution, this requires a method applicable even when the theoretical conditions for convergence are not met. This is the reason why the empirical quantiles and a classical type I risk of 5% are chosen. For a given data set, a first estimation of the HMM parameters using all the emission data (observed and MD) was done. Then, the estimated parameters of the hidden Markov chain were used to create simulated samples of the latent variable. Finally, data sets of emitted variable were generated using the latent variable samples and forcing the emission law to follow the MCAR constraint (all the emission probabilities equal to the total proportion of MD in the original data set). A similar procedure was applied to generate data set under the MAR mechanism (the proportions of

MD for each value of the covariable were used).

For the MCAR hypothesis test, the values of the test statistic T^1 were computed for each sample simulated with the MCAR mechanism and the 95% quantile was extracted from this empirical distribution. For the ignorable hypothesis test, the values of the test statistic T^2 were computed for each sample simulated with the MCAR or the MAR mechanisms.

5. Simulations

Because the study that motivated our interest in MD for HMM was linked to pig farming, the simulation study was conducted with data as close as possible to the the pig farming data. Diarrhea is a common health problem in post-weaning piglets. An experiment was organized to describe the way this occurs. Piglets were observed during their early age and their health status was assessed by a visual inspection of their feces. For each piglet, the feces were evaluated each day after the end of its weaning. Because the observation time was limited to 20 min per day and a group of piglets, not all the feces of a group were observed. This resulted in a lot of missing values in the time series of observations. Beside the assessment of the feces, real-time water and food consumption and weight measurements were recorded for each piglet. In what follows, these variables are referred to as the "behavior covariables".

The first question that arose in this study was to determine whether or not the missing observations were ignorable, in other words, did the unknown health status affect the occurrence of missing values ?

An HMM with covariables was used to model the feces evaluation. The latent Markov variable can be interpreted as the actual health status of the piglet. To simplify, only a binary health status is considered: the pig can be healthy or sick (1 or 2). Of course, the health status is probably a phenomenon much more regular than a binary variable, but in practice, veterinarian has to make a binary decision: to treat or not to treat the piglet. The

feces evaluation is the emitted variable. This evaluation was defined as normal or diarrheic (1 or 2), plus the missing observation, denoted by NA. The behaviour covariables were included in the initial and transition probabilities of the Markov chain. The feces evaluations were performed by three different observers. The variable representing the observer is included as a covariable in the emission process, i.e., $O_t \in \{1, 2, 3\}$.

In order to evaluate performance of the test, simulations were carried out. An estimation of the distribution of the initial values and growth coefficients for each type of behavior data (water, food, and weight), are obtained from healthy piglets. A total of 10000 piglets, i.e., behavior data and health status over 35 days, were simulated using these estimates. First, in order to control for the total number of observations with a sick health status, a sequence of health statuses was generated. For each piglet, two illness episodes of length five days were randomly spread over the time period, in order to obtain about 30% sick days. Secondly, the corresponding behavior data were generated by modifying each growth coefficient as follows. Compared to "healthy" growth coefficients, for each illness episode, for days 1 and 2, the growth coefficients were reduced by 80%, for days 3 and 4, the growth coefficients were reduced by 50% and for day 5, the growth coefficients were reduced by 70%.

Since the true health status of the piglet was not known before the study, the length of an illness episode and the way the growth coefficients were modified were fixed and visually estimated from the data. Centered Gaussian perturbations were added to the growth curves obtained (the standard deviations of these perturbations are the estimates of the standard deviations of the empirical distribution for each growth coefficient). Figure 2 shows an example of such simulation.

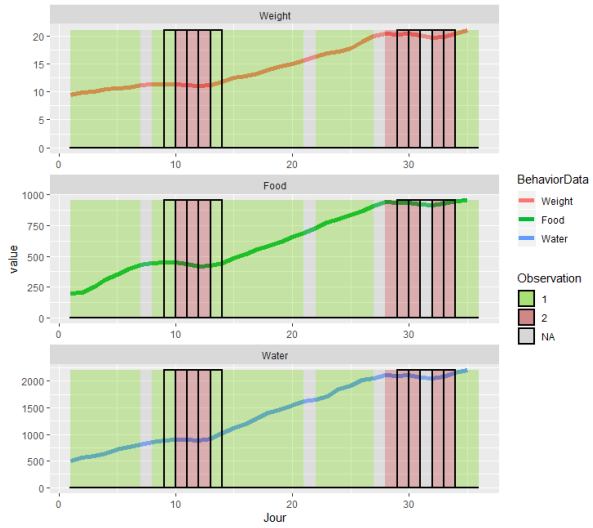

Figure 2: Example of simulated behaviour data for one individual. The color of the background indicates the value of the emitted variable, i.e., the feces evaluations Y_t: green for normal, red for pathological and gray for NA. The days into black lines indicates the true sick days ($U_t = 2$).

