Identifiability of discrete Input-output hidden Markov models with external signals

Etienne David, Jean Bellot, Sylvain Le Corff, Luc Lehéricy

Identifiability of discrete Input-Output hidden Markov models with external signals

Étienne David[1,2], Jean Bellot[2], Sylvain Le Corff[3], and Luc Lehéricy[4]

[1]Samovar, Département CITI, Telecom SudParis, Evry, 91011, France
[2]Heuritech, Paris, 75003, France
[3]LPSM, Sorbonne Université, UMR CNRS 8001, 75005, Paris
[4]Laboratoire J.A Dieudonné, Université Côte d'Azur, CNRS, Nice, 06100, France

Abstract

In this paper, we consider a bivariate process $(X_t, Y_t)_{t\in\mathbb{Z}}$ which, conditionally on a signal $(W_t)_{t\in\mathbb{Z}}$, is a hidden Markov model whose transition and emission kernels depend on $(W_t)_{t\in\mathbb{Z}}$. The resulting process $(X_t, Y_t, W_t)_{t\in\mathbb{Z}}$ is referred to as an input-output hidden Markov model or hidden Markov model with external signals. We prove that this model is identifiable and that the associated maximum likelihood estimator is consistent. Introducing an Expectation Maximization-based algorithm, we train and evaluate the performance of this model in several frameworks. In addition to learning dependencies between $(X_t, Y_t)_{t\in\mathbb{Z}}$ and $(W_t)_{t\in\mathbb{Z}}$, our approach based on hidden Markov models with external signals also outperforms state-of-the-art algorithms on real-world fashion sequences.

Keywords: Hidden Markov Model, Identifiability, Consistency, Expectation-Maximization, Fashion time series

1 Introduction

A hidden Markov model (HMM) is a bivariate process $(X_t, Y_t)_{t\in\mathbb{Z}}$ where $(X_t)_{t\in\mathbb{Z}}$ is a hidden Markov process and $(Y_t)_{t\in\mathbb{Z}}$ is an observed process such that at each time $s \in \mathbb{Z}$, the conditional law of Y_s given $(X_t)_{t\in\mathbb{Z}}$ depends only on X_s. Such models, introduced in the late 1960s, have been largely studied and applied in many disciplines, see for instance [Douc et al., 2014, Chopin et al., 2020, Särkkä, 2013] and references therein. As the process $(X_t)_{t\in\mathbb{Z}}$ is not observed, the maximum likelihood estimator (MLE) is intractable in most cases. The Expectation Maximization (EM) algorithm, introduced in [Dempster et al., 1977], overcomes this issue and provides a very appealing framework to infer these models with latent states. Variants of the EM algorithms have also been proposed to perform for instance online learning, see [Andrieu and Doucet, 2003, Cappé and Moulines, 2009, Le Corff and Fort, 2013], or inference of seasonal hidden Markov models [Touron, 2019].

Numerous theoretical results have been provided for hidden Markov models and their extensions. General identifiability results have been first obtained in [Gassiat et al., 2015] for HMMs with finite state space using the spectral method introduced in [Hsu et al., 2012]. This result establishes that given the law of a

triplet of observations, the transition matrix of the hidden states and the emission densities, i.e. the conditional densities of the observations given the states, can be identified up to a common permutation. This result was then extended by [Touron, 2019, Gassiat and Rousseau, 2016] to provide a theoretical justification of the use of nonparametric finite translation HMMs and of HMMs with seasonality. Finally, the work of [Gassiat et al., 2020] generalizes the identifiability guarantees to nonparametric translation HMMs with continuous state space, without any assumption on the distribution of the noise, and under a light tail assumption on the distribution of the latent variables.

The consistency of the MLE for HMMs has also been widely studied since the first result of [Baum and Petrie, 1966], where consistency is proved when both $(Y_t)_{t\in\mathbb{Z}}$ and $(X_t)_{t\in\mathbb{Z}}$ take values in a discrete space. A notable extension is proved in [Leroux, 1992] in the case where only $(X_t)_{t\in\mathbb{Z}}$ is assumed to be discrete. A general result, valid for a large class of nonlinear state space models and encompassing linear Gaussian state space models and finite state models, is then established in [Douc et al., 2011]. Additional results have also been proposed to analyze several extensions of HMMs. For instance, the authors of [Juang and Rabiner, 1985] introduce the autoregressive hidden Markov model. In this model, at each time t, the conditional law of Y_t given all the available information depends of X_t but also on some past values denoted $Y_{s:t-1}$ with $s < t$. Another HMM variant can be found in [Touron, 2019] where seasonal components are included in the law of $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$. For both extensions, identifiability of the model and consistency of the MLE have been proved, see [Touron, 2019, Douc et al., 2004].

Despite all these results, recent state-of-the-art forecasting models are for a large part based on recurrent neural networks [Hochreiter and Schmidhuber, 1997, Salinas et al., 2020] or sequence to sequence deep learning architectures [Vaswani et al., 2017, Li et al., 2019]. Intrinsically designed to deal with numerous heterogeneous data and include external signals, these new approaches overcome some limitations of HMMs and reach unprecedented accuracy levels in various frameworks and for numerous data sets, see for instance [Lim et al., 2021, Salinas et al., 2020, David et al., 2022]. However, these results have a cost: i) most of the recent state-of-the-art models are black boxes as the final forecast usually cannot be explained ; ii) very few theoretical guarantees exist for such deep learning architecture-based algorithms.

Regarding this new signal processing context, a main limitation of HMMs is the absence of theoretical results concerning the inclusion of meaningful external signals in the transition and emission kernels. In this paper, we consider bivariate processes $(Y_t, X_t)_{t\in\mathbb{Z}}$ with $(Y_t)_{t\in\mathbb{Z}}$ the observation process and $(X_t)_{t\in\mathbb{Z}}$ a discrete hidden process. Conditionally on an observed external signal $(W_t)_{t\in\mathbb{Z}}$, it is assumed that $(Y_t, X_t)_{t\in\mathbb{Z}}$ is a hidden Markov model so that at each time $t \in \mathbb{Z}$, the transition matrix of the hidden process and the emission law of the observation sequence depend on (W_t, W_{t+1}). Such models are inspired by the Input-Ouput models introduced in [Bengio and Frasconi, 1994], where a recurrent architecture is used to combine a discrete hidden state representing a past context and some input variables for sequence processing. The contextual HMMs of [Radenen and Artieres, 2012] also provide numerical insights of the benefit of adding external variables in the setting of Gaussian HMMs.

In this paper, we prove the identifiability of HMMs with external signals (Theorem 1) and the consistency of the associated MLE (Theorem 2). Then, we implement the MLE using the EM algorithm and show on a synthetic data set that it can recover the true set of parameters and that the addition of external signals does not prevent an efficient training process. Finally, we evaluate the proposed method on real world retail times series using the fashion data set introduced in [David et al., 2022]. This data set gathers the evolution of thousands of fashion items on social media and provides for each of them an external signal representing the behaviour of influencers. Using this additional influencers signal as an external signal in our new framework, we run experiments on a sample of fashion time series with challenging dynamics. Our approach outperforms state-of-the-art algorithms, including deep learning architectures on several time series and illustrates the potential of HMMs with external signals.

The paper is organized as follows. Section 2 extends the identifiability result of HMMs to HMMs with external signals following [Touron, 2019]. Then, the consistency of the MLE is proved in Section 3 following [Douc et al., 2014, Chapter 13]. Section 4 describes our experiments using synthetic data and some real-world fashion time series. Finally, a general conclusion and some research perspectives are given in Section 5.

Notations. For any vector v of size $m \geqslant 1$, $\mathrm{diag}(v)$ is the diagonal matrix in $\mathbb{R}^{m\times m}$ whose diagonal is given by v. By convention, one-dimensional vectors are row vectors in this paper. Given a sequence $(Y_t)_{t\in\mathbb{Z}}$, for any $s \in \mathbb{Z}$ and all $r \in \mathbb{Z}$ such that $r < s$, write $Y_{r:s} = (Y_r, \cdots, Y_s)$ with the convention $Y_{s:s} = Y_s$. For any finite set A, let $\mid A \mid$ be the cardinality of A. Consider a finite measurable space $(\mathsf{X}, \mathcal{X})$. For any transition matrix Q defined on $\mathsf{X}\times\mathsf{X}$, any measurable function h defined on X and any $A \in \mathcal{X}$, write for all $x, x' \in \mathsf{X}$,

$$Q(x,h) = Qh(x) = \sum_{x'\in\mathsf{X}} Q(x,x')h(x') \quad \text{and} \quad Q(x,A) = \sum_{x'\in\mathsf{X}} Q(x,x')\mathbb{1}_A(x')\,,$$

where $\mathbb{1}_A$ is the indicator function of the set A. In addition, for all sequences of transition matrices $\{Q_k\}_{k\in\mathbb{Z}}$, and all $r \leqslant s$, write for all $x_r \in \mathsf{X}$ and any measurable function h defined on X,

$$Q_{r,s}(x_r,h) = Q_{r,s}h(x_r) = \sum_{x_{r+1:s}\in\mathsf{X}^{s-r}} Q_{r+1}(x_r,x_{r+1})\cdots Q_s(x_{s-1},x_s)h(x_s)\,,$$

with the convention $Q_{s,s} = \mathrm{Id}$ and $Q_k = Q_{k,k+1}$.

2 Identifiability of HMMs with external signals

Let $(W_t)_{t\in\mathbb{Z}}$ be a sequence of external variables taking values in a measurable space $(\mathsf{W}, \mathcal{W})$. We assume that all variables W_t, $t \in \mathbb{Z}$, have the same support, and without loss of generality, we assume this support to be the whole space W. These auxiliary variables may account for the history of some additional time series, or any other available information. Let $(Y_t)_{t\in\mathbb{Z}}$ be the sequence of observations taking values in a measurable space $(\mathsf{Y}, \mathcal{Y})$ with Y a Polish space. We consider models in which there exists a hidden process $(X_t)_{t\in\mathbb{Z}}$ taking values in a finite space X such that if $\mathbb{P}$ is the distribution of the process $(W_t, X_t, Y_t)_{t\in\mathbb{Z}}$, the pair $(\mathsf{X}, \mathbb{P})$ satisfies the following assumptions.

H1 The conditional law of $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ given $(W_t)_{t\in\mathbb{Z}}$ satisfies

- for all $t \in \mathbb{Z}$, for all $k \in \mathsf{X}$,

$$\mathbb{P}(X_{t+1} = k \mid (X_s)_{s\leqslant t}, (W_s)_{s\in\mathbb{Z}}) = \mathbb{P}(X_{t+1} = k \mid X_t, W_t, W_{t+1})\,,$$

- For all $t \in \mathbb{Z}$ and for all measurable set $A \in \mathcal{Y}$,

$$\mathbb{P}\left(Y_t \in A \mid (X_s)_{s\in\mathbb{Z}}, (W_s)_{s\in\mathbb{Z}}\right) = \mathbb{P}(Y_t \in A \mid X_t, W_t)\,.$$

A graphical model to illustrate H1 is displayed in Figure 1.

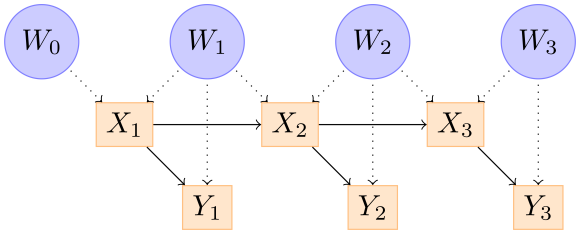

Figure 1: Graphical model of a latent data models with external signals satisfying H1.

H2 For all $t \in \mathbb{Z}$, for all w_{t-1}, w_t, $w_{t+1} \in \mathsf{W}$, all x_{t-1}, x_t, x_{t+1} in X and all $A \in \mathcal{Y}^{\otimes 3}$, the limit $\lim_{\varepsilon\to 0} \mathbb{P}(X_{t-1:t+1} = x_{t-1:t+1}, Y_{t-1:t+1} \in A \mid W_{t-1:t+1} \in B(w_{t-1:t+1}, \varepsilon))$ exists and there exist functions

$$(w, x) \in \mathsf{W} \times \mathsf{X} \longmapsto \pi_{t,w}(x),$$
$$(w, w', x, x') \in \mathsf{W} \times \mathsf{W} \times \mathsf{X} \times \mathsf{X} \longmapsto Q_{t|w,w'}(x, x')$$

and

$$(w, x, A) \in \mathsf{W} \times \mathsf{X} \times \mathcal{Y} \longmapsto \nu_{t|w,x}(A)$$

such that $\pi_{t,w}$ is a probability vector on X, $Q_{t|w,w'}$ is a transition matrix on $\mathsf{X}\times\mathsf{X}$, $\nu_{t|w,x}$ is a probability measure on $(\mathsf{Y}, \mathcal{Y})$ and

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \mathbb{P}(X_{t-1:t+1} = x_{t-1:t+1}, Y_{t-1:t+1} \in A \mid W_{t-1:t+1} \in B(w_{t-1:t+1}, \varepsilon)) \\
&\quad =: \mathbb{P}(X_{t-1:t+1} = x_{t-1:t+1}, Y_{t-1:t+1} \in A \mid W_{t-1:t+1} = w_{t-1:t+1}) \\
&\quad = \pi_{t-1,w_{t-1}}(x_{t-1}) Q_{t-1|w_{t-1},w_t}(x_{t-1}, x_t) Q_{t|w_t,w_{t+1}}(x_t, x_{t+1}) \\
&\qquad\qquad \times \int_A \otimes_{s=t-1}^{t+1} \nu_{s|w_s,x_s}(dy_{t-1:t+1}). \qquad (1)
\end{aligned}$$

In particular, for all $t \in \mathbb{Z}$, for all $x_t, x_{t+1} \in \mathsf{X}$ and for all $w_t, w_{t+1} \in \mathsf{W}$,

$$\pi_{t,w_t}(x_t) = \lim_{\varepsilon\to 0} \mathbb{P}(X_t = x_t \mid W_t \in B(w_t, \varepsilon)),$$

$$Q_{t|w_t,w_{t+1}}(x_t, x_{t+1}) = \lim_{\varepsilon\to 0} \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t, W_{t:t+1} \in B(w_{t:t+1}, \varepsilon)),$$

and for all measurable set $A \in \mathcal{Y}$,

$$\nu_{t|w_t,x_t}(A) = \lim_{\varepsilon\to 0} \mathbb{P}(Y_t \in A \mid X_t = x_t, W_t \in B(w_t, \varepsilon)).$$

The conditional law of the process $(X_t, Y_t)_{t\in\mathbb{Z}}$ given $(W_t)_{t\in\mathbb{Z}}$ is the law of a hidden Markov model with transition matrices $Q_{t|W_t,W_{t+1}}$ and emission densities $(\nu_{t|W_t,k})_{k\in\mathsf{X}}$, $t \in \mathbb{Z}$.

H3 For all $t \in \mathbb{Z}$, for all $w_t, w_{t+1} \in \mathsf{W}$, $Q_{t|w_t,w_{t+1}}$ is invertible, and for all $w_t \in \mathsf{W}$, $(\nu_{t|w_t,x_t})_{x_t\in\mathsf{X}}$ are linearly independent.

H4 For all $t \in \mathbb{Z}$ and for all $k \in \mathsf{X}$ and $w_t \in \mathsf{W}$, $\pi_{t,w_t}(k) > 0$.

Theorem 1. *Assume that* $(\mathsf{X}, \mathbb{P})$ *satisfies H1-4 with parameters*

$$\vartheta = \{\pi_{t,w_t}, Q_{t-1|w_{t-1},w_t}, \nu_{t|w_t,x}\}_{t\in\mathbb{Z},x\in\mathsf{X},w_{t-1},w_t\in\mathsf{W}} .$$

Let $(\widetilde{\mathsf{X}}, \widetilde{\mathbb{P}})$ *be such that* $|\widetilde{\mathsf{X}}| \leqslant |\mathsf{X}|$ *and satisfying Assumptions H1-2 with parameters*

$$\tilde{\vartheta} = \{\widetilde{\pi}_{t,w_t}, \widetilde{Q}_{t-1|w_{t-1},w_t}, \widetilde{\nu}_{t|w_t,x}\}_{t\in\mathbb{Z},x\in\mathsf{X},w_{t-1},w_t\in\mathsf{W}} .$$

If for all $t \in \mathbb{Z}$*, the distribution of* $(W_{t-1:t+1}, Y_{t-1:t+1})$ *is the same under* $\mathbb{P}$ *and* $\widetilde{\mathbb{P}}$*, then* $|\widetilde{\mathsf{X}}| = |\mathsf{X}|$ *and there exists a family of bijections* $(\sigma_{t,w})_{t\in\mathbb{Z},w\in\mathsf{W}}$*, where* $\sigma_{t,w} : \mathsf{X} \longrightarrow \widetilde{\mathsf{X}}$*, such that for all* $t \in \mathbb{Z}$*,* $x, x' \in \mathsf{X}$ *and* w_{t-1}*,* $w_t \in \mathsf{W}$*,*

$$\begin{cases} \pi_{t-1,w_{t-1}}(x) = \widetilde{\pi}_{t-1,w_{t-1}}(\sigma_{t-1,w_{t-1}}(x)), \\ Q_{t-1|w_{t-1},w_t}(x, x') = \widetilde{Q}_{t-1|w_{t-1},w_t}(\sigma_{t-1,w_{t-1}}(x), \sigma_{t,w_t}(x')), \\ \nu_{t|w_t,x} = \widetilde{\nu}_{t|w_t,\sigma_{t,w_t}(x)} . \end{cases}$$

Proof. For all $t \in \mathbb{Z}$ and w_{t-1}, w_t, w_{t+1} in W, the limits $\mathbb{P}(\cdot \mid W_{t-1:t+1} = w_{t-1:t+1})$ and $\widetilde{\mathbb{P}}(\cdot \mid W_{t-1:t+1} = w_{t-1:t+1})$ from (1) define probability distributions of $(X_{t-1:t+1}, Y_{t-1:t+1})$ depending on $w_{t-1:t+1}$. We write $\mathbb{E}[\cdot \mid W_{t-1:t+1} = w_{t-1:t+1}]$ the conditional expectation under the probability distribution $\mathbb{P}(\cdot \mid W_{t-1:t+1} = w_{t-1:t+1})$. Moreover, since $\mathbb{P}(Y_{t-1:t+1} \in A \mid W_{t-1:t+1} \in B(w_{t-1:t+1}, \varepsilon)) = \widetilde{\mathbb{P}}(Y_{t-1:t+1} \in A \mid W_{t-1:t+1} \in B(w_{t-1:t+1}, \varepsilon))$ for all w_{t-1}, w_t, w_{t+1} in W and $\varepsilon > 0$, the distribution of $Y_{t-1:t+1}$ is the same under under $\mathbb{P}(\cdot \mid W_{t-1:t+1} = w_{t-1:t+1})$ and $\widetilde{\mathbb{P}}(\cdot \mid W_{t-1:t+1} = w_{t-1:t+1})$. We may then extend the identifiability results for hidden Markov models given in [De Castro et al., 2017, Gassiat et al., 2015] to hidden Markov models with external signals following the same steps as in the proofs introduced in [Touron, 2019].

Let $(\phi_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions on $(\mathsf{Y}, \mathcal{Y})$ such that for all probability measures ν_1, ν_2 on $(\mathsf{Y}, \mathcal{Y})$, if for all $n \in \mathbb{N}$, $\int_{\mathsf{Y}} \phi_n d\nu_1 = \int_{\mathsf{Y}} \phi_n d\nu_2$, then $\nu_1 = \nu_2$. As Y is a Polish space, the existence of such a sequence $(\phi_n)_{n\in\mathbb{N}}$ is ensured. Let $k \in \mathsf{X}$, $m \geqslant 1$ and $a, b, c \in \{1, ..., m\}$. For all w_{t-1}, w_t, w_{t+1} in W, all $t \in \mathbb{Z}$, define the matrix $O_{t,w_t} \in \mathbb{R}^{m\times K}$ by

$$(O_{t,w_t})_{a,k} = \mathbb{E}[\phi_a(Y_t) \mid X_t = k, W_t = w_t] . \tag{2}$$

For all $a, b, c \in \{1, ..., m\}$, consider also:

$$\begin{aligned} L_{t,w_t}(a) &= \mathbb{E}[\phi_a(Y_t) \mid W_t = w_t] , \\ N_{t,w_{t-1:t}}(a, b) &= \mathbb{E}[\phi_a(Y_{t-1})\phi_b(Y_t) \mid W_{t-1:t} = w_{t-1:t}] , \\ P_{t,w_{t-1:t+1}}(a, c) &= \mathbb{E}[\phi_a(Y_{t-1})\phi_c(Y_{t+1}) \mid W_{t-1:t+1} = w_{t-1:t+1}] , \\ M_{t,w_{t-1:t+1}}(a, b, c) &= \mathbb{E}[\phi_a(Y_{t-1})\phi_b(Y_t)\phi_c(Y_{t+1}) \mid W_{t-1:t+1} = w_{t-1:t+1}] . \end{aligned}$$

For greater conciseness, the dependency on m of all these matrices is kept implicit. The first step of the proof is to write, for all w_{t-1}, w_t and w_{t+1} in W and $b \in \{1, \dots, m\}$, the known quantities L_{t,w_t}, $N_{t,w_{t-1:t}}$, $P_{t,w_{t-1:t+1}}$ and $M_{t,w_{t-1:t+1}}(., b, .)$ as functions of the quantities to be identified: π_{t,w_t}, $(O_{s,w_s})_{t-1\leqslant s\leqslant t+1}$, $Q_{t-1|w_{t-1},w_t}$, $Q_{t|w_t,w_{t+1}}$ and $(\nu_{s|w_s,x})_{t-1\leqslant s\leqslant t+1}$. For all $a \in \{1, \dots, M\}$, $w_t \in \mathsf{W}$,

$$L_{t,w_t}(a) = \sum_{x_t\in\mathsf{X}} \pi_{t,w_t}(x_t)\mathbb{E}[\phi_a(Y_t) \mid X_t = x_t, W_t = w_t] = \sum_{x_t\in\mathsf{X}} \pi_{t,w_t}(x_t)(O_{t,w_t})_{a,x_t} .$$

We obtain similarly, for all w_{t-1}, w_t and w_{t+1} in W,

$$L_{t,w_t} = O_{t,w_t}\pi_{t,w_t}^\top\,, \tag{3}$$

$$N_{t,w_{t-1:t}} = O_{t-1,w_{t-1}}\mathrm{diag}(\pi_{t-1,w_{t-1}})Q_{t-1|w_{t-1},w_t}O_{t,w_t}^\top\,, \tag{4}$$

$$P_{t,w_{t-1:t+1}} = O_{t-1,w_{t-1}}\mathrm{diag}(\pi_{t-1,w_{t-1}})Q_{t-1|w_{t-1},w_t}Q_{t|w_t,w_{t+1}}O_{t+1,w_{t+1}}^\top\,, \tag{5}$$

and for all $b \in \{1,\dots,m\}$,

$$M_{t,w_{t-1:t+1}}(.,b,.) = O_{t-1,w_{t-1}}\mathrm{diag}(\pi_{t-1,w_{t-1}})Q_{t-1|w_{t-1},w_t}\mathrm{diag}(O_{t,w_t}(b,.))Q_{t|w_t,w_{t+1}}O_{t+1,w_{t+1}}^\top\,. \tag{6}$$

The second step is to prove that O_{t,w_t} can be computed using the known quantities $L_{t-1,w_{t-1}}$, $N_{t,w_{t-1:t}}$, $P_{t,w_{t-1:t+1}}$ and $M_{t,w_{t-1:t+1}}$. Assumption H3 and the definition of the sequence $(\phi_n)_{n\in\mathbb{N}}$ yield that for all w_{t-1}, w_t and w_{t+1} in W, there exists $m_0 > K$, such that for all $m \geqslant m_0$, $O_{t-1,w_{t-1}}$, O_{t,w_t} and $O_{t+1,w_{t+1}}$ have full rank. Consider now that $m \geqslant m_0$. Under H3 and H4, $Q_{t-1|w_{t-1},w_t}$, $Q_{t|w_t,w_{t+1}}$ and $\mathrm{diag}(\pi_{w_{t-1}})$ are invertible. Then, using (5), it follows that the matrix $P_{t,w_{t-1:t+1}}$ has rank K. Write the singular value decomposition of the matrix $P_{t,w_{t-1:t+1}}$:

$$P_{t,w_{t-1:t+1}} = U\Sigma V^\top\,, \tag{7}$$

where U and V are matrices in $\mathbb{R}^{m\times K}$ containing the singular vectors associated with non-zero singular values of $P_{t,w_{t-1:t+1}}$ and Σ is a $K\times K$ diagonal matrix. As $P_{t,w_{t-1:t+1}}$ has rank K, the diagonal matrix $\Sigma = U^\top P_{t,w_{t-1:t+1}}V$ contains the K nonzero singular values of $P_{t,w_{t-1:t+1}}$. It is important to note that this decomposition is not unique as the order of the singular values is not fixed. For all $b \in \{1,...,m\}$ and for all w_{t-1}, w_t and w_{t+1} in W, define:

$$B_{t,w_{t-1:t+1}}(b) = (U^\top P_{t,w_{t-1:t+1}}V)^{-1}U^\top M_{t,w_{t-1:t+1}}(.,b,.)V\,. \tag{8}$$

Using (5) and (6), for all $b \in \{1,...,m\}$,

$$\begin{aligned}
B_{t,w_{t-1:t+1}}(b) &= (U^\top O_{t-1,w_{t-1}}\mathrm{diag}(\pi_{t-1,w_{t-1}})Q_{t-1|w_{t-1},w_t}Q_{t|w_t,w_{t+1}}O_{t+1,w_{t+1}}^\top V)^{-1}U^\top O_{t-1,w_{t-1}} \\
&\qquad\times \mathrm{diag}(\pi_{t-1,w_{t-1}})Q_{t-1|w_{t-1},w_t}\mathrm{diag}(O_{t,w_t}(b,.))Q_{t|w_t,w_{t+1}}O_{t+1,w_{t+1}}^\top V \\
&= (Q_{t|w_t,w_{t+1}}O_{t+1,w_{t+1}}^\top V)^{-1}\mathrm{diag}(O_{t,w_t}(b,.))Q_{t|w_t,w_{t+1}}O_{t+1,w_{t+1}}^\top V\,.
\end{aligned}$$

Defining $R = (Q_{t|w_t,w_{t+1}}O_{t+1,w_{t+1}}^\top V)^{-1}$, yields, for all $b \in \{1,...,m\}$,

$$\mathrm{diag}(O_{t,w_t}(b,.)) = R^{-1}B_{t,w_{t-1:t+1}}(b)R\,. \tag{9}$$

By (9), for all $(\alpha_1,...,\alpha_m) \in \mathbb{R}^m$, the eigenvalues of $\sum_{b=1}^m \alpha_b B_{t,w_{t-1:t+1}}(b)$ are the diagonal values of the matrix $\mathrm{diag}(\sum_{b=1}^m \alpha_b O_{t,w_t}(b,.)) = \mathrm{diag}(\alpha O_{t,w_t})$. Since O_{t,w_t} has rank K, there exist $\alpha \in \mathbb{R}^m$ such that $B_{t,w_{t-1:t+1}} = \sum_{b=1}^m \alpha_b B_{t,w_{t-1:t+1}}(b)$ has distinct eigenvalues. Therefore, the eigenvalue decomposition of $B_{t,w_{t-1:t+1}}$ is unique up to permutation and scaling of the columns of R. By computing the eigenvectors of $B_{t,w_{t-1:t+1}}$, we can finally compute R up to permutation and scaling of its columns. Therefore, the vectors $O_{t,w_t}(b,.)$ can be recovered for all $b \in \{1,...,m\}$ up to a common permutation of their components. As $M_{t,w_{t-1:t+1}}(.,b,.)$ and $P_{t,w_{t-1:t+1}}$ are the same when computed under $\mathbb{P}$ and $\tilde{\mathbb{P}}$,

$$(O_{t,w_t})_{.,x} = (\widetilde{O}_{t,w_t})_{.,\sigma_{t,w_{t-1:t+1}}(x)}$$

for some permutation $\sigma_{t,w_{t-1:t+1}}$. From its definition in (2), the matrix O_{t,w_t} does not depend of w_{t-1} and w_{t+1}. Consequently, $\sigma_{t,w_{t-1:t+1}}$ only depends on w_t and we write σ_{t,w_t} so that

$$(O_{t,w_t})_{\cdot,x} = (\widetilde{O}_{t,w_t})_{\cdot,\sigma_{t,w_t}(x)} \,. \tag{10}$$

We can similarly recover $O_{t-1,w_{t-1}}$ and $O_{t+1,w_{t+1}}$ up to permutations $\sigma_{t-1,w_{t-1}}$ and $\sigma_{t+1,w_{t+1}}$ by considering the conditional law of the triplets $Y_{t-2:t}$ and $Y_{t:t+2}$ given $W_{t-2:t}$ and $W_{t:t+2}$.

The last part of the proof is to show that the remaining quantities $\pi_{t-1,w_{t-1}}$, $Q_{t-1|w_{t-1},w_t}$ and $(\nu_{s|w_s,x})_{t-1\leqslant s\leqslant t+1}$ can also be identified up to the permutations $(\sigma_{s,w_s})_{t-1\leqslant s\leqslant t+1}$. For all $s \in \{t-1, t, t+1\}$, let U_s be a $K \times m$ matrix such that $U_s^\top O_{s,w_s}$ is invertible. Such a matrix exists as soon as O_{s,w_s} has full rank, which is the case since we assumed $m \geqslant m_0$. Using (3),

$$L_{t-1,w_{t-1}} = O_{t-1,w_{t-1}} \pi_{t-1,w_{t-1}}^\top$$

and

$$\pi_{t-1,w_{t-1}} = ((U_{t-1}^\top O_{t-1,w_{t-1}})^{-1} U_{t-1}^\top L_{t-1,w_{t-1}})^\top \,.$$

Given a permutation σ, we write Π_σ the associated permutation matrix, that is the matrix whose j-th column has a 1 in row $\sigma(j)$ and 0 elsewhere for all j. In particular, given a matrix A, the columns of $A\Pi_\sigma$ are the columns of A permuted according to σ. Since the matrix $L_{t-1,w_{t-1}}$ is the same under $\mathbb{P}$ and $\tilde{\mathbb{P}}$, Equation(10) yields

$$L_{t-1,w_{t-1}} = O_{t-1,w_{t-1}} \pi_{t-1,w_{t-1}}^\top = \widetilde{O}_{t-1,w_{t-1}} \widetilde{\pi}_{t-1,w_{t-1}}^\top = O_{t-1,w_{t-1}} \Pi_{\sigma_{t-1,w_{t-1}}} \widetilde{\pi}_{t-1,w_{t-1}}^\top \,,$$

which brings that $\widetilde{\pi}_{t-1,w_{t-1}} = \pi_{t-1,w_{t-1}} \Pi_{\sigma_{t-1,w_{t-1}}}$. Likewise,

$$N_{t,w_{t-1:t}} = O_{t-1,w_{t-1}} \mathrm{diag}(\pi_{t-1,w_{t-1}}) Q_{t-1|w_{t-1},w_t} O_{t,w_t}^\top \,,$$

which yields

$$Q_{t-1|w_{t-1},w_t} = \mathrm{diag}(\pi_{t-1,w_{t-1}})^{-1} (U_{t-1}^\top O_{t-1,w_{t-1}})^{-1} U_{t-1}^\top N_{t,w_{t-1:t}} U_t (O_{t,w_t}^\top U_t)^{-1} \,.$$

Moreover, since the matrix $N_{t-1,w_{t-1:t}}$ is the same when computed under $\mathbb{P}$ and $\tilde{\mathbb{P}}$ and noting that $\mathrm{diag}(\pi_{t-1,w_{t-1}} \Pi_{\sigma_{t-1,w_{t-1}}}) = \Pi_{\sigma_{t-1,w_{t-1}}}^\top \mathrm{diag}(\pi_{t-1,w_{t-1}})\, \Pi_{\sigma_{t-1,w_{t-1}}}$,

$$\begin{aligned}
N_{t,w_{t-1:t}} &= O_{t-1,w_{t-1}} \mathrm{diag}(\pi_{t-1,w_{t-1}}) Q_{t-1|w_{t-1},w_t} O_{t,w_t}^\top \\
&= \widetilde{O}_{t-1,w_{t-1}} \mathrm{diag}(\widetilde{\pi}_{t-1,w_{t-1}}) \widetilde{Q}_{t-1|w_{t-1},w_t} \widetilde{O}_{t,w_t}^\top \\
&= O_{t-1,w_{t-1}} \Pi_{\sigma_{t-1,w_{t-1}}} \mathrm{diag}(\pi_{t-1,w_{t-1}} \Pi_{\sigma_{t-1,w_{t-1}}}) \widetilde{Q}_{t-1|w_{t-1},w_t} (O_{t,w_t} \Pi_{\sigma_{t,w_t}})^\top \,, \\
&= O_{t-1,w_{t-1}} \mathrm{diag}(\pi_{t-1,w_{t-1}}) \Pi_{\sigma_{t-1,w_{t-1}}} \widetilde{Q}_{t-1|w_{t-1},w_t} \Pi_{\sigma_{t,w_t}}^\top O_{t,w_t}^\top \,,
\end{aligned}$$

which gives $\widetilde{Q}_{t-1|w_{t-1},w_t} = \Pi_{\sigma_{t-1,w_{t-1}}}^\top Q_{t-1|w_{t-1},w_t} \Pi_{\sigma_{t,w_t}}$.

Therefore, under H1-4, if for all t, all w_{t-1}, w_t, w_{t+1} in W, the distribution of $Y_{t-1:t+1}$ given $W_{t-1:t+1} = w_{t-1:t+1}$ is the same under two sets of parameters, then for all $x, x' \in \mathsf{X}$,

$$\begin{aligned}
\pi_{t,w_t}(x) &= \widetilde{\pi}_{t-1,w_{t-1}}(\sigma_{t-1,w_{t-1}}(x)) \,, \\
Q_{t-1|w_{t-1},w_t}(x,x') &= \widetilde{Q}_{t-1|w_{t-1},w_t}(\sigma_{t-1,w_{t-1}}(x), \sigma_{t,w_t}(x')) \,, \\
(O_{t,w_t})_{\cdot,x} &= (\widetilde{O}_{t,w_t})_{\cdot,\sigma_{t,w_t}(x)} \,.
\end{aligned}$$

The last equality provides that for every $m \geqslant m_0$, $O_{t,w_t} = \widetilde{O}_{t,w_t}\Pi_{\sigma_{t,w_t}}$. By definition of the sequence $(\phi_n)_{n\in\mathbb{N}}$, this implies that for all $w_t \in \mathsf{W}$, for all $k \in \mathsf{X}$:

$$\nu_{t|w_t,k} = \widetilde{\nu}_{t|w_t,\sigma_{t,w_t}(k)}$$

and this result concludes the proof. □

3 Consistency

We consider hidden Markov models with external variables as defined in Section 2 and satisfying Assumptions H1-4. We assume that for all $t \in \mathbb{Z}$, w_t, w_{t+1} in W, and $k \in \mathsf{X}$, the transition matrices and emission distributions of the model are entirely parameterized by some $\theta \in \Theta$, where Θ is a closed parameter space, and written $Q^\theta_{t|w_t,w_{t+1}}$ and $\nu^\theta_{t|w_t,k}$. In addition, we introduce the following assumptions.