Finally, for each simulated piglet, several vectors of feces evaluations were generated: one for each MD mechanism (see below) and for different proportions of missing values in the dataset. In the piglet example, the emission law is the conditional distribution of the feces evaluation given the latent health status and the observer who made the evaluation. That is, the following discrete probabilities have to be fixed for each $u \in \{1, 2\}$ and each $o \in \{1, 2, 3\}$:

$$\mathbb{P}(Y_t^i = 1 | U_t^i = u, O_t^i = o)\,;\, \mathbb{P}(Y_t^i = 2 | U_t^i = u, O_t^i = o) \text{ and } \mathbb{P}(Y_t^i = NA | U_t^i = u, O_t^i = o).$$

For the MCAR mechanism, the probability of MD is the same for all the observers and health statuses. The probabilities of observing $Y = 1$ or $Y = 2$ were fixed so that a diseased piglet was more likely to have pathological feces than normal ones and so that a healthy piglet was more likely to have normal feces than pathological ones. The probabilities of $Y = 1$ or $Y = 2$ were allowed to vary between observers. An example of an emission law for MCAR is presented in Table 1 for a total proportion of missing values equal to 5%.

$\mathbb{P}(Y=y\|U=u,O=o)$		o=a		o=b		o=c	
		u		u		u	
MCAR case		1	2	1	2	1	2
y	1	0.931	0.0475	0.9025	0.095	0.855	0.1425
	2	0.019	0.9025	0.0475	0.855	0.095	0.8075
	NA	0.050	0.050	0.050	0.050	0.050	0.050

Table 1: Emission law of Y conditional on U and O, for the MCAR mechanism, for 5% of missing values in the simulated dataset. The last line gives the conditional probabilities of MD, which are all equal.

In the MAR mechanism, the probability of having MD depends on the observer variable. In our case, this corresponds to having the proportion of missing values varying between observers, but being the same for the two latent health statuses for a given observer. An example of the emission probabilities for a total proportion of missing values of 15% is given in Table 2.

$\mathbb{P}(Y=y\|U=u,O=o)$		o=a		o=b		o=c	
		u		u		u	
MAR case		1	2	1	2	1	2
y	1	0.8624	0.044	0.8075	0.085	0.738	0.123
	2	0.0176	0.836	0.425	0.765	0.082	0.697
	NA	0.12	0.12	0.15	0.15	0.18	0.18

Table 2: Emission law of Y conditional on U and O, for the MAR mechanism, for 15% of missing values in the simulated dataset. The last line gives the conditional probabilities of MD, which are equal for a given o.

The MNAR mechanism allows the missing values to depend on both observed and non observed variables. In our piglet example, this is equivalent to saying that the probability of observing a missing value is not the same whether the pig is healthy or sick, and eventually differs from one observer to another. Two subtypes of the MNAR mechanism were investigated: MNAR1 where the probability of a value's being missing depends only on the latent health status, and MNAR2 where it also depends on the observer.

Tables 3 and 4 give examples of emission laws for 5, respectively, 30, percent of missing values.

$\mathbb{P}(Y=y\|U=u,O=o)$		o=a		o=b		o=c	
		u		u		u	
MNAR1 case		1	2	1	2	1	2
y	1	0.926	0.0483	0.8978	0.0965	0.8505	0.1448
	2	0.019	0.9167	0.0472	0.8685	0.0945	0.8202
	NA	0.055	0.035	0.055	0.035	0.0.055	0.035

Table 3: Emission law of Y conditional on U and O, for the MNAR1 mechanism, for 5% of missing values in the simulated dataset. The last line gives the conditional probabilities of MD, which differ for $u = 1$ and $u = 2$.