H5 The process $(Y_t, W_t)_{t\in\mathbb{Z}}$ is stationary and ergodic.

Note that we do not assume the model parameters $Q^\theta_{t|w_t,w_{t+1}}$ and $\nu^\theta_{t|w_t,k}$ to be the same for all $t \in \mathbb{Z}$. Most works on maximum likelihood estimation assume that the distribution of the observations belongs to the proposed parametric family of distributions. In many cases, it is unlikely that this assumption is satisfied. In this section, we only assume that H5 holds but we do not assume that $(Y_t)_{t\in\mathbb{Z}}$ has a distribution satisfying Assumptions H1-4. Consequently, we introduce $\mathbb{P}^*$ the true distribution of $\{(Y_t, W_t)\}_{t\in\mathbb{Z}}$ and $\mathbb{E}^*$, the expectation under this distribution.

H6
a) There exists a probability measure λ on $(\mathsf{Y}, \mathcal{Y})$ such that for all $x \in \mathsf{X}$, all $w \in \mathsf{W}$, all $t \in \mathbb{Z}$ and all $\theta \in \Theta$, $\nu^\theta_{t|w,x}$ has a density with respect to λ denoted by $f^\theta_{t|w,x}$.

b) There exists $\sigma_- > 0$ such that, for all $x, x' \in \mathsf{X}$, all $(w, w') \in \mathsf{W}^2$, all $t \in \mathbb{Z}$ and all $\theta \in \Theta$, $Q^\theta_{t|w,w'}(x, x') > \sigma_-$.

c) For all $t \in \mathbb{Z}$, $y \in \mathsf{Y}$, all $w \in \mathsf{W}$ and all $\theta \in \Theta$: $0 < \sum_{x\in\mathsf{X}} f^\theta_{t|w,x}(y) < \infty$.

H7 $b^+ := \sup_{\theta,t} \sup_{x_t,y_t,w_t} f^\theta_{t|w_t,x_t}(y_t) < \infty$ and for all $t \in \mathbb{Z}$, $\mathbb{E}^*[\,|\ln(b^-(t, Y_t, W_t)\,|\,] < \infty$, where $b^-(t, y_t, w_t) := \inf_\theta \sum_{x_t\in\mathsf{X}} f^\theta_{t|w_t,x_t}(y_t)$.

Consider also the following family of probability distributions on X:

$$\mathcal{D} = \{\pi \text{ probability distribution on } \mathsf{X}\,;\, \forall x \in \mathsf{X}, \pi(x) > \sigma_-\}\ .$$

For any initial distribution $\pi \in \mathcal{D}$, any $\theta \in \Theta$ and any $r, s \in \mathbb{Z}$ such that $r < s$, let $L^\theta_{\pi,r:s}$ be the conditional likelihood function of the $s - r + 1$ first observations of the hidden Markov model with external variables associated with initial distribution π at time r and parameter θ:

$$L^\theta_{\pi,r:s}(y_{r:s} \mid w_{r:s}) = \sum_{x_{r:s}\in\mathsf{X}^{r-s+1}} \pi(x_r) f^\theta_{r|w_r,x_r}(y_r) \prod_{p=r+1}^{s} Q^\theta_{p-1|w_{p-1},w_p}(x_{p-1}, x_p) f^\theta_{p|w_p,x_p}(y_p)\,.$$

Under H1-4, and following the demonstration introduced in [Douc et al., 2014, Chapter 13], it is possible to establish the strong consistency of the maximum likelihood estimator conditionally to the external variables defined as

$$\hat{\theta}_{n,\pi,W_{0:n-1}} \in \operatorname*{argmax}_{\theta\in\Theta} L^\theta_{\pi,0:n-1}(Y_{0:n-1} \mid W_{0:n-1})\,. \tag{11}$$

Since $L^\theta_{\pi,r:s}(Y_{r:s} \mid W_{r:s})$ is the likelihood of a hidden Markov model, the loglikelihood of the observations $Y_{0:n-1}$ conditionally to the external signals $W_{0:n-1}$, denoted by $\ell_{\pi,n}(\theta) = \ln L^\theta_{\pi,0:n-1}(Y_{0:n-1} \mid W_{0:n-1})$, can be decomposed as follows using H1:

$$\begin{aligned}\ell_{\pi,n}(\theta) = \ln L^\theta_{\pi,0:n-1}(Y_{0:n-1} \mid W_{0:n-1}) &= \sum_{t=0}^{n-1} \ln L^\theta_{\pi,0:t}(Y_t \mid Y_{0:t-1}, W_{0:n-1}) \\ &= \sum_{t=0}^{n-1} \ln L^\theta_{\pi,0:t}(Y_t \mid Y_{0:t-1}, W_{0:t}) \,, \end{aligned} \tag{12}$$

with the convention $L^\theta_{\pi,0:0}(Y_0 \mid Y_{0:-1}, W_0) = L^\theta_{\pi,0}(Y_0 \mid W_0) = \sum_{x_0 \in \mathsf{X}} \pi(x_0) f^\theta_{0|W_0,x_0}(Y_0)$. We first show that the limit $\lim_{m\to\infty} L^\theta_{\pi,-m:t}(Y_t \mid Y_{-m:t-1}, W_{-m:t})$ exists $\mathbb{P}^*$-a.s. and does not depend on π. Writing $L^\theta(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t})$ this limit, it forms an ergodic stationary sequence and we introduce the following approximation of $\ell_{\pi,n}(\theta)$:

$$\ell^s_n(\theta) = \sum_{t=0}^{n-1} \ln L^\theta(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t}) \,, \tag{13}$$

where the superscript s stands for stationary. Using Birkoff's theorem with this new sequence, we can prove that there is a constant $\ell(\theta)$ such that $\lim_{n\to\infty} n^{-1} \ln L^\theta_\pi(Y_{0:n-1} \mid W_{0:n-1}) = \ell(\theta)$, $\mathbb{P}^*$-a.s.. The last step of the proof amounts to establishing that $\lim_{n\to\infty} d(\hat\theta_{n,\pi,w_{0:n-1}}, \Theta^*) = 0$ with $\Theta^* = \operatorname{argmax}_{\theta\in\Theta} \ell(\theta)$.

Proposition 1.1. *Under Assumptions H5-H7, for all $(y_t, w_t)_{t\in\mathbb{Z}}$, $\pi \in \mathcal{D}$, $(L^\theta_{\pi,-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}))_{m\geqslant 0}$ has a finite limit, which does not depend on the initial distribution π, denoted by $L^\theta(y_t \mid y_{-\infty:t-1}, w_{-\infty:t})$. Moreover, the limit $\lim_{n\to\infty} n^{-1} \sum_{t=0}^{n-1} \ln L^\theta(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t})$ exists $\mathbb{P}^*$-a.s. and*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \sup_{\pi\in\mathcal{D}} n^{-1} \mid \ell_{\pi,n}(\theta) - \ell^s_n(\theta) \mid = 0 \,, \quad \mathbb{P}^*\text{-a.s.} \,, \tag{14}$$

where $\ell_{\pi,n}(\theta) = \sum_{t=0}^{n-1} \ln L^\theta_{\pi,0:t}(Y_t \mid Y_{0:t-1}, W_{0:t})$ and $\ell^s_n(\theta) = \sum_{t=0}^{n-1} \ln L^\theta(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t})$.

Proof. The proof follows the same steps as the proof of [Douc et al., 2014, Proposition 13.5] and is given in Supplementary material B.1. □

The last part of the proof is to use the Birkhoff ergodic theorem so as to conclude the strong consistency of the conditional MLE $\hat\theta_{n,\pi,w_{0:n-1}}$.

H8 for all $t \in \mathbb{Z}$, all $(x_{t-1}, x_t) \in \mathsf{X}^2$, all $(w_{t-1}, w_t) \in \mathsf{W}^2$ and all $y_t \in \mathsf{Y}$, the functions $\theta \mapsto Q^\theta_{t-1|w_{t-1},w_t}(x_{t-1}, x_t)$ and $\theta \mapsto f^\theta_{t|w_t,x_t}(y_t)$ are continuous.

Theorem 2. *Under H5-H8, For any sequence of estimators $(\hat\pi_n)_n$ taking values in $\mathcal{D}$, $\mathbb{P}^*$-a.s.,*

$$\lim_{n\to\infty} d(\hat\theta_{n,\hat\pi_n,w_{0:n-1}}, \Theta^*) = 0 \,, \tag{15}$$

with $\Theta^* = \operatorname*{argmax}_{\theta\in\Theta} \ell(\theta)$ and $\ell(\theta) = \mathbb{E}^*[\ln L^\theta(Y_0 \mid Y_{-\infty:-1}, W_{-\infty:0})]$.

Proof. The proof follows the same steps as the proofs of [Douc et al., 2014, Theorems 13.7 and 8.42] and is given in Supplementary material B.2. □

4 Experiments

We propose to use an Expectation Maximization algorithm to learn the parameters of the proposed models, see [Dempster et al., 1977]. Let $(X_t, Y_t, W_t)_{t\geqslant 1}$ be a HMM with external signals and K hidden states. An EM-based algorithm can be derived to maximize the loglikelihood function $\theta \mapsto \log p_\theta(Y_{1:N} \mid W_{1:N})$ for some $N \geqslant 1$.

Given a current parameter estimate θ_{k-1}, the pivotal idea of this algorithm is to replace the loglikelihood of the observations by the surrogate quantity:

$$\theta \mapsto \mathcal{Q}(\theta, \theta_{k-1}) = \mathbb{E}_{\theta_{k-1}} \left[\log p_\theta(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}) \mid \mathbf{Y}, \mathbf{W}\right] ,$$

with $\mathbf{X} = (X_t)_{1\leqslant t\leqslant N}$, $\mathbf{Y} = (Y_t)_{1\leqslant t\leqslant N}$, $\mathbf{W} = (W_t)_{1\leqslant t\leqslant N}$ and $p_\theta(\mathbf{X}, \mathbf{Y} \mid \mathbf{W})$ the joint density of $(\mathbf{X}, \mathbf{Y})$ conditionally to the external signal $\mathbf{W}$. The new parameter estimate is then obtained following the two steps:

(i) compute $\theta \mapsto Q(\theta, \theta_{k-1})$;

(ii) set θ_k as one of the maximizers of $\theta \mapsto Q(\theta, \theta_{k-1})$.

As the latent states take values in $\{1, \dots, K\}$, the conditional distribution of $\mathbf{X}$ given $(\mathbf{Y}, \mathbf{W})$ can be computed explicitly using the Baum-Welch forward-backward algorithm, see for instance [Douc et al., 2014]. Therefore, step (i) can be performed as it is. For step (ii), as a maximizer of $\theta \mapsto Q(\theta, \theta_{k-1})$ is not always straightforward to compute, the generalized EM (GEM) approach [Dempster et al., 1977] is used. Given θ_{k-1}, an optimizer is used to find a θ_k verifying $Q(\theta_k, \theta_{k-1}) \geqslant Q(\theta_{k-1}, \theta_{k-1})$. This less restrictive variation, despite a potential slowdown, still ensures the convergence of the EM algorithm.

4.1 Simulated data

Assume first that $K = 2$ and consider the following hidden Markov models.

Hidden Markov Model (*hmm*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}$ with $Q_{i1} = \exp(P_{i1})/(1 + \exp(P_{i1}))$ and $P_{i1} = \omega_{i1} \in \mathbb{R}$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ is Gaussian with mean $\mu_k \in \mathbb{R}$ and variance $\sigma_k^2 > 0$.

Seasonal Hidden Markov Model (*shmm*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}(t)$ with $Q_{i1}(t) = \exp(P_{i1}(t))/(1 + \exp(P_{i1}(t)))$ and $P_{i1}(t) = \omega_{i1} + \omega_{i3}\cos(2\pi t/T) + \omega_{i4}\sin(2\pi t/T)$, with ω_{i1}, ω_{i3}, ω_{i4} in $\mathbb{R}$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ is Gaussian with mean $\mu_k(t) = \delta_{k1} + \delta_{k3}\cos(2\pi t/T) + \delta_{k4}\sin(2\pi t/T)$, with δ_{k1}, δ_{k3}, δ_{k4} in $\mathbb{R}$, and variance $\sigma_k^2 > 0$.

Hidden Markov Model with External Signals (*hmm-es*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}(W_t)$ with $Q_{i1}(W_t) = \exp(P_{i1}(W_t))/(1 + \exp(P_{i1}(W_t)))$ and $P_{i1}(W_t) = \omega_{i1} + \omega_{i2}W_t$, with ω_{i1}, ω_{i2} in $\mathbb{R}$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ and W_t is Gaussian with mean $\mu_k(W_t) = \delta_{k1} + \delta_{k2}W_t$, with δ_{k1}, δ_{k2} in $\mathbb{R}$, and variance $\sigma_k^2 > 0$.

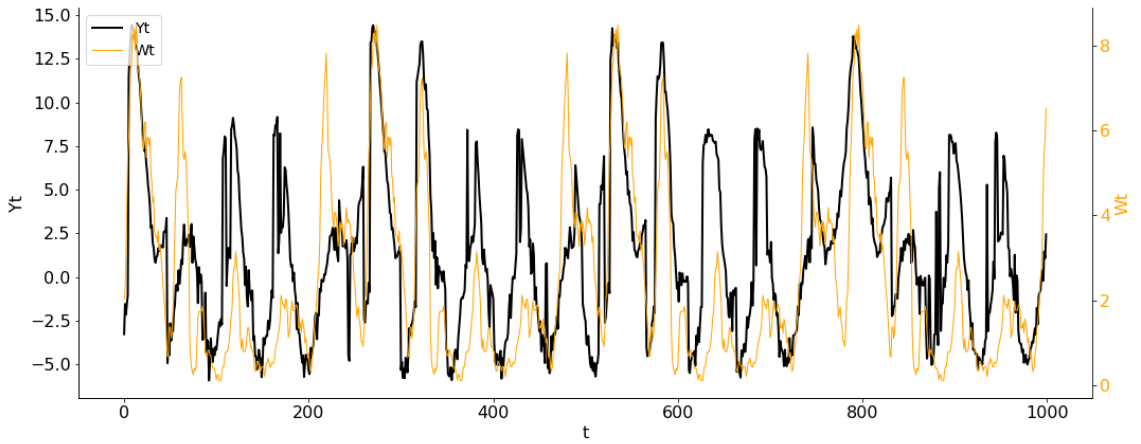

Figure 2: Example of external signal sequence (orange) and associated simulated sequence using a *shmm-es* model with parameter θ^* (black).

Seasonal Hidden Markov Model with External Signals (*shmm-es*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i, W_t) = Q_{ij}(t, W_t)$ with $Q_{i1}(t, W_t) = \exp(P_{i1}(t, W_t))/(1+\exp(P_{i1}(t, W_t)))$ and $P_{i1}(t, W_t) = \omega_{i1} + \omega_{i2} W_t + \omega_{i3} \cos(2\pi t/T) + \omega_{i4} \sin(2\pi t/T)$, with ω_{i1}, ω_{i2}, ω_{i3}, ω_{i4} in $\mathbb{R}$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ and W_t is Gaussian with mean $\mu_k(t, W_t) = \delta_{k1} + \delta_{k2} W_t \delta_{k3} \cos(2\pi t/T) + \delta_{k4} \sin(2\pi t/T)$, with δ_{k1}, δ_{k2}, δ_{k3}, δ_{k4} in $\mathbb{R}$, and variance $\sigma_k^2 > 0 > 0$.

A first simulated time series is generated using a hidden Markov model with external signals and seasonal components as defined above (*shmm-es*). As external signal $(W_t)_{t \geqslant 1}$, a signal of the fashion dataset introduced in [David et al., 2022] is used to provide a realistic setting. The external sequence is smoothed using a moving average with a sliding window of length 8 and divided by the mean of the first year to rescale the signal. In addition, the smoothed signal is duplicated to simulate an arbitrary long sequence $(Y_t)_{t \geqslant 1}$. Figure 2 displays the resulting external signal duplicated 4 times to reach the length of 1000 time steps. We define a set of parameters θ^* of a *shmm-es* with $T = 52$:

$$
\begin{aligned}
\pi^* &= \begin{pmatrix} \pi_1^* & 1 - \pi_1^* \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix} \\
\delta^* &= \begin{pmatrix} \delta_{11}^* & \delta_{12}^* & \delta_{13}^* & \delta_{14}^* \\ \delta_{21}^* & \delta_{22}^* & \delta_{23}^* & \delta_{24}^* \end{pmatrix} = \begin{pmatrix} 3. & 0.8 & 2.5 & 4. \\ -1.1 & -0.1 & -1.5 & 3.5 \end{pmatrix} \\
\sigma^* &= \begin{pmatrix} \sigma_1^* & \sigma_2^* \end{pmatrix} = \begin{pmatrix} 0.5 & 0.25 \end{pmatrix} \\
\omega^* &= \begin{pmatrix} \omega_{11}^* & \omega_{12}^* & \omega_{13}^* & \omega_{14}^* \\ \omega_{21}^* & \omega_{22}^* & \omega_{23}^* & \omega_{24}^* \end{pmatrix} = \begin{pmatrix} 0.5 & 0.9 & 0.7 & 0.5 \\ -2. & -0.2 & -0.6 & 0.7 \end{pmatrix}.
\end{aligned}
$$

We set $T = 52$ according to the weekly seasonality of the external signal. Using this set of parameters and the external sequence $(W_t)_{t \geqslant 1}$, a sequence of $(X_t)_{t \geqslant 1}$ and $(Y_t)_{t \geqslant 1}$ is generated, see Figure 2. Using the Expectation Maximization algorithm, 10 *hmm*, *shmm*, *hmm-es* and *shmm-es* are fitted on a training set of length 10000 with different initial parameters $\theta_0 = (\omega^0, \delta^0, \sigma^0)$. For each initial parameter estimate, the EM algorithm is run for 1000 iterations. A test set of length 250 is generated and used to evaluate the different approaches. As we provide a single simulated sequence, the parameter π is not learned and fixed at the

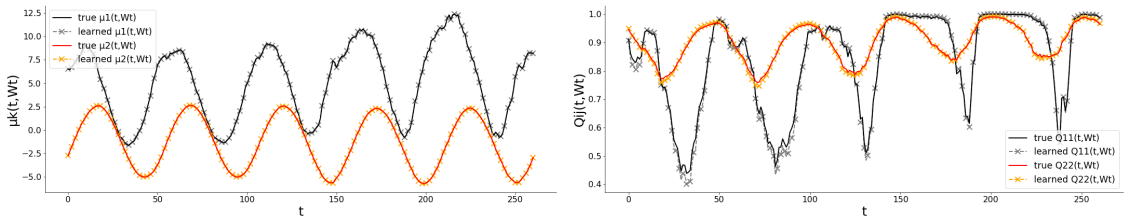

Figure 3: Comparison of the true functions $\mu_1(t, Wt)$ and $\mu_2(t, Wt)$ and the estimations with the *shmm-es* model (left). Comparison of the true functions $Q_1(t, Wt)$ and $Q_2(t, Wt)$ and the estimations with the *shmm-es* model (right). For the functions $\mu_1(t, Wt)$ and $\mu_2(t, Wt)$, estimations are almost perfect: the true functions and the learned ones are combined.

true parameter $\pi*$ during the training. In order to reproduce results and trainings on simulated sequences, a complete code in Python is publicly provided[1].

We first evaluate the performance of the EM algorithm to estimate the true set of parameters θ^*. Figure 3 displays the true functions $t \mapsto \mu_1(t, Wt)$, $t \mapsto \mu_2(t, Wt)$, $t \mapsto Q_{11}(t, Wt)$ and $t \mapsto Q_{22}(t, Wt)$ and the estimations by the *shmm-es*. A complete overview of the final learned parameters by the EM-based algorithm for the *shmm-es* model can be found in Table 4. Additional experiments to analyze the impact of the sequence length and the initial parameters are also summarized in Appendix A.1. Secondly, the forecasting accuracy of *hmm*, *shmm*, *hmm-es* and *shmm-es* is evaluated. For each trained model. A set of 1000 predictions of the test set is generated, the average prediction is computed and evaluated using the mean absolute error (MAE), the mean squared errror (MSE) and the mean absolute scaled error (MASE):

$$\text{MAE} = \frac{1}{h}\sum_{j=1}^{h} \mid Y_{N+j} - \hat{Y}_{N+j} \mid \qquad \text{MSE} = \frac{1}{h}\sum_{j=1}^{h}(Y_{N+j} - \hat{Y}_{N+j})^2$$

$$\text{MASE} = \frac{N-T}{h}\frac{\sum_{j=1}^{h} \mid Y_{N+j} - \hat{Y}_{N+j} \mid}{\sum_{i=1}^{N-T} \mid Y_i - Y_{i-T} \mid},$$

where N represents the train set length, h the horizon and T the seasonality length. Table 1 summarizes the mean and the standard deviation over the 10 repetitions for the 4 approaches and the 3 metrics. Finally, Figure 4 shows predictions of each method on the test set. This first experiment illustrates two main results. Firstly, as the true set of parameters was correctly recovered for the *shmm-es* model, the EM algorithm is efficient to estimate parameters of models with external signals. Secondly, Table 1 shows that *hmm-es* and *shmm-es* methods are able to leverage the external signal and outperform their concurrent methods *hmm* and *shmm*.

4.2 Fashion time series forecasting

An interesting application of HMM with external signals can be found in the fashion and retail industries. It is crucial for these domains to accurately forecast the consumers future behaviours in order to make optimal inventory decisions and avoid massive wastes. However, fashion dynamics appear to be really volatile with nonlinear changes of dynamics resulting from the apparition of new tendencies. By taking into account behaviour of influencers as external signals, it becomes possible to better anticipate these changes.

[1] `https://github.com/etidav/hmm_with_external_signals`

Table 1: MASE, MSE and MAE accuracy of the 4 hmm approaches considered using a synthetic time series. For each method, 10 trainings are done with different initialisation parameter. The mean and the standard deviation over the 10 iteration is displayed for each approach

	MASE		**MAE**		**MSE**	
	mean	*std*	*mean*	*std*	*mean*	*std*
hmm	1.354	0.016	4.833	0.056	31.124	0.368
shmm	0.903	0.005	3.222	0.017	15.582	0.193
hmm-es	1.245	0.008	4.446	0.027	26.346	0.327
shmm-es	**0.737**	0.008	**2.630**	0.029	**14.102**	0.346

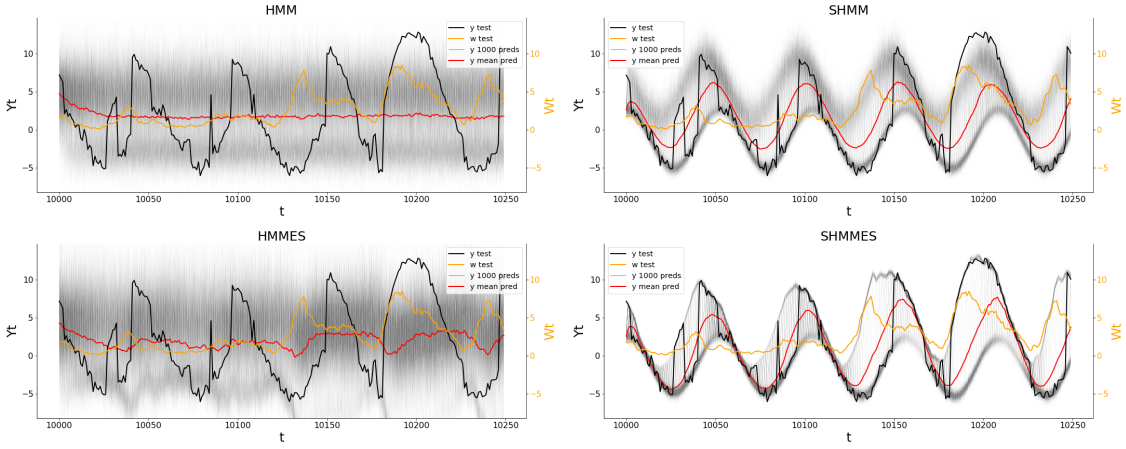

Figure 4: 1000 simulation and average prediction of *hmm*, *shmm*, *hmm-es* and *shmm-es* models on the test set.

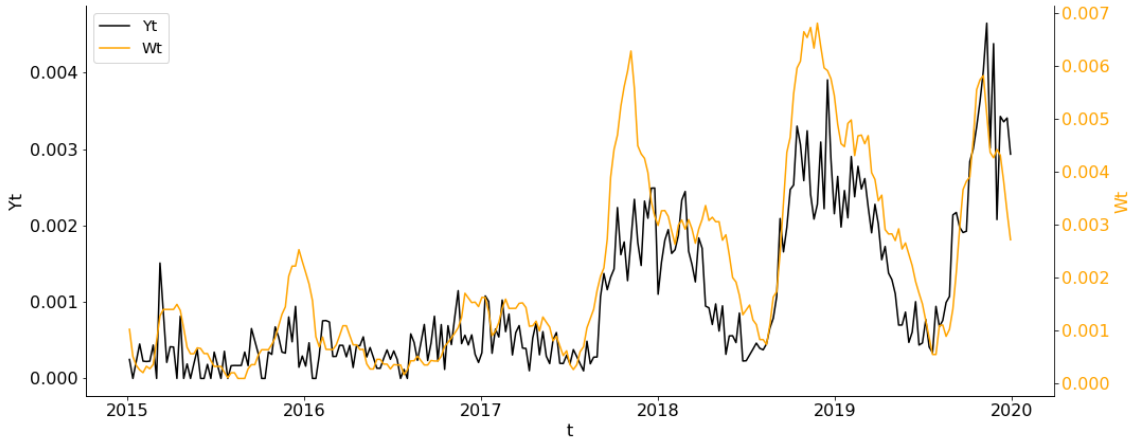

Figure 5: Time series "eu-female-top-325" from [David et al., 2022] representing an emerging fashion trend on social media with its linked influencers external signal. The influencers sequence is smoothed using a moving average with a sliding window of length 8.

4.2.1 Application to a single time series

A sequence [1] from the fashion dataset introduced in [David et al., 2022] was selected as it shows a sudden change of level, seasonality and noise intensity as illustrated in Figure 5. Several models are trained using this time series. *hmm*, *shmm*, *hmm-es* and *shmm-es* described in Section 4.1 are considered in this first application. We also consider an autoregressive HMM (*ar-hmm*), an autoregressive HMM with seasonal components (*ar-shmm*), an autoregressive HMM with external signals (*ar-hmm-es*) and an autoregressive HMM with seasonal components and external signals (*ar-shmm-es*). A complete description of these additional models can be found in Appendix A.3.2. To fairly compare and evaluate HMM-based methods, several benchmarks are also evaluated. Four statistical benchmarks are proposed: *snaive*, *thetam*, *tbats*, *ets*. Complete descriptions and references for these models can be found in [Hyndman et al., 2015]. A recurrent neural network (RNN) model [Hochreiter and Schmidhuber, 1997] denoted *lstm* is also considered. Finally the hybrid model *hermes* introduced in [David et al., 2022] is added in the pool of benchmarks. Combining the strengths of statistical approaches and RNNs, this model achieved impressive results on the fashion dataset and demonstrated the benefit of the inclusion of influencers signal. As all the previous benchmarks do not include the external signal, variations of *lstm* and *hermes* using external signal and called *lstm-es* and *hermes-es* are also considered.

For HMM-based models and statistical benchmarks, a complete code is publicly available[2]. Concerning the training of the HMM approaches, additional information can be found in Appendix A.4. For *lstm*, *lstm-es*, *hermes* and *hermes-es*, pre-trained models on the fashion dataset introduced in [David et al., 2022] are directly used without retraining. We evaluate our candidates on a 1-year forecasting task. The 52 points of 2020 are hidden during the training procedure and used to evaluate the accuracy of the models. For each model, a forecast of the test set is computed and the three metrics MAE, MSE and MASE are used to evaluate the prediction. In addition, for HMM-based models, ten independent training procedures are performed

[1] This sequence is referred to as "eu-female-top-325".

[2] `https://github.com/etidav/hmm_with_external_signals`

Table 2: MASE, MSE and MAE accuracy on the fashion time series eu-female-top-325. Bold values provide the best performance for the benchmarks and for the HMMs-based approaches.

	MASE		**MAE**(10^{-3})		**MSE**(10^{-6})	
	mean	*std*	*mean*	*std*	*mean*	*std*
thetam	1.73	-	0.87	-	1.04	-
ets	1.59	-	0.80	-	0.89	-
tbats	1.25	-	0.63	-	0.68	-
snaive	1.09	-	0.55	-	0.51	-
lstm-es	0.97	0.20	0.49	0.10	0.50	0.19
lstm	0.78	0.11	0.39	0.06	0.28	0.09
hermes	0.70	0.05	0.35	0.02	0.23	0.02
hermes-es	**0.67**	0.04	**0.34**	0.02	**0.22**	0.01
hmm	1.99	0.01	1.01	0.01	1.78	0.03
shmm	1.95	0.01	0.99	0.01	1.61	0.02
hmm-es	0.98	0.07	0.60	0.04	0.52	0.06
ar-hmm	0.95	0.01	0.58	0.01	0.62	0.01
ar-hmm-es	0.80	0.13	0.49	0.08	0.40	0.09
ar-shmm	0.77	0.07	0.47	0.04	0.43	0.09
shmm-es	0.62	0.04	0.38	0.02	0.24	0.02
ar-shmm-es	**0.56**	0.08	**0.35**	0.05	**0.24**	0.07

and standard variations over the 10 replications are provided for the three metrics. Results are summarized in Table 2. In this challenging framework, the 4 statistical benchmarks do not reach the same performance as models with external signals. The same remark can be made with the *hmm* and *shmm* approaches. However, the inclusion of the external signal considerably improves the performance of the HMM-based model and they achieve a level of accuracy comparable to RNN-based models *hermes-es* and *lstm-es* trained on the whole fashion dataset and using external signal. Predictions of the best HMM model *ar-shmm-es*, the statistical method *tbats* and the two state-of-the-art models *hermes-es* and *lstm-es* are displayed in Figures 6-7.

4.2.2 Application on a sample of time series

In this section, the different approaches are compared using 10 sequences from the fashion dataset introduced in [David et al., 2022]. Name and additional materials concerning these trends can be found in Appendix A.3. For each time series, the same models and training process as in Section 4.2.1 are considered. Table 3 summarizes the results on each sequence in terms of MASE. The inclusion of the external signal always largely improves the accuracy of the HMM models. *ar-hmm-es* method reaches the highest level of accuracy over the 10 time series followed by *hmm-es*, *lstm-es*,*hermes-es* and *ar-shmm-es*.

- On the first fashion sequence, HMMs including the influencers signal leveraged the influencers external signals resulting in a significant improvement of the accuracy compared to the benchmark models.
- Conversely, on the seventh fashion sequence, as the first increase of the external signal led to a decrease of the main sequence in the train set, HMM models using the influencers signal showed diffi-

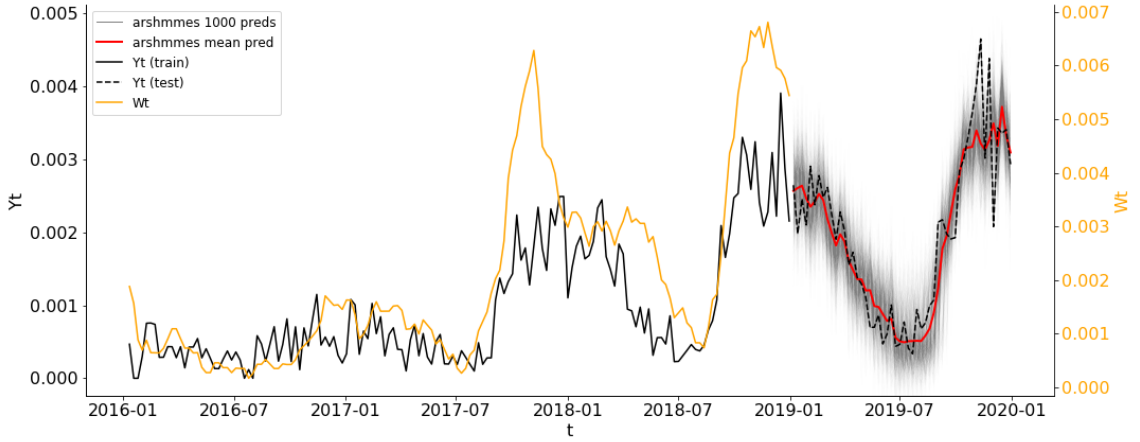

Figure 6: *ar-shmm-es* predictions of the last year of the time series eu-female-top-325. 1000 simulation are calculated and displayed in grey and for each point, the mean over the 1000 predictions is displayed in red.

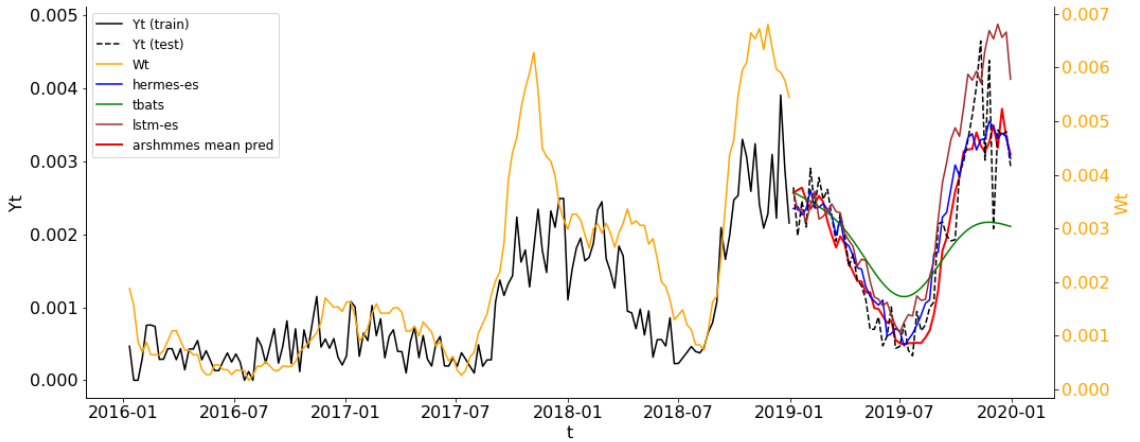

Figure 7: *tbats*, *hermes-es*, *lstm-es* and *ar-shmm-es* predictions on the last year of the time series eu-female-top-325. for *ar-shmm-es*, 1000 predictions are calculated and for each point, the mean (in red) is displayed.