$\mathbb{P}(Y=y\|U=u,O=o)$		o=a		o=b		o=c	
		u		u		u	
MNAR2 case		1	2	1	2	1	2
y	1	0.735	0.0425	0.6175	0.075	0.54	0.105
	2	0.015	0.8075	0.0325	0.675	0.06	0.595
	NA	0.250	0.1500	0.3500	0.250	0.40	0.300

Table 4: Emission law of Y conditional on U and O, for the MNAR2 mechanism, for 30% of missing values in the simulated dataset. The last line gives the conditional probabilities of MD, which differ for each value of u and o.

For each MD mechanism and each total proportion of MD, 500 random samples of $n = 300$ individuals were drawn from the 10000 individuals previously generated. For each dataset, the HMM parameters were estimated and the test statistics under the null hypothesis ignorable (M3) and under the null hypothesis MCAR (M1) were computed.

The empirical statistics distribution (grey lines) and the corresponding theoretical χ^2 law (black lines) are represented in Figure 3. Note that the asymptotic distribution of the test statistics is not always reached for some total proportions of MD. It is not surprising that the proportion of MD affects the convergence of the distribution of the statistics since it affects the estimation of the HMM parameters, see Dedieu et al. (2014) for an example. In our study, it can be seen that when there were too few MD (like 5%), then, as would be the case with any emission category rarely observed, the estimation is bad because there were not enough observations. At the opposite extreme, when there are too many MD (40% or 50%) the estimation of the more likely latent states sequence for a given observation sequence

is made with very uninformative data. Indeed, under the two null hypotheses (MCAR or ignorable), the MD could be emitted from each latent state with the same probability. For the two tests, it seems that there is an "ideal" proportion of MD, around 20%, resulting in a quicker convergence of the actual distribution to the theoretical one.

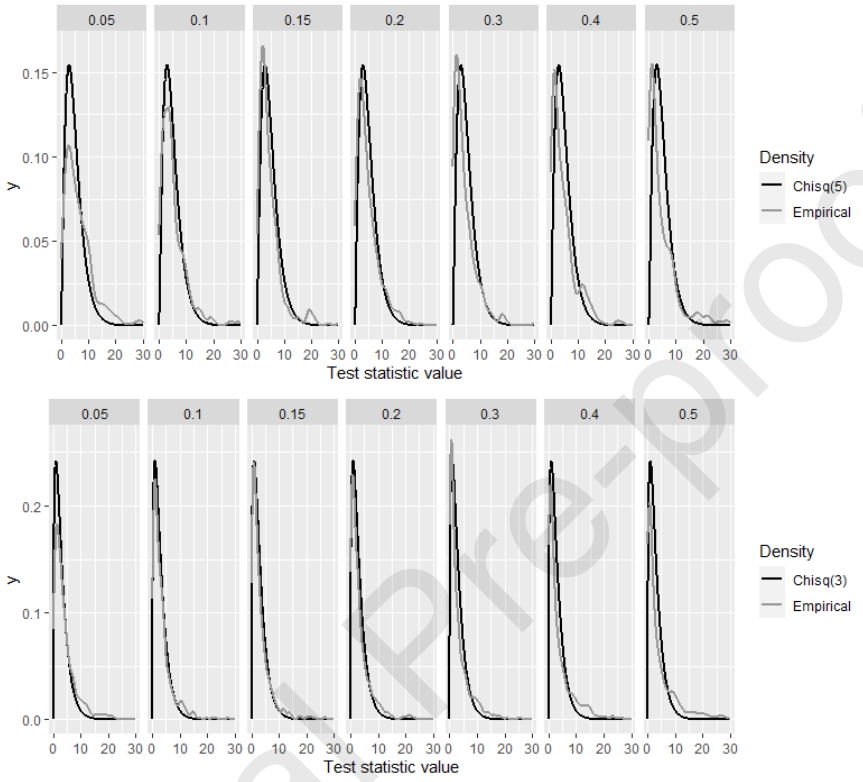

Figure 3: Empirical distribution of the test statistic versus its asymptotic χ^2 distribution (black plain line) for different proportions of missing values in the simulated data set. Distribution under $\mathcal{H}_0$=MCAR (top) and under $\mathcal{H}_0$= ignorable MD (bottom). When there are too many or too few MD, the asymptotic is not reached. 20% of MD seems to be the "best" proportion.