Table 3: MASE of the benchmarks and HMM models on the 10 fashion time series. For the RNN-based and HMM-based approaches, as 10 training were done, standard variation of the final MASE over the 10 replications is also provided. Bold values provide the best performance for the benchmarks and for the HMMs-based approaches.

	ts 1	**ts 2**	**ts 3**	**ts 4**	**ts 5**	**ts 6**	**ts 7**	**ts 8**	**ts 9**	**ts 10**	**total**
snaive	5.91	1.25	1.36	**0.46**	1.09	0.98	2.44	0.73	0.87	0.45	1.55
thetam	4.93	0.76	**0.63**	1.15	1.73	0.84	**0.86**	1.33	**0.38**	0.57	1 .32
tbats	5.04	0.80	0.73	0.61	1.25	0.65	1.40	0.83	0.51	0.32	1.21
ets	4.90	**0.45**	0.69	0.64	1.59	0.62	1.33	0.79	0.48	0.44	1.19
lstm	5.35 ± 0.53	0.69 ± 0.20	0.92 ± 0.20	0.83 ± 0.22	0.78 ± 0.11	0.70 ± 0.05	1.54 ± 0.13	1.22 ± 0.42	0.59 ± 0.05	0.31 ± 0.02	1.29
hermes	5.50 ± 0.12	0.54 ± 0.06	0.85 ± 0.15	0.55 ± 0.06	0.70 ± 0.05	0.75 ± 0.03	1.98 ± 0.13	0.73 ± 0.05	0.60 ± 0.03	0.28 ± 0.01	1.25
hermes-es	4.70 ± 0.38	0.68 ± 0.27	0.76 ± 0.11	0.65 ± 0.11	**0.67** ± 0.04	0.61 ± 0.05	1.64 ± 0.23	**0.70** ± 0.07	0.59 ± 0.03	**0.27** ± 0.01	1.13
lstm-es	**4.18** ± 0.54	0.80 ± 0.21	0.88 ± 0.17	0.77 ± 0.30	0.97 ± 0.20	**0.54** ± 0.08	1.18 ± 0.18	0.87 ± 0.22	0.51 ± 0.03	0.29 ± 0.02	1.10
shmm	6.80 ± 0.44	0.56 ± 0.02	1.05 ± 0.27	0.95 ± 0.01	1.95 ± 0.01	0.46 ± 0.07	2.69 ± 0.01	0.84 ± 0.01	0.62 ± 0.01	**0.36** ± 0.04	1.63
hmm	5.48 ± 0.01	0.83 ± 0.02	0.75 ± 0.02	1.09 ± 0.01	1.99 ± 0.01	0.65 ± 0.01	2.61 ± 0.01	1.32 ± 0.01	0.75 ± 0.13	0.65 ± 0.01	1.61
ar-shmm	5.54 ± 0.23	**0.31** ± 0.01	0.80 ± 0.15	0.42 ± 0.01	0.77 ± 0.07	0.99 ± 0.02	2.60 ± 0.04	0.75 ± 0.01	0.84 ± 0.08	0.81 ± 0.05	1.38
ar-hmm	5.55 ± 0.01	0.48 ± 0.05	**0.54** ± 0.01	0.60 ± 0.01	0.95 ± 0.01	1.22 ± 0.04	**1.84** ± 0.81	**0.64** ± 0.08	0.79 ± 0.16	0.69 ± 0.01	1.33
shmm-es	4.67 ± 0.99	0.36 ± 0.03	0.66 ± 0.06	0.43 ± 0.04	0.62 ± 0.04	1.11 ± 0.13	2.89 ± 0.66	0.81 ± 0.06	0.58 ± 0.07	0.59 ± 0.15	1.27
ar-shmm-es	4.57 ± 0.53	0.46 ± 0.22	0.64 ± 0.07	0.46 ± 0.02	**0.56** ± 0.08	0.62 ± 0.04	2.72 ± 0.59	0.79 ± 0.05	**0.51** ± 0.06	0.52 ± 0.15	1.18
hmm-es	**2.89** ± 0.34	0.52 ± 0.01	**0.54** ± 0.02	0.42 ± 0.02	0.98 ± 0.07	**0.43** ± 0.12	2.84 ± 0.26	0.92 ± 0.01	0.53 ± 0.06	0.61 ± 0.01	1.07
ar-hmm-es	3.04 ± 0.50	0.41 ± 0.05	0.64 ± 0.16	**0.40** ± 0.01	0.80 ± 0.13	0.81 ± 0.03	2.21 ± 0.74	0.90 ± 0.02	0.54 ± 0.04	0.62 ± 0.01	**1.04**

culties to leverage the external signal. Consequently, even simple statistical models like *ets* or *tbats* outperformed the HMM-based models on this specific sequence.

- As the 10 fashion sequences are short and some of them do not have a strong seasonal component, seasonal variations of HMMs did not reach the best global level of accuracy.

Finally, over the 10 time series, HMMs including the external signal show the same high level of accuracy than benchmarck models like *hermes-es* and *lstm-es* while these two benchmarks have several thousand of parameters, have been train on the whole fashion dataset gathering 10000 time series and include the external signal. It reveals that on these 10 specific sequences, the new HMM approach is better suited and better leverages the influencers signal while maintaining theoretical properties and interpretability. Multiples figures displaying predictions of the different models on the fashion sequences can be found in Appendix A.5.

5 Conclusion

The motivation of this paper is to establish theoretical guarantees for a family of latent data models including external signals in the transition matrices and the emissions laws. We showed that under several assumptions, the identifiability and convergence results known in the HMM literature can be extended to these models. In addition of the theoretical guarantees, numerous experiments were done on simulated data and real world time series. Several HMMs including external signal were tested on a long-term forecasting task and compared to statistical and RNN-based alternatives. Final results highlighted that including external signals in HMMs allows to learn meaningful dependencies and improve forecasting performance. On some real world sequences, they even outperformed state-of-the-art models.

However, as a HMM with external signals has to be trained for each new sequences, our approach remains computationally expensive to train. Consequently, a future work will be to design a HMM framework able to be trained on a large dataset, to learn complex shared dynamics and finally leverage high dimensional external signals.

References

[Andrieu and Doucet, 2003] Andrieu, C. and Doucet, A. (2003). Online Expectation-Maximization type algorithms for parameter estimation in general state space models. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 6, pages VI–69. IEEE.

[Baum and Petrie, 1966] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.

[Bengio and Frasconi, 1994] Bengio, Y. and Frasconi, P. (1994). An Input Output HMM architecture. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

[Cappé and Moulines, 2009] Cappé, O. and Moulines, É. (2009). Online Expectation-Maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

[Chopin et al., 2020] Chopin, N., Papaspiliopoulos, O., et al. (2020). *An introduction to sequential Monte Carlo*. Springer.

[David et al., 2022] David, É., Bellot, J., and Le Corff, S. (2022). HERMES: Hybrid error-corrector model with inclusion of external signals for nonstationary fashion time series. preprint.

[De Castro et al., 2017] De Castro, Y., Gassiat, É., and Le Corff, S. (2017). Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Transactions on Information Theory*, 63(8):4758 – 47777.

[Dempster et al., 1977] 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*, 39:1–38.

[Douc et al., 2011] Douc, R., Moulines, É., Olsson, J., and Van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *the Annals of Statistics*, 39(1):474–513.

[Douc et al., 2004] Douc, R., Moulines, É., and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, 32(5):2254 – 2304.

[Douc et al., 2014] Douc, R., Moulines, É., and Stoffer, D. (2014). *Nonlinear time series: theory, methods and applications with R examples*. Chapman and Hall/CRC.

[Gassiat et al., 2015] Gassiat, É., Cleynen, A., and Robin, S. (2015). Inference in finite state space non parametric hidden Markov models and applications. *Statistics and Computing*, 26(1-2):np.

[Gassiat et al., 2020] Gassiat, É., Le Corff, S., and Lehéricy, L. (2020). Identifiability and consistent estimation of nonparametric translation hidden Markov models with general state space. *Journal of Machine Learning Research*, 21(115):1–40.

[Gassiat and Rousseau, 2016] Gassiat, É. and Rousseau, J. (2016). Nonparametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193 – 212.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

[Hsu et al., 2012] Hsu, D., Kakade, S. M., and Zhang, T. (2012). A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480.

[Hyndman et al., 2015] Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., and Razbash, S. (2015). Forecasting functions for time series and linear models. *R package version*, 6.

[Juang and Rabiner, 1985] Juang, B.-H. and Rabiner, L. (1985). Mixture autoregressive hidden Markov models for speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1404–1413.

[Le Corff and Fort, 2013] Le Corff, S. and Fort, G. (2013). Online Expectation-Maximization based algorithms for inference in hidden Markov models. *Electronic Journal of Statistics*, 7:763–792.

[Leroux, 1992] Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1):127–143.

[Li et al., 2019] Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., and Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of Transformer on time series forecasting. *Advances in neural information processing systems*, 32.

[Lim et al., 2021] Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

[Radenen and Artieres, 2012] Radenen, M. and Artieres, T. (2012). Contextual hidden Markov models. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2113–2116. IEEE.

[Salinas et al., 2020] Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191.

[Särkkä, 2013] Särkkä, S. (2013). *Bayesian filtering and smoothing*. Cambridge university press.

[Touron, 2019] Touron, A. (2019). Consistency of the maximum likelihood estimator in seasonal hidden markov models. *Statistics and Computing*, 29(5):1055–1075.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Table 4: *shmm-es* final parameters depending of the train set size. the final values are displayed and percentages of errors compared to the true parameters are computed.

	true parameter	**length=1000**		**length=10000**		**length=100000**	
		mean	*% error*	*mean*	*% error*	*mean*	*% error*
$\delta 11$	3.0	3.0	$<1\%$	2.99	$<1\%$	3.0	$<1\%$
$\delta 12$	0.8	0.8	$<1\%$	0.8	$<1\%$	0.8	$<1\%$
$\delta 13$	2.5	2.5	$<1\%$	2.49	$<1\%$	2.5	$<1\%$
$\delta 14$	4.0	3.98	$<1\%$	4.0	$<1\%$	4.0	$<1\%$
$\delta 21$	-1.1	-1.12	1%	-1.07	3%	-1.1	$<1\%$
$\delta 22$	-0.1	-0.09	8%	-0.11	8%	-0.1	1%
$\delta 23$	-1.5	-1.54	3%	-1.51	1%	-1.5	$<1\%$
$\delta 24$	3.5	3.57	2%	3.49	$<1\%$	3.49	$<1\%$
$\sigma 1$	0.25	0.24	4%	0.25	1%	0.25	$<1\%$
$\sigma 2$	0.5	0.5	1%	0.51	1%	0.5	$<1\%$
$\omega 11$	0.5	0.52	3%	0.32	35%	0.46	8%
$\omega 12$	0.9	0.91	1%	1.02	13%	0.92	2%
$\omega 13$	0.7	0.49	30%	0.66	6%	0.7	$<1\%$
$\omega 14$	0.5	0.51	3%	0.53	6%	0.56	12%
$\omega 21$	-2.0	-2.48	24%	-2.0	$<1\%$	-2.0	$<1\%$
$\omega 22$	-0.2	-0.01	95%	-0.21	7%	-0.19	5%
$\omega 23$	-0.6	-0.8	34%	-0.63	6%	-0.62	4%
$\omega 24$	0.7	0.88	26%	0.79	12%	0.72	2%

A Additional numerical results

A.1 *shmm-es*-generated time series

In addition to Figure 3, Table 4 provides a complete description of the final parameters recovered by the *shmm-es* depending on the length of the sequence used during training. In each case, final parameters values and percentages of errors compared to the real parameters are displayed. Consider a parameter x^* and $\hat{x}$ its estimate, we call percentage of errors the following quantity: $100 \times (\mid x^* - \hat{x} \mid /x^*)$. In each scenario $(\delta_{11}, \delta_{12}, ..., \delta_{22})$ and $(\sigma_{11}, \sigma_{22})$ are efficiently recovered. Some parameters of the transition matrices, even in the case where a sequence of 100000 time steps is used, are not perfectly learned. However, using larger training sequences considerably improves the estimation for most of them.

As the EM algorithm is strongly impacted by the initialisation of the parameters, a second experiment is done so as to evaluate the impact of the initialisation on the *shmm-es* learning. Three trainings are run with a sequence of length 10000. For the first one, an initialisation of θ_0 is sampled with Gaussian distributions with mean $(\omega^*, \delta^*, \sigma^*)$ and standard deviations equal to 0.5. For the second one, the Gaussian standard deviations are set to 1 and for the last one, increased to 2. Table 5 displays final parameters values recovered by the *shmm-es* in the 3 scenarios as well as a percentage of error defined above. In all scenarios, the EM algorithm converges and accurately retrieves the true set of parameters of the *shmm-es* model.

Table 5: *shmm-es* final parameters depending of the initialization method. The first method use Gaussian distribution centred at the true parameter values with the standard deviation set at 0.5. For the second method, the standard deviation is increased to 1 and for the last one, increased to 2. A simulated sequence with length equal to 10000 is used to train a *shmm-es* model for each initialization method. With the 3 resulting set of parameters, the final values are displayed and percentages of errors compared to the true set of parameters are computed.

	true parameter	**Init. std=0.5**		**Init. std=1**		**Init. std=2**	
		mean	*% error*	*mean*	*% error*	*mean*	*% error*
$\delta 11$	3.0	2.99	$<1\%$	2.99	$<1\%$	2.99	$<1\%$
$\delta 12$	0.8	0.8	$<1\%$	0.8	$<1\%$	0.8	$<1\%$
$\delta 13$	2.5	2.49	$<1\%$	2.49	$<1\%$	2.49	$<1\%$
$\delta 14$	4.0	4.0	$<1\%$	4.0	$<1\%$	4.0	$<1\%$
$\delta 21$	-1.1	-1.07	3%	-1.07	3%	-1.07	3%
$\delta 22$	-0.1	-0.11	8%	-0.11	7%	-0.11	7%
$\delta 23$	-1.5	-1.51	1%	-1.51	1%	-1.51	1%
$\delta 24$	3.5	3.49	$<1\%$	3.49	$<1\%$	3.49	$<1\%$
$\sigma 1$	0.25	0.25	1%	0.25	1%	0.25	1%
$\sigma 2$	0.5	0.51	1%	0.5	$<1\%$	0.5	$<1\%$
$\omega 11$	0.5	0.32	35%	0.32	35%	0.32	35%
$\omega 12$	0.9	1.02	13%	1.02	13%	1.02	13%
$\omega 13$	0.7	0.66	6%	0.66	6%	0.66	6%
$\omega 14$	0.5	0.53	6%	0.53	6%	0.53	6%
$\omega 21$	-2.0	-2.0	$<1\%$	-2.0	$<1\%$	-2.0	$<1\%$
$\omega 22$	-0.2	-0.21	7%	-0.21	7%	-0.21	7%
$\omega 23$	-0.6	-0.63	6%	-0.64	6%	-0.64	6%
$\omega 24$	0.7	0.79	12%	0.79	12%	0.79	12%

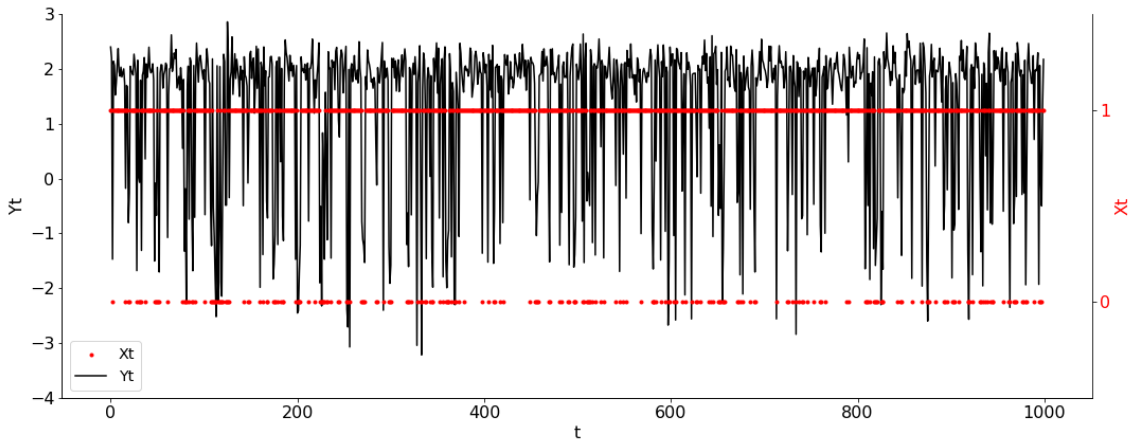

Figure 8: Simulated hidden sequence $(X_t)_{t\geqslant 1}$ (red) and simulated main sequence $(Y_t)_{t\geqslant 1}$ (black) using a HMM model with parameter θ^*.

A.2 hmm-generated time series

A second simulated time series is generated using a simple HMM as defined in Section 4.1. The true set of parameters of the model θ^* is:

$$\pi^* = \begin{pmatrix}\pi_1^* & 1-\pi_1^*\end{pmatrix} = \begin{pmatrix}0 & 1\end{pmatrix}$$
$$\delta^* = \begin{pmatrix}\delta_{11}^* \\ \delta_{21}^*\end{pmatrix} = \begin{pmatrix}-1. \\ 2.\end{pmatrix}$$
$$\sigma^* = \begin{pmatrix}\sigma_1^* & \sigma_2^*\end{pmatrix} = \begin{pmatrix}1 & 0.25\end{pmatrix}$$
$$\omega^* = \begin{pmatrix}\omega_{11}^* \\ \omega_{21}^*\end{pmatrix} = \begin{pmatrix}-0.8 \\ -1.4\end{pmatrix}.$$

Using this set of parameters, a sequence of $(X_t)_{t\geqslant 1}$ and $(Y_t)_{t\geqslant 1}$ of length 10000 is generated and a sample of length 1000 is displayed in Figure 8. Using the Expectation Maximization algorithm and with the same protocol as in Section 4.1, a *hmm*, *shmm*, *hmm-es* and *shmm-es* are trained. In order to fit the *hmm-es* and the *shmm-es* approach, the generated external signal $(W_t)_{t\leqslant 1}$ displayed in Figure 2 is used. With this experiment, we evaluate the ability of HMM with external signals to recover the true set of parameters in the case where the external signal $(W_t)_{t\geqslant 1}$ is not involved in the HMM dynamics. Table 6 shows the mean of the final parameters over the 4 trains and the percentage of error associated. The true set of parameters is well recovered by the four candidates. As the time series is generated with a simple HMM, the two models using the external signal learned parameters near 0 for the external signal dependencies and seasonal variations did not learn artificial seasonal effects.