Computed as described in Section 4.3, the empirical quantiles of level 95%, for the two tests and for the different proportions of missing values, are shown in Table 5. As for the convergence of the test statistic law, the quantile varies with the total proportion of MD. This is an additional argument for the use of the empirical quantile obtained by simulation.

Prop NA	0.05	0.1	0.15	0.2	0.3	0.4	0.5
quantile under MCAR hyp. (theo. 11.07)	17.88	14.72	13.63	13.46	12.54	14.51	21.46
quantile under Ignorable hyp. (theo. 7.81)	12.87	11.48	11.22	11.41	11.76	15.06	20.88

Table 5: Empirical quantiles for the two test statistic distributions according to the total proportion of NA in the dataset. The theoretical values correspond to the 95% quantile of the $\chi^2(5)$ distribution for the MCAR hypothesis and $\chi^2(3)$ for the ignorable hypothesis. The quantiles vary with the proportion of NA, showing that the convergence towards the asymptotic is affected by this proportion.

The empirical power of the test and the results, presented in Table 6, were better than expected. Once again, a total proportion of MD around 20% seems to give the best results, and the power drastically decreases when the proportion of NA is higher than 30%. These conclusions should be taken carefully. Indeed, the total structure of the emission law also plays an important role in the quality of the parameter estimation: if the two latent states could emit the two informative evaluations ($Y = 1$ or 2) with probabilities that are close to each other, the estimation of the most likely sequence of latent states would be difficult. Consequently, the understanding of the MD mechanism will be poor.

One can also see that the power decreases for datasets with a total proportion of MD of 15%. A possible explanation for this non-intuitive phenomenon is given in Section 7.

6. Real data application

A total of $n = 153$ piglets were followed during the first week of weaning. This results in seven days of behavior data records (from day 1 to 7) and in five days of feces visual inspections (days 3 to 7). There were three different observers, so $O_t \in \{1, 2, 3\}$. Inn order to include some history in the covariables , the relative variations for each of the three behavior covariables for day t with respect to one and two days before were used as X_t. In other words, X_t contains the relative variations of the three behavior data (daily weight, food, and water) between day t and $t-1$ and between day t and $t-2$. An example may be seen in Figure 4.

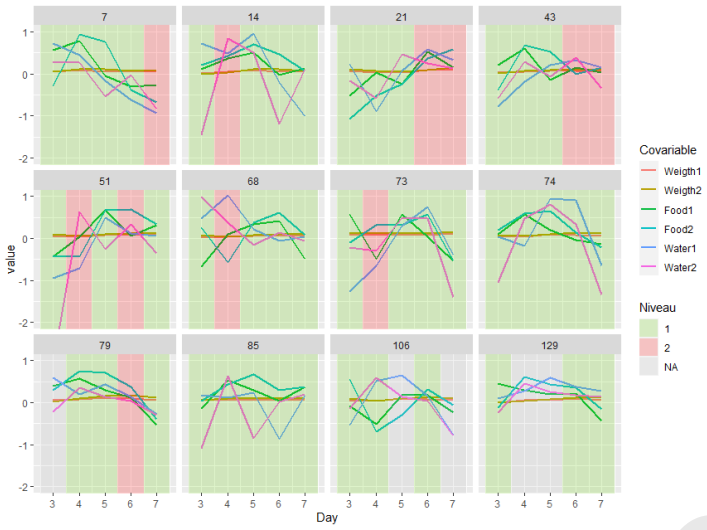

Figure 4: Example of covariables (standardized) along observation time, computed from real behaviour data. Background colour indicates the feces evaluation Y_t (red for sick, green for healthy, and gray for NA). "Cov1" indicates to the relative variations of the covariable "Cov" between day t and $t-1$, and "Cov2" indicates to the relative variations between day t and $t-2$. The relation between behaviour data and illness is not trivial.

The parameters of the HMM were estimated a first time, in order to have reference values for the emission probabilities of the evaluations $Y = 1$ or 2 for the MCAR and MAR mechanisms. For the MAR mechanism, the proportion of MD was fixed for each observer, and estimated from the real data. In the dataset, the total proportion of missing values was 8%. Simulations were performed to determine the empirical quantiles of the test statistics distribution under both the ignorable and MCAR hypotheses.

In this application, the ignorable MD hypothesis was rejected, concluding that the MD were of the non-ignorable type: the health status of a piglet affects the occurrence of a missing observation. The missing information cannot be recovered using the observed data alone. With this first result in hand, it is not useful to perform the second test (MCAR) but if it is done anyway, as expected, the MCAR hypothesis was also rejected. If the ignorable hypothesis had not been rejected, it would have been tempting to perform in sequence the test of the MCAR hypothesis. This way of proceeding is sequential and will be discussed in the last section. Figure 5 shows the differences between the estimated MD probabilities

and their expected counterparts under the ignorable and MCAR hypotheses. Although this overrides the conclusion allowed by the tests, it seems that the healthy piglets are more prone to producing MD than those with digestive troubles, which is rather realistic.

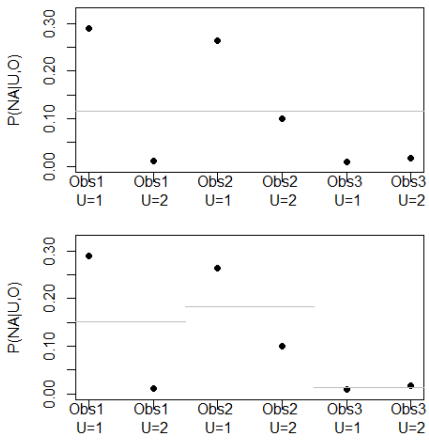

Figure 5: Estimated probabilities of emitting missing values, by observer and latent state ($\circ$). The grey horizontal lines correspond to the expected values under the null hypotheses: MCAR MD (top) and ignorable MD (bottom). These values correspond to the plug-in variables defined in $T_N^1(g(\alpha))$ and $T_N^2(g(\alpha))$. The estimates are too far from the expected values, leading to the rejection of the two hypotheses.

7. Discussion and conclusion

This section contains a comment on the application of the tests to real data. The parameters of the HMM should be estimated a first time for the simulated data to be generated. Currently, the variance of this first estimation is not considered. But as it is done in parametric bootstrap, it is taken into account when the parameters α, β, and γ are re-estimated for each simulation sample.

A sequential test could be built as follows: first, use test 1 to test the ignorable null hypothesis, and if it is not rejected, use test 2 to test the MCAR null hypothesis. At the end, this test would be able to classify a dataset with MD as MCAR, MAR, or MNAR. This procedure has some advantages but currently suffers some drawbacks. Indeed, concerning the null hypothesis MCAR, the final risk of type I (i.e., the proportion of datasets with MCAR missing data that are wrongly classified as MAR or MNAR) can be fixed to 5%

by choosing a couple of adequate quantiles for tests 1 and 2. The use of the sequential test allows slightly increasing the power of the MCAR null hypothesis test compared to the direct test of MCAR (test 2), see Table 6.

Prop NA	0.05	0.1	0.15	0.2	0.3	0.4	0.5
Power of MCAR direct test (no.2)	0.809	0.869	0.81	0.928	0.741	0.737	0.541
Power of MCAR sequential test	0.823	0.876	0.806	0.929	0.741	0.737	0.541
Power of MAR sequential test	0.74	0.96	0.67	0.91	-	-	-
Power of ignorable test (no.1)	0.76	0.76	0.71	0.92	0.78	0.66	0.37

Table 6: Power of the different tests for a type I risk of 5%. First line, direct test of the MCAR null hypothesis (test 2) ; second line, sequential test of the MCAR null hypothesis ; third line, sequential test of the MAR null hypothesis ; fourth line, test of the ignorable null hypothesis (test 1). For the tests 1 and 2, the best power is reached for 20% of MD. The performance strongly decreases when the proportion of MD exceeds 30%. The power for the MCAR hypothesis test is slightly improved by the use of the sequential test. For the MAR hypothesis test, it is not possible to guarantee a risk of 5% for proportions of MD above 30%. The unexpected decrease of power at 15% MD is discussed in the text.

Nevertheless, the control of the overall type I risk for the MAR hypothesis requires the control of the type II risk of the first test and of the type I risk for the second test. In our simulations, it has not been possible to find suitable pairs of levels for tests 1 and 2 in order to obtain a controlled risk of 5% for both the MCAR and MAR hypotheses. Moreover, in our simulations, the risk for the MAR hypothesis cannot always be set to 5% (even if considered alone, without fixing the risk for the MCAR hypothesis). This issue occurs when there is too much MD in the dataset (a proportion greater than 30%, as can be seen in Table 6). However, the sequential test procedure appears interesting since it would allow testing the MAR hypothesis, for which only a few tests exist.