Table 6: *hmm*, *shmm*, *hmm-es* and *shmm-es* final parameters. In each case, a training is done using a simulated sequence of length 10000. The resulting set of parameters and the percentages of errors as defined in Appendix A.1 are displayed. Rows of parameters representing external signal dependencies are highlighted.

	true parameter	*hmm*		*shmm*		*hmm-es*		*shmm-es*	
		mean	*% error*	*mean*	*% error*	*mean*	*% error*	*mean*	*% error*
$\delta 11$	-1.0	-1.01	1%	-1.01	1%	-1.01	1%	-1.0	<1%
$\delta 12$	-	-	-	-	-	-0.0	-	-0.0	-
$\delta 13$	-	-	-	-0.06	-	-	-	-0.06	-
$\delta 14$	-	-	-	0.02	-	-	-	0.02	-
$\delta 21$	2.0	2.01	<1%	2.01	<1%	2.01	1%	2.01	1%
$\delta 22$	-	-	-	-	-	-0.0	-	-0.0	-
$\delta 23$	-	-	-	-0.0	-	-	-	-0.0	-
$\delta 24$	-	-	-	0.0	-	-	-	0.0	-
$\omega 11$	-0.8	-0.87	8%	-0.87	8%	-0.83	4%	-0.82	3%
$\omega 12$	-	-	-	-	-	-0.01	-	-0.02	-
$\omega 13$	-	-	-	0.03	-	-	-	0.02	-
$\omega 14$	-	-	-	-0.04	-	-	-	-0.05	-
$\omega 21$	-1.4	-1.42	2%	-1.42	2%	-1.44	3%	-1.44	3%
$\omega 22$	-	-	-	-	-	0.01	-	0.01	-
$\omega 23$	-	-	-	0.01	-	-	-	0.02	-
$\omega 24$	-	-	-	0.03	-	-	-	0.03	-
$\sigma 1$	1.0	1.0	<1%	1.0	<1%	1.0	<1%	1.0	<1%
$\sigma 2$	0.25	0.25	<1%	0.25	<1%	0.25	<1%	0.25	<1%

A.3 Fashion time series forecasting

A.3.1 Fashion sequences

In Section 4.2.2, ten sequences from the fashion dataset were selected. [3] These time series were not totally randomly selected but for the fact that they display various dynamics including abrupt changes of behaviours, which are difficult to forecast. A smoothing is applied using a moving average with a sliding window of length 8 to the external signal associated with each time series. Figure 9 displays the 10 sequences and their external signals.

A.3.2 Model descriptions

In this section, a complete description of the HMM approaches introduced in Section 4.2 is given.

Hidden Markov Model (*hmm*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}$ with $Q_{i1} = \exp(P_{i1})/(1 + \exp(P_{i1}))$ and $P_{i1} = \omega_{i1} \in \mathbb{R}$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ is Gaussian with mean $\mu_k \in \mathbb{R}$ and variance σ_k^2.

Seasonal Hidden Markov Model (*shmm*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}(t)$ with $Q_{i1}(t) = \exp(P_{i1}(t))/(1 + \exp(P_{i1}(t)))$ and $P_{i1}(t) =$

[3] They are respectively named br-female-shoes-262, br-female-texture-59, br-female-texture-82, eu-female-outerwear-177, eu-female-top-325, eu-female-top-394, eu-female-texture-80, us-female-outerwear-171, us-female-shoes-76, and us-female-top-79.

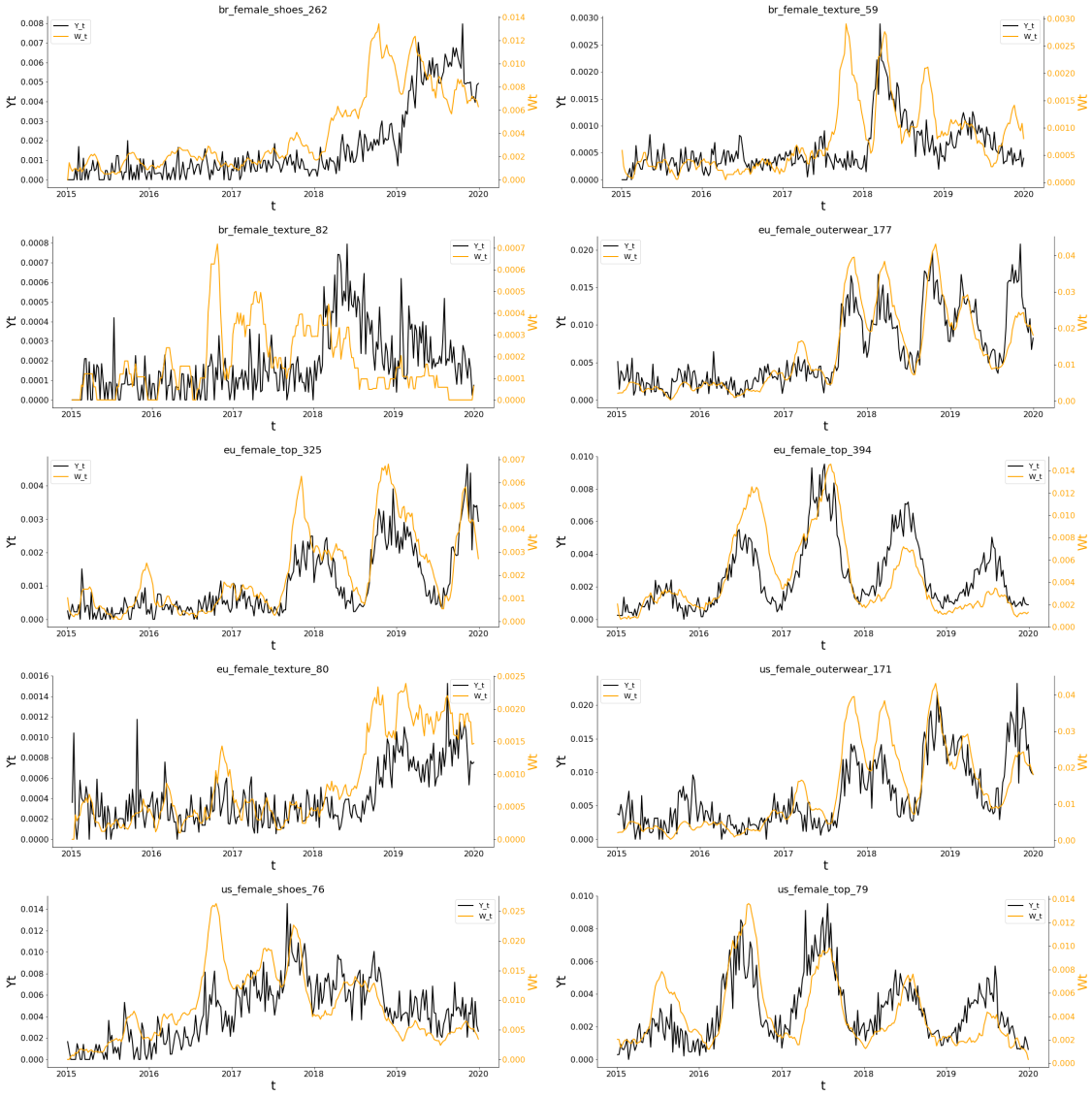

Figure 9: 10 fashion time series (black) and their associated external signals (orange).

$\omega_{i1} + \omega_{i3}\cos(2\pi t/T) + \omega_{i4}\sin(2\pi t/T)$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ is Gaussian with mean $\mu_k(t) = \delta_{k1} + \delta_{k4}\cos(2\pi t/T) + \delta_{k5}\sin(2\pi t/T)$ and variance σ_k^2.

Auto Regressive Hidden Markov Model (*ar-hmm*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}$ with $Q_{i1} = \exp(P_{i1})/(1 + \exp(P_{i1}))$ and $P_{i1}(Y_{t-52}) = \omega_{i1} + \omega_{i2}Y_{t-52}$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ and Y_{t-52} is Gaussian with mean $\mu_k(Y_{t-52}) = \delta_{k1} + \delta_{k2}Y_{t-52}$ and variance σ_k^2.

Auto Regressive Seasonal Hidden Markov Model (*ar-shmm*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}(t)$ with $Q_{i1}(t) = \exp(P_{i1}(t))/(1 + \exp(P_{i1}(t)))$ and $P_{i1}(t) = \omega_{i1} + \omega_{i2}\cos(2\pi t/T) + \omega_{i3}\sin(2\pi t/T)$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ and Y_{t-52} is Gaussian with mean $\mu_k(t, Y_{t-52}) = \delta_{k1} + \delta_{k2}Y_{t-52} + \delta_{k4}\cos(2\pi t/T) + \delta_{k5}\sin(2\pi t/T)$ and variance σ_k^2.

Hidden Markov Model with External Signals (*hmm-es*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}(W_{t-52})$ with $Q_{i1}(W_{t-52}) = \exp(P_{i1}(W_{t-52})$ $)/(1 + \exp(P_{i1}(W_{t-52})))$ and $P_{i1}(W_{t-52}) = \omega_{i1} + \omega_{i2}W_{t-52}$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ and W_{t-52} is Gaussian with mean $\mu_k(W_{t-52}) = \delta_{k1} + \delta_{k3}W_{t-52}$ and variance σ_k^2.

Seasonal Hidden Markov Model with External Signals (*shmm-es*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i, W_{t-52}) = Q_{ij}(t, W_{t-52})$ with $Q_{i1}(t, W_{t-52}) = \exp(P_{i1}(t, W_{t-52}))/(1 + \exp(P_{i1}(t, W_{t-52})))$ and $P_{i1}(t, W_{t-52}) = \omega_{i1} + \omega_{i2}W_{t-52} + \omega_{i3}\cos(2\pi t/T) + \omega_{i4}\sin(2\pi t/T)$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$ and W_{t-52} is Gaussian with mean $\mu_k(t, W_{t-52}) = \delta_{k1} + \delta_{k3}W_{t-52} + \delta_{k4}\cos(2\pi t/T) + \delta_{k5}\sin(2\pi t/T)$ and variance σ_k^2.

Auto Regressive Hidden Markov Model with External Signals (*ar-hmm-es*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}(W_{t-52})$ with $Q_{i1}(W_{t-52}) = \exp(P_{i1}(W_{t-52}))/(1+\exp(P_{i1}(W_{t-52})))$ and $P_{i1}(W_{t-52}) = \omega_{i1}+\omega_{i2}W_{t-52}$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$, Y_{t-52} and W_{t-52} is Gaussian with mean $\mu_k(Y_{t-52}, W_{t-52}) = \delta_{k1} + \delta_{k2}Y_{t-52} + \delta_{k3}W_{t-52}$ and variance σ_k^2.

Auto Regressive Seasonal Hidden Markov Model with External Signals (*ar-shmm-es*). For all $i, j, k \in \{1, 2\}$, $\mathbb{P}(X_1 = k) = \pi_k \in (0, 1)$ and for $t \geqslant 1$, $\mathbb{P}(X_{t+1} = j \mid X_t = i) = Q_{ij}(t, W_{t-52})$ with $Q_{i1}(t, W_{t-52}) = \exp(P_{i1}(t, W_{t-52}))/(1 + \exp(P_{i1}(t, W_{t-52})))$ and $P_{i1}(t) = \omega_{i1} + \omega_{i2}W_{t-52} + \omega_{i3}\cos(2\pi t/T) + \omega_{i4}\sin(2\pi t/T)$. For all $t \geqslant 1$, the conditional distribution of Y_t given $\{X_t = k\}$, Y_{t-52} and W_{t-52} is Gaussian with mean $\mu_k(t, Y_{t-52}, W_{t-52}) = \delta_{k1} + \delta_{k2}Y_{t-52} + \delta_{k3}W_{t-52} + \delta_{k4}\cos(2\pi t/T) + \delta_{k5}\sin(2\pi t/T)$ and variance σ_k^2.

In a real world situation, the external signal, depending on influencers, is not known in advance. Consequently, a lag of one year (here 52 time steps) is introduced i.e. the distribution of the HMM at time t depends on W_{t-52}. The same lag is used in the Auto-regressive HMM as almost all the fashion sequences show a strong yearly seasonality.

A.4 HMM-based model training

We propose a complete overview of the training process used in Section 4.2 for the HMM approaches. For each HMM-based model, given a fashion time series, the following estimation procedure is used.

1. Parameter θ_0 is randomly initialized.
2. A GEM is run for 10 iterations.
3. Using the resulting parameter, 10 predictions of the last year of the train set are computed and evaluated using a MSE.
4. The average MSE over the 10 forecasts is computed
5. Steps 1-4 are repeated 30 times and the best run is saved based on the average MSE computed in step 4.
6. Starting with the initial parameter of the best run, 500 iterations of the EM algorithm are run and the final parameter $\hat{\theta}$ is saved.

The complete code is developed in Python and Tensorflow and available at `https://github.com/etidav/hmm_with_external_signals`.

A.5 HMM-based model predictions

In this last section, in addition to Table 3, predictions of the best models on the 10 considered fashion time series are displayed in Figure 10.

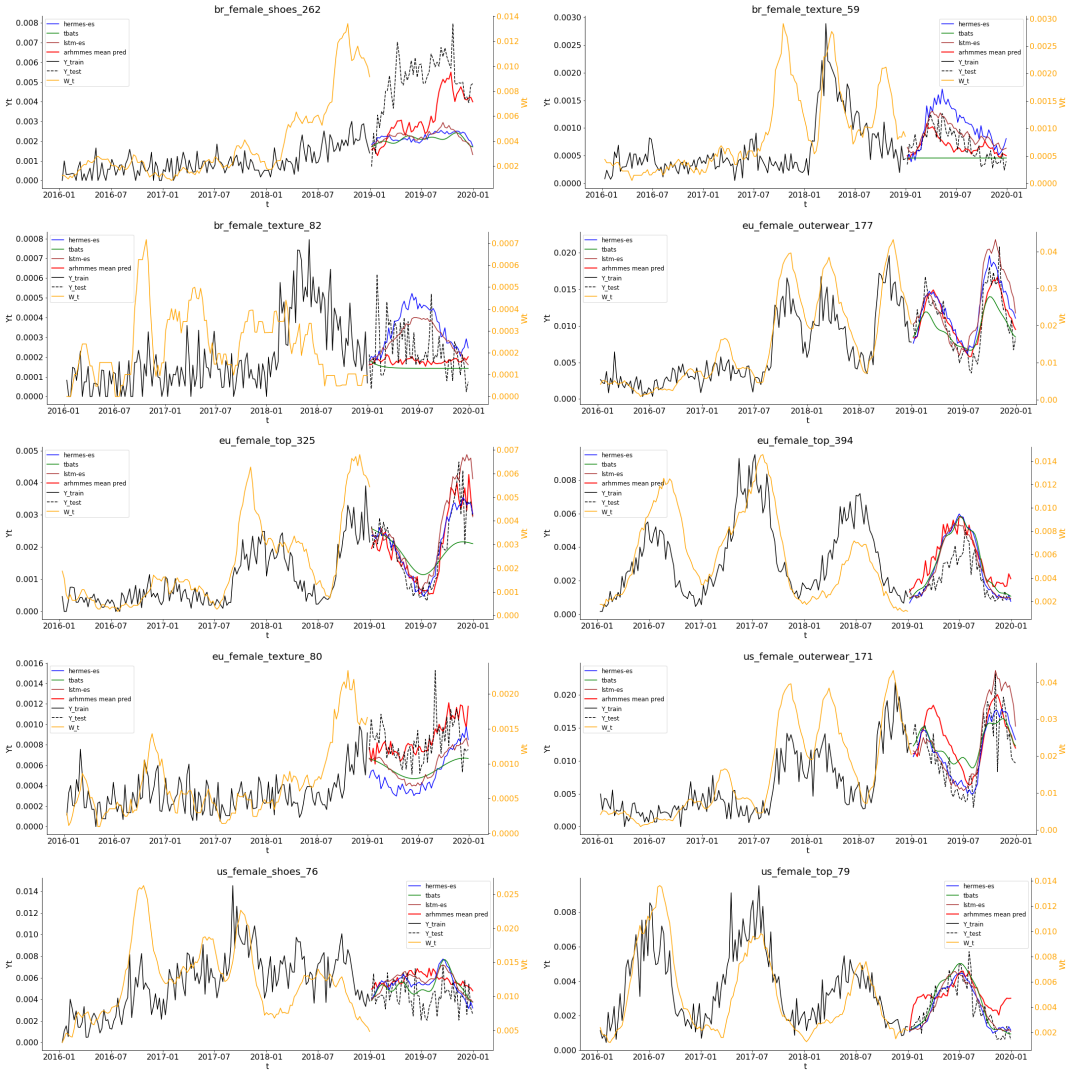

Figure 10: *tbats*, *hermes-es*, *lstm-es* and *hmmes* predictions on the last year of the 10 fashion time series. for *hmmes*, 1000 predictions are calculated and for each point, the mean is displayed

Supplementary material

B Additional proofs

Additional notations. For any finite signed measure λ on $(\mathsf{X}, \mathcal{X})$, $\mid \lambda \mid$ refers to the total variation and $\mid \lambda \mid_{\mathrm{TV}} = \mid \lambda \mid (\mathsf{X})$ denotes the total variation norm. We denote by $d_{\mathrm{TV}}(.)$ the total variation distance and recall that for every signed measures λ and λ' on $(\mathsf{X}, \mathcal{X})$, $d_{\mathrm{TV}}(\lambda, \lambda') = \mid \lambda - \lambda' \mid_{\mathrm{TV}} /2$. Finally, for every Markov kernel M on $(\mathsf{X}, \mathcal{X})$, we denote by $\Delta_{TV}(M)$ the Dobrushin coefficient of M: $\Delta_{TV}(M) = \sup_{x,x' \in \mathsf{X}} \mid M(x, \cdot) - M(x', \cdot) \mid_{\mathrm{TV}} /2$. The proofs given here follow closely [Douc et al., 2014] and are given for completeness.