For both tests considered in this article, a decreased power was observed when the proportion of MD is 15%. The power of a test is directly linked to the distance between the null and alternative hypotheses. In the HMM framework, two things, among others, can explain a bad separation of the hypotheses. Firstly, the theoretical distance between the hypotheses is small, i.e., the emission probabilities of Y =NA fixed for the simulation are not sufficiently different (the values chosen for the simulations can be seen in the last line in

Table 2). Secondly, the emission probability estimates are far from their actual values and consequently, the test is made on values that do not reflect the reality of the MD mechanism. As mentioned earlier, bad performances of estimators can be encountered when the probabilities of the different values of the outcome ($Y = 1, 2$,NA) are too close for different values of the latent variable. This implies that Y gives poor information on the latent Markov process.

The link between the choice of the values in the emission law and the distance between the null and alternative hypotheses of the test is then very complex and depends on many parameters. This interesting point is not investigated further. Nevertheless, the reader should bear in mind that for tests 1 and 2, the distance between $\mathcal{H}_0$ and $\mathcal{H}_1$ is not the same for the different proportions of MD explored by the simulations, explaining the non-monotonic variation of the test power.

The test methodology is only described for categorical variables. In case of a continuous emitted variable, the hidden variable should also be continuous, according to the method proposed here. In presence of MD, the distribution of the emitted variable is a mixture of a continuous distribution and a Dirac measure on $Y = \text{NA}$. It would then be necessary to ensure that certain theoretical results still hold since the distribution of Y is no longer dominated by the Lebesgue measure. Hopefully, the estimation of the mixture weight for the Dirac measure can be obtained. This estimation is a function of u, the value of the hidden state. The extension of the test proposed here could be to test if this function is a constant function.

In conclusion, a method to test the ignorable and MCAR mechanisms in an HMM framework was proposed. Their asymptotic laws were derived, but a method based on simulations was proposed in order to avoid theoretical constraints that are not always satisfied in real data applications. The two proposed tests have good power. The main advantage of the method proposed here is that it does not require additional assumptions other than those

that are made to use an HMM to model the observed variable. Indeed, the test statistics only use the estimates of the HMM that would have been obtained in a classical fitting of an HMM. A promising sequential procedure, making it possible to test the MAR hypothesis has also been explored but more work is needed to make it usable.

Acknowledgements

The authors thank the PigletDetect project co-funded by the Carnot Institute, IFIP and INRAe.

References

Bartolucci, F. and Farcomeni, A. (2015). Information matrix for hidden markov models with covariates. *Statistics and Computing*, 25(3):515–526.

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). LMest: An R package for latent markov models for longitudinal categorical data. *Journal of Statistical Software*, 81(4):1–38.

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

Breunig, C. (2019). Testing missing at random using instrumental variables. *Journal of Business & Economic Statistics*, 37(2):223–234.

Bulla, J., Lagona, F., Maruotti, A., and Picone, M. (2012). A multivariate hidden markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(4):544–567.

Dedieu, D., Delpierre, C., Gadat, S., Lang, T., Lepage, B., and Savy, N. (2014). Mixed hidden markov model for heterogeneous longitudinal data with missingness and errors in the outcome variable. *Journal de la Société Française de Statistique*, 155(1):73–98.

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

Dixon, W. J. and Brown, M. B. (1983). *BMDP statistical software*, volume 1. Univ of California Press.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power. *The American Statistician*, 55(1):19–24.

Jamshidian, M. and Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674.

Jamshidian, M., Jalal, S., and Jansen, C. (2014). Missmech: an r package for testing homoscedasticity, multivariate normality, and missing completely at random (mcar). *Journal of Statistical software*, 56(6):1–31.

Kim, K. H. and Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67(4):609–623.

Lagona, F. and Picone, M. (2013). Maximum likelihood estimation of bivariate circular hidden markov models from incomplete data. *Journal of Statistical Computation and Simulation*, 83(7):1223–1237.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

Little, R. J. A. and Rubin, D. B. (2014). *Introduction*, chapter 1, pages 1–23. John Wiley & Sons, Ltd.

Park, T. and Davis, C. S. (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics*, 49(2):631–638.

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

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.