B.1 Proof of Proposition 1.1

Lemma 2.2 provides that for a fixed $t \in \mathbb{Z}$ and for all $\pi \in \mathcal{D}$, the sequence $(L^\theta_{\pi,-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}))_{m \geqslant 0}$ is a uniform Cauchy sequence. Consequently, it has a finite limit and Equation (28) provides that this limit does not depend of the initial distribution π. We write $L^\theta(y_t \mid y_{-\infty:t-1}, w_{-\infty:t})$ this limit. Using Equation (27), for all $\theta \in \Theta$, for all $\pi \in \mathcal{D}$, for all $t \geqslant 1$ and $m \geqslant 0$,

$$\ln(\sigma_- b^-(t, y_t, w_t)) \leqslant \ln L^\theta_{\pi,-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) \leqslant \ln(b^+) \, .$$

Therefore, $\mid \ln L^\theta_{\pi,-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) \mid \leqslant \mid \ln(\sigma_- b^-(t, y_t, w_t)) \mid \vee \mid \ln(b^+) \mid$. Taking the limit as m grows to infinity,

$$\mid \ln L^\theta(y_t \mid y_{-\infty:t-1}, w_{-\infty:t}) \mid \leqslant \mid \ln(\sigma_- b^-(t, y_t, w_t)) \mid \vee \mid \ln(b^+) \mid , \tag{16}$$

and, under Assumption H7,

$$\mathbb{E}^*[\mid \ln L^\theta(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t}) \mid] < \infty \, . \tag{17}$$

Thus, as we assumed the process $(Y_t, W_t)_{t \in \mathbb{Z}}$ stationary and ergodic with H5, Birkhoff's ergodic theorem can be used and establishes that $n^{-1} \sum_{t=0}^{n-1} \ln L^\theta(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t})$ exists $\mathbb{P}^*$-a.s. Finally, using Equation (29) and letting k grow to infinity yields for all $t \geqslant 1$ and all $\pi \in \mathcal{D}$,

$$\sup_{\theta \in \Theta} \mid \ln L^\theta_\pi(y_t \mid y_{0:t-1}, w_{0:t}) - \ln L^\theta(y_t \mid y_{-\infty:t-1}, w_{-\infty:t}) \mid \leqslant \frac{\rho^{t-1}}{\sigma_-} \, .$$

In addition, for $t = 0$ and $m > 0$,

$$\mid L^\theta_{\pi,0}(y_0 \mid w_0) - L^\theta_{\pi,-m:0}(y_0 \mid y_{-m:-1}, w_{-m:0}) \mid \leqslant 2 \sum_{x_0 \in \mathsf{X}} f^\theta_{0|w_0,x_0}(y_0) \, .$$

On the other hand, $L^\theta_{\pi,-m:0}(y_0 \mid y_{-m:-1}, w_{-m:0}) \geqslant \sigma_- \sum_{x_0 \in \mathsf{X}} f^\theta_{0|w_0,x_0}(y_0)$ and noting that $\pi \in \mathcal{D}$, $L^\theta_{\pi,0}(y_0 \mid w_0) \geqslant \sigma_- \sum_{x_0 \in \mathsf{X}} f^\theta_{0|w_0,x_0}(y_0)$. Therefore, since $\mid \ln a - \ln b \mid \leqslant \mid a - b \mid /(a \wedge b)$,

$$\mid \ln L^\theta_{\pi,0}(y_0 \mid w_0) - \ln L^\theta_{\pi,-m:0}(y_0 \mid y_{-m:-1}, w_{-m:0}) \mid \leqslant \frac{2}{\sigma_-} \, ,$$

which yields $\mid \ln L^{\theta}_{\pi,0}(y_0 \mid w_0) - \ln L^{\theta}(y_0 \mid y_{-\infty:-1}, w_{-\infty:t}) \mid \leqslant 2/\sigma_-$. Consequently, for all $\theta \in \Theta$ and all $\pi \in \mathcal{D}$:

$$
\begin{aligned}
n^{-1} \mid \ell_{\pi,n}(\theta) - \ell^s_n(\theta) \mid &= n^{-1} \left| \sum_{t=0}^{n-1} \ln L^{\theta}_{\pi,0:t}(Y_t \mid Y_{0:t-1}, W_{0:t}) - \sum_{t=0}^{n-1} \ln L^{\theta}(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t}) \right| \\
&= n^{-1} \left| \sum_{t=0}^{n-1} \ln L^{\theta}_{\pi,0:t}(Y_t \mid Y_{0:t-1}, W_{0:t}) - \ln L^{\theta}(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t}) \right| \\
&\leqslant n^{-1} \sum_{t=0}^{n-1} \left| \ln L^{\theta}_{\pi,0:t}(Y_t \mid Y_{0:t-1}, W_{0:t}) - \ln L^{\theta}(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t}) \right| \\
&\leqslant n^{-1} \frac{2 + \sum_{t=1}^{n-1} \rho^{t-1}}{\sigma_-} \, .
\end{aligned}
$$

The proof of Equation (14) follows.

B.2 Proof of Theorem 2

We introduce the notation $\Delta_t(\theta) = \ln L^{\theta}(Y_t \mid Y_{-\infty:t-1}, W_{-\infty:t})$. Then, under Assumption H7 and using Equation (27), $\Delta_0(\theta) \in L^1$, see (17) in the proof of Proposition 1.1. Thus, Birkhoff's theorem can be used to establish that $\lim_{n\to\infty} n^{-1}\ell^s_n(\theta)$ exists $\mathbb{P}^*$-a.s., and

$$
\lim_{n\to\infty} n^{-1}\ell^s_n(\theta) = \lim_{n\to\infty} n^{-1} \sum_{t=0}^{n-1} \Delta_t(\theta) = \mathbb{E}^*[\Delta_0(\theta)] = \ell(\theta), \quad \mathbb{P}^*\text{-a.s.} \tag{18}
$$

After this first result, the rest of the proof is divided in three steps.

i) First, using H8 and Lemma 2.2, we show that $\theta \mapsto \mathbb{E}^*[\Delta_0(\theta)]$ is upper-semicontinuous.

ii) Then, introducing for all $n \geqslant 1$, $\bar{\theta}_n \in \operatorname{argmax}_{\theta\in\Theta} n^{-1}\ell^s_n(\theta)$, we establish that $\lim_{n\to\infty} d(\bar{\theta}_n, \Theta^*) = 0$, $\mathbb{P}^*$-a.s..

iii) Finally, combining i) and ii), we prove that $\lim_{n\to\infty} d(\hat{\theta}_{n,\hat{\pi}_n,w_{0:n-1}}, \Theta^*) = 0$, $\mathbb{P}^*$-a.s..

Let K a compact subset of Θ. For all $\theta_0 \in K$, $\rho > 0$,

$$
\limsup_{\rho\to 0} \limsup_{n\to\infty} \sup_{\theta\in B(\theta_0,\rho)} n^{-1} \sum_{k=0}^{n-1} \Delta_k(\theta) \leqslant \limsup_{\rho\to 0} \limsup_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \sup_{\theta\in B(\theta_0,\rho)} \Delta_k(\theta) \, .
$$

In the proof of Proposition 1.1, Equation (16) shows that for all $\theta \in \Theta$, $\mid \Delta_0(\theta) \mid \leqslant \mid \ln(\sigma_- b^-(0, Y_0, W_0)) \mid \vee \mid \ln(b^+) \mid$. Hence, $\sup_{\theta\in B(\theta_0,\rho)} \mid \Delta_0(\theta) \mid \leqslant \mid \ln(\sigma_- b^-(0, Y_0, W_0)) \mid \vee \mid \ln(b^+) \mid$ and under Assumption H 7, $\mathbb{E}[\sup_{\theta\in B(\theta_0,\rho)} \mid \Delta_0(\theta) \mid] < \infty$. Thus, $\sup_{\theta\in B(\theta_0,\rho)} \Delta_0(\theta)$ belongs to L^1 and by Birkhoff's ergodic theorem,

$$
\begin{aligned}
\limsup_{\rho\to 0} \limsup_{n\to\infty} \sup_{\theta\in B(\theta_0,\rho)} n^{-1} \sum_{k=0}^{n-1} \Delta_k(\theta) &\leqslant \limsup_{\rho\to 0} \limsup_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \sup_{\theta\in B(\theta_0,\rho)} \Delta_k(\theta) \\
&= \limsup_{\rho\to 0} \mathbb{E}^*\Big[\sup_{\theta\in B(\theta_0,\rho)} \Delta_0(\theta) \Big] \, .
\end{aligned}
$$

As the function $\rho \mapsto \sup_{\theta \in B(\theta_0,\rho)} \Delta_0(\theta)$ is non-decreasing, using the monotone convergence theorem,

$$\limsup_{\rho \to 0} \mathbb{E}^* \left[\sup_{\theta \in B(\theta_0,\rho)} \Delta_0(\theta) \right] = \mathbb{E}^* \left[\limsup_{\rho \to 0} \sup_{\theta \in B(\theta_0,\rho)} \Delta_0(\theta) \right] .$$

Under Assumption H8, for all $t \in \mathbb{Z}$, all $(x_{t-1}, x_t) \in \mathsf{X}^2$, all $(w_{t-1}, w_t) \in \mathsf{W}^2$ and all $y_t \in \mathsf{Y}$, $\theta \mapsto Q^\theta_{t-1|w_{t-1},w_t}(x_{t-1}, x_t)$ and $\theta \mapsto f^\theta_{t|w_t,x_t}(y_t)$ are continuous. Consequently, the function $\theta \mapsto \ln L^\theta_n(Y_{-m:t} \mid W_{-m:t})$ is continuous. Moreover, using (29), for all $\pi \in \mathcal{D}$ and all $\varepsilon > 0$, there exists $n_0 \geqslant 1$ such that for all $n, m \geqslant n_0$ with $m \geqslant n$:

$$\sup_{\theta \in \Theta} | \ln L^\theta_\pi(y_t \mid y_{-n:t-1}, w_{-n:t}) - \ln L^\theta_\pi(y_t \mid y_{-m:t-1}, w_{-m:t}) | \leqslant \frac{\rho^{t+n-1}}{\sigma_-} \leqslant \frac{\rho^{t+n_0-1}}{\sigma_-} \leqslant \varepsilon .$$

This result provides that $\{\ln L^\theta_\pi(Y_{-m:t} \mid W_{-m:t})\}_{m \geqslant 0}$ is an uniform Cauchy sequence. Therefore, the sequence $\theta \mapsto \ln L^\theta_n(Y_{-m:t} \mid W_{-m:t})$ is a sequence of continuous function and converges uniformly to $\theta \mapsto \ln L^\theta(Y_{-\infty:t} \mid W_{-\infty:t})$ that yields that this limit is continuous. Thus,

$$\limsup_{\rho \to 0} \mathbb{E}^* \left[\sup_{\theta \in B(\theta_0,\rho)} \Delta_0(\theta) \right] = \mathbb{E}^* \left[\limsup_{\rho \to 0} \sup_{\theta \in B(\theta_0,\rho)} \Delta_0(\theta) \right] = \mathbb{E}^*[\Delta_0(\theta_0)] .$$

We finally have:

$$\limsup_{\rho \to 0} \sup_{\theta \in B(\theta_0,\rho)} \mathbb{E}^*[\Delta_0(\theta)] \leqslant \limsup_{\rho \to 0} \mathbb{E}^*[\sup_{\theta \in B(\theta_0,\rho)} \Delta_0(\theta)] = \mathbb{E}^*[\Delta_0(\theta_0)] . \tag{19}$$

This result establishes that $\theta \mapsto \mathbb{E}^*[\Delta_0(\theta)]$ is upper-semicontinuous.

Consequently, $\Theta^* := \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}^*[\Delta_0(\theta)]$ is a closed and nonempty subset of Θ and for all $\epsilon > 0$, $K_\epsilon := \{\theta \in \Theta; d(\theta, \Theta^*) \geqslant \epsilon\}$ is a compact subset of Θ. The upper-semicontinuity of $\theta \mapsto \mathbb{E}^*[\Delta_0(\theta)]$ provides that there exists $\theta_\epsilon \in K_\epsilon$ such that, for all $\theta^* \in \Theta^*$,

$$\sup_{\theta \in K_\epsilon} \mathbb{E}^*[\Delta_0(\theta)] = \mathbb{E}^*[\Delta_0(\theta_\epsilon)] < \mathbb{E}^*[\Delta_0(\theta_*)] .$$

Consider now $\bar{\theta}_n$ a parameter such that $\{\bar{\theta}_n : n \in \mathbb{N}^*\} \subset \Theta$ and for all $n \geqslant 1$, $\bar{\theta}_n \in \operatorname{argmax}_{\theta \in \Theta} n^{-1} \ell^s_n(\theta)$. We want to prove the intermediate result $\lim_{n \to \infty} d(\bar{\theta}_n, \Theta^*) = 0$, $\mathbb{P}^*$-a.s. For all $\eta > 0$ and $\tilde{\theta} \in K$, we can find a $\rho_{\tilde{\theta}} > 0$ such that,

$$\limsup_{n \to \infty} \sup_{\theta \in B(\tilde{\theta}, \rho_{\tilde{\theta}})} n^{-1} \sum_{k=0}^{n-1} \Delta_k(\theta) \leqslant \mathbb{E}^*[\Delta_0(\tilde{\theta})] + \eta \leqslant \sup_{\theta \in K} \mathbb{E}^*[\Delta_0(\theta)] + \eta \quad \mathbb{P}^*\text{-a.s.}$$

Since K is a compact subset, there exist $p \geqslant 1$ and a collection of $\theta_i \in K$ for $i \in \{1, \cdots, p\}$ such that $K \subset \bigcup_{i=1}^p B(\theta_i, \rho_{\theta_i})$ and

$$\limsup_{n \to \infty} \max_{1 \leqslant i \leqslant p} \sup_{\theta \in B(\theta_i, \rho_{\theta_i})} n^{-1} \sum_{k=0}^{n-1} \Delta_k(\theta) \leqslant \sup_{\theta \in K} \mathbb{E}^*[\Delta_0(\theta)] + \eta \quad \mathbb{P}^*\text{-a.s.}$$

The previous inequality is then equivalent to

$$\limsup_{n \to \infty} \sup_{\theta \in K} n^{-1} \sum_{k=0}^{n-1} \Delta_k(\theta) \leqslant \sup_{\theta \in K} \mathbb{E}^*[\Delta_0(\theta)] + \eta \quad \mathbb{P}^*\text{-a.s.}$$

Since η is arbitrary,

$$\limsup_{n\to\infty} \sup_{\theta\in K} n^{-1} \sum_{k=0}^{n-1} \Delta_k(\theta) \leqslant \sup_{\theta\in K} \mathbb{E}^*[\Delta_0(\theta)] \quad \mathbb{P}^*\text{-a.s.}$$

We have finally that,

$$\begin{aligned}\limsup_{n\to\infty} \sup_{\theta\in K_\epsilon} n^{-1}\ell_n^s(\theta) &= \limsup_{n\to\infty} \sup_{\theta\in K_\epsilon} n^{-1} \sum_{t=0}^{n-1} \Delta_k(\theta) \\ &\leqslant \sup_{\theta\in K_\epsilon} \mathbb{E}^*[\Delta_0(\theta)] \\ &< \mathbb{E}^*[\Delta_0(\theta^*)] = \lim_{n\to\infty} n^{-1}\ell_n^s(\theta^*) \leqslant \liminf_{n\to\infty} n^{-1}\ell_n^s(\bar\theta_n) \end{aligned} \tag{20}$$

The last result insures that $\bar\theta_n \notin K_\epsilon$ for all n greater that a certain random rank N and as ϵ is arbitrary, we prove:

$$\lim_{n\to\infty} d(\bar\theta_n, \Theta^*) = 0, \ \mathbb{P}^*\text{-a.s.} \tag{21}$$

Using the result above and the upper-semicontinuity of $\theta \mapsto \mathbb{E}^*[\Delta_0(\theta)]$, the final result can be proved. For all $\theta^* \in \Theta^*$, $\mathbb{P}^*$-a.s.,

$$\begin{aligned}\mathbb{E}^*[\Delta_0(\theta^*)] = \liminf_{n\to\infty} n^{-1}\ell_n^s(\theta^*) \leqslant \liminf_{n\to\infty} n^{-1}\ell_n^s(\bar\theta_n) &\leqslant \limsup_{n\to\infty} n^{-1}\ell_n^s(\bar\theta_n) \\ &= \limsup_{n\to\infty} \sup_{\theta\in\Theta} n^{-1}\ell_n^s(\theta) \leqslant \sup_{\theta\in\Theta} \mathbb{E}^*[\Delta_0(\theta)]\end{aligned}$$

Since $\theta \mapsto \mathbb{E}^*[\Delta_0(\theta)]$ is upper-semicontinuous and Θ^* is a closed and nonempty subset of Θ, $\theta \mapsto \mathbb{E}^*[\Delta_0(\theta)]$ reaches is maximum and we have:

$$\sup_{\theta\in\Theta} \mathbb{E}^*[\Delta_0(\theta)] = \max_{\theta\in\Theta} \mathbb{E}^*[\Delta_0(\theta)] = \mathbb{E}^*[\Delta_0(\theta^*)]\,.$$

Consequently,

$$\lim_{n\to\infty} n^{-1}\ell_n^s(\bar\theta_n) = \mathbb{E}^*[\Delta_0(\theta^*)], \quad \mathbb{P}^*\text{-a.s.}$$

We introduce $\delta_n := \sup_{\theta\in\Theta} \sup_{\pi\in\mathcal{D}} n^{-1} \mid \ell_{\pi,n}(\theta) - \ell_n^s(\theta) \mid$. Then,

$$n^{-1}\ell_n^s(\bar\theta_n) - \delta_n \leqslant n^{-1}\ell_{\hat\pi_n,n}(\bar\theta_n) \leqslant n^{-1}\ell_{\hat\pi_n,n}(\hat\theta_{n,\hat\pi_n,w_{0:n-1}}) \leqslant n^{-1}\ell_n^s(\hat\theta_{n,\hat\pi_n,w_{0:n-1}}) + \delta_n \leqslant n^{-1}\ell_n^s(\bar\theta_n) + \delta_n\,.$$

Thus, using Proposition1.1,

$$\lim_{n\to\infty} n^{-1}\ell_n^s(\hat\theta_{n,\hat\pi_n,w_{0:n-1}}) = \mathbb{E}^*[\Delta_0(\theta^*)], \quad \mathbb{P}^*\text{-a.s.}$$

Then, as for Equation (20),

$$\begin{aligned}\limsup_{n\to\infty} \sup_{\theta\in K_\epsilon} n^{-1}\ell_n^s(\theta) = \limsup_{n\to\infty} \sup_{\theta\in K_\epsilon} n^{-1} \sum_{t=0}^{n-1} \Delta_k(\theta) &\leqslant \sup_{\theta\in K_\epsilon} \mathbb{E}^*[\Delta_0(\theta)] \\ &< \mathbb{E}^*[\Delta_0(\theta^*)] = \lim_{n\to\infty} n^{-1}\ell_n^s(\hat\theta_{n,\hat\pi_n,w_{0:n-1}})\,.\end{aligned}$$

Therefore, $\hat\theta_{n,\hat\pi_n,w_{0:n-1}} \notin K_\epsilon$. As ϵ is arbitrary, the proof follows.

B.3 Technical results

Define, for all $x, x' \in \mathsf{X}$, $y \in \mathsf{Y}$, $w, w' \in \mathsf{W}$, $\theta \in \Theta$, and $t \in \mathbb{Z}$:

$$M^{\theta,y}_{t,w,w'}(x,x') = Q^{\theta}_{t-1|w,w'}(x,x') f^{\theta}_{t|w,x'}(y)\,.$$

For every $m \in \mathbb{Z}$ such that $-m \leqslant t$, we introduce the distribution of the hidden state X_t given the observations $(Y_{-m:t}, W_{-m:t})$ and starting at time $-m$ with $X_{-m} \sim \pi$ and π a discrete distribution on X. For all $A \in \mathcal{X}$,

$$\mathbb{P}^{\theta}_{\pi,t|-m}(X_t \in A \mid Y_{-m:t}, W_{-m:t}) = \frac{\sum_{x_{-m:t}\in\mathsf{X}^{t+m+1}} \pi(x_{-m}) f^{\theta}_{-m|W_{-m},x_{-m}}(Y_{-m}) \prod_{p=-m+1}^{t} M^{\theta,Y_p}_{p,W_{p-1},W_p}(x_{p-1},x_p) \mathbb{1}_A(x_t)}{\sum_{x_{-m:t}\in\mathsf{X}^{t+m+1}} \pi(x_{-m}) f^{\theta}_{-m|W_{-m},x_{-m}}(Y_{-m}) \prod_{p=-m+1}^{t} M^{\theta,Y_p}_{p,W_{p-1},W_p}(x_{p-1},x_p)}\,. \tag{22}$$

Using Assumption H6-(b), Lemma 2.1 establishes the forgetting of the initial distribution π for $\mathbb{P}^{\theta}_{\pi,t|-m}(X_t \in . \mid Y_{t:-m}, W_{t:-m})$.

Lemma 2.1. *Under H6, for all probability measures π, π', $k \geqslant 1$ and $-\infty < r \leqslant s < \infty$,*

$$\sup_{\theta\in\Theta} d_{\mathrm{TV}}(\mathbb{P}^{\theta}_{\pi,s|r}(X_s \in . \mid Y_{r:s}, W_{r:s}), \mathbb{P}^{\theta}_{\pi',s|r}(X_s \in . \mid Y_{r:s}, W_{r:s})) \leqslant \rho^{s-r}\,, \tag{23}$$

$$\sup_{\theta\in\Theta} d_{\mathrm{TV}}(\mathbb{P}^{\theta}_{\pi,s|r}(X_s \in . \mid Y_{r:s}, W_{r:s}), \mathbb{P}^{\theta}_{\pi,s|r-k}(X_s \in . \mid Y_{r-k:s}, W_{r-k:s})) \leqslant \rho^{s-r} \quad \mathbb{P}^*\text{-a.s.}\,, \tag{24}$$

where $\rho = 1 - \sigma_-$ with σ_- is defined in Assumption H6-(b).

Proof. Assuming that Equation (23) holds, for all discrete distributions π, π' on X and all $-\infty < r \leqslant s < \infty$:

$$d_{\mathrm{TV}}(\mathbb{P}^{\theta}_{\pi,s|r}(X_s \in . \mid Y_{r:s}W_{r:s}), \mathbb{P}^{\theta}_{\pi',s|r}(X_s \in . \mid Y_{r:s}W_{r:s})) \leqslant \rho^{s-r}\,.$$

For all $x_r \in \mathsf{X}$, define $\tilde{\pi}(x_r) \propto \sum_{(x_{r-k},\cdots,x_{r-1})\in\mathsf{X}^k} \pi(x_{r-k}) \prod_{p=r-k}^{r-1} f^{\theta}_{p|W_p,x_p}(Y_p) Q^{\theta}_{p|W_p,W_{p+1}}(x_p, x_{p+1})$. Then, for all $A \in \mathcal{X}$, using Equation (22):

$$\begin{aligned}
&\mathbb{P}^{\theta}_{\pi,s|r-k}(X_s \in A \mid Y_{r-k:s}, W_{r-k:s}) \\
&\quad= \frac{\sum_{x_{r-k:s}\in\mathsf{X}^{s-r+k+1}} \pi(x_{r-k}) f^{\theta}_{r-k|W_{r-k},x_{r-k}}(Y_{r-k}) \prod_{p=r-k+1}^{s} M^{\theta,Y_p}_{p,W_{p-1},W_p}(x_{p-1},x_p) \mathbb{1}_A(x_s)}{\sum_{x_{r-k:s}\in\mathsf{X}^{s-r+k+1}} \pi(x_{r-k}) f^{\theta}_{r-k|W_{r-k},x_{r-k}}(Y_{r-k}) \prod_{p=r-k+1}^{s} M^{\theta,Y_p}_{p,W_{p-1},W_p}(x_{p-1},x_p)} \\
&\quad= \frac{\sum_{x_{r:s}\in\mathsf{X}^{s-r+1}} \tilde{\pi}(x_r) f^{\theta}_{r|W_r,x_r}(Y_r) \prod_{p=r+1}^{s} M^{\theta,Y_p}_{p,W_{p-1},W_p}(x_{p-1},x_p) \mathbb{1}_A(x_s)}{\sum_{x_{r:s}\in\mathsf{X}^{s-r+1}} \tilde{\pi}(x_r) f^{\theta}_{r|W_r,x_r}(Y_r) \prod_{p=r+1}^{s} M^{\theta,Y_p}_{p,W_{p-1},W_p}(x_{p-1},x_p)} \\
&\quad= \mathbb{P}^{\theta}_{\tilde{\pi},s|r}(X_s \in A \mid Y_{r:s}, W_{r:s})\,.
\end{aligned}$$

Combining Equation (23) and the previous result yields (24). For a better readability, we introduce the following quantity:

$$M^{\theta}_{r,s}(x_r, x_s) = \sum_{x_{r+1:s-1}\in\mathsf{X}^{s-r-1}} \prod_{p=r+1}^{s} M^{\theta,Y_p}_{p,W_{p-1},W_p}(x_{p-1},x_p)\,.$$

Then, we can notice that for all $\theta \in \Theta$, if we denote $h : x_s \mapsto \mathbb{1}_A(x_s)$, $\bar{\pi}(x_r) \propto \pi(x_r) f^{\theta}_{r|W_r,x_r}(Y_r)$ and $\bar{\pi}'(x_r) \propto \pi'(x_r) f^{\theta}_{r|W_r,x_r}(Y_r)$, we have $\mathbb{P}^{\theta}_{\pi,r|s}(X_s \in A \mid Y_{r:s}W_{r:s}) = \bar{\pi} M^{\theta}_{r,s} h / \bar{\pi} M^{\theta}_{r,s} \mathbb{1}$ and $\mathbb{P}^{\theta}_{\pi',r|s}(X_s \in A \mid Y_{r:s}W_{r:s}) = \bar{\pi}' M^{\theta}_{r,s} h / \bar{\pi}' M^{\theta}_{r,s} \mathbb{1}$. Then, proving Equation (23) is equivalent to prove:

$$\left| \frac{\bar{\pi} M^{\theta}_{r,s} h}{\bar{\pi} M^{\theta}_{r,s} \mathbb{1}} - \frac{\bar{\pi}' M^{\theta}_{r,s} h}{\bar{\pi}' M^{\theta}_{r,s} \mathbb{1}} \right| \leqslant \rho^{s-r} \mathrm{osc}(h) \,,$$

where $\rho = 1 - \sigma_-$. Introducing for all $A \in \mathcal{P}(\mathsf{X})$,

$$\widetilde{M}^{\theta}_{r,s}(x_r, A) = \frac{M^{\theta}_{r,s}(x_r, A)}{M^{\theta}_{r,s}(x_r, \mathsf{X})} \,,$$

we obtain

$$\frac{\bar{\pi} M^{\theta}_{r,s} h}{\bar{\pi} M^{\theta}_{r,s} \mathbb{1}} = \frac{\sum_{x_r \in \mathsf{X}} \bar{\pi}(x_r) M^{\theta}_{r,s}(x_r, \mathsf{X}) \widetilde{M}^{\theta}_{r,s} h(x_r)}{\sum_{x_r \in \mathsf{X}} \bar{\pi}(x_r) M^{\theta}_{r,s}(x_r, \mathsf{X})} = \bar{\pi}_{r,s} \widetilde{M}^{\theta}_{r,s} h \,, \tag{25}$$

where for all $B \in \mathcal{X}$,

$$\bar{\pi}_{r,s}(B) = \frac{\sum_{x_r \in X} \bar{\pi}(x_r) M^{\theta}_{r,s}(x_r, \mathsf{X}) \mathbb{1}_B(x_r)}{\sum_{x_r \in \mathsf{X}} \bar{\pi}(x_r) M^{\theta}_{r,s}(x_r, \mathsf{X})} \,.$$

Then, we have

$$\begin{aligned} \widetilde{M}^{\theta}_{r,s}(x_r, A) = \frac{M^{\theta}_{r,s}(x_r, A)}{M^{\theta}_{r,s}(x_r, \mathsf{X})} &= \frac{\sum_{x_{r+1} \in \mathsf{X}} M^{\theta}_{r+1}(x_r, x_{r+1}) M^{\theta}_{r+1,s}(x_{r+1}, \mathsf{X}) \widetilde{M}^{\theta}_{r+1,s}(x_{r+1}, A)}{M^{\theta}_{r,s}(x_r, \mathsf{X})} \\ &= R^{\theta}_{r,s} \widetilde{M}^{\theta}_{r+1,s}(x_r, A) \,, \end{aligned}$$

where the kernel $R^{\theta}_{r,s}$ is defined, for all $x_r \in \mathsf{X}$, $A \in \mathcal{X}$, by:

$$R^{\theta}_{r,s}(x_r, A) = \frac{\sum_{x_{r+1} \in A} M^{\theta}_{r+1}(x_r, x_{r+1}) M^{\theta}_{r+1,s}(x_{r+1}, \mathsf{X})}{M^{\theta}_{r,s}(x_r, \mathsf{X})} \,.$$

Consequently, by induction, the quantity $\widetilde{M}^{\theta}_{r,s}$ can be expressed as

$$\widetilde{M}^{\theta}_{r,s} = R^{\theta}_{r,s} R^{\theta}_{r+1,s} \cdots R^{\theta}_{s-1,s} \,. \tag{26}$$

Using Assumption H6-(b), for any $x_s \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$R^{\theta}_{r,s}(x_s, A) = \frac{\sum_{x_{r+1} \in A} M^{\theta}_{r+1}(x_r, x_{r+1}) M^{\theta}_{r+1,s}(x_{r+1}, \mathsf{X})}{\sum_{x_{r+1} \in \mathsf{X}} M^{\theta}_{r+1}(x_r, x_{r+1}) M^{\theta}_{r+1,s}(x_{r+1}, \mathsf{X})} \geqslant \sigma_- v_{r,s}(A) \,,$$

with:

$$v_{r,s}(A) = \frac{\sum_{x_{r+1} \in A} f^{\theta}_{r+1|W_{r+1},x_{r+1}}(Y_{r+1}) M^{\theta}_{r+1,s}(x_{r+1}, \mathsf{X})}{\sum_{x_{r+1} \in \mathsf{X}} f^{\theta}_{r+1|W_{r+1},x_{r+1}}(Y_{r+1}) M^{\theta}_{r+1,s}(x_{r+1}, \mathsf{X})} \,.$$

Consequently, the kernel $R_{r,s}$ verifies the Doeblin condition and its Dobrushin coefficient satisfies $\Delta_{TV}(R^{\theta}_{r,t}) \leqslant \rho$, see [Douc et al., 2014] (definition 6.9 and Lemma 6.10). Moreover, as the Dobrushin coefficient is submultiplicative and using the decomposition (26) yields

$$\Delta_{TV}(\widetilde{M}^{\theta}_{r,s}) \leqslant \Delta_{TV}(R^{\theta}_{r,s}) \Delta_{TV}(R^{\theta}_{r+1,s}) \cdots \Delta_{TV}(R^{\theta}_{s-1,s}) \leqslant \rho^{s-r} \,.$$

Finally, for any $\pi, \pi' \in \mathbb{M}_1(\mathsf{X})$, by Lemma 6.5 introduced in [Douc et al., 2014] and (25),

$$\left| \frac{\bar{\pi} M^\theta_{r,s}}{\bar{\pi} M^\theta_{r,s} \mathbb{1}} - \frac{\bar{\pi}' M^\theta_{r,s}}{\bar{\pi}' M^\theta_{r,s} \mathbb{1}} \right|_{\mathrm{TV}} = | \, \bar{\pi}_{r,s} \widetilde{M}^\theta_{r,s} - \bar{\pi}'_{r,s} \widetilde{M}^\theta_{r,s} \, |_{\mathrm{TV}} \leqslant \Delta_{\mathrm{TV}}(\widetilde{M}^\theta_{r,s}) \mid \bar{\pi}_{r,s} - \bar{\pi}'_{r,s} \mid_{\mathrm{TV}}$$
$$\leqslant \rho^{s-r} \mid \bar{\pi}_{r,s} - \bar{\pi}'_{r,s} \mid_{\mathrm{TV}} .$$

Then,

$$\left| \frac{\bar{\pi} M^\theta_{r,s} h}{\bar{\pi} M^\theta_{r,s} \mathbb{1}} - \frac{\bar{\pi}' M^\theta_{r,s} h}{\bar{\pi}' M^\theta_{r,s} \mathbb{1}} \right| \leqslant \frac{1}{2} \left| \frac{\bar{\pi} M^\theta_{r,s}}{\bar{\pi} M^\theta_{r,s} \mathbb{1}} - \frac{\bar{\pi}' M^\theta_{r,s}}{\bar{\pi}' M^\theta_{r,s} \mathbb{1}} \right|_{\mathrm{TV}} \operatorname{osc}(h) \leqslant \rho^{s-r} \frac{1}{2} \left| \bar{\pi}_{r,s} - \bar{\pi}'_{r,s} \right|_{\mathrm{TV}} \operatorname{osc}(h)$$
$$\leqslant \rho^{s-r} \operatorname{osc}(h) ,$$

which concludes the proof. □

For all $t, m \in \mathbb{Z}$ verifying $-m < t$, the conditional loglikelihood can be written as follows

$$L^\theta_{\pi,-m:t}(Y_t \mid Y_{-m:t-1}, W_{-m:t}) = \sum_{(x_{t-1},x_t) \in \mathsf{X}^2} \mathbb{P}^\theta_{\pi,-m}(X_{t-1} = x_{t-1} \mid Y_{-m:t-1}, W_{-m:t}) Q^\theta_{t-1|W_{t-1},W_t}(x_{t-1}, x_t) f^\theta_{t|W_t,x_t}(Y_t) . \quad (27)$$

Lemma 2.2. *Under Assumptions H5-H7, for all probability measures π, π', all $t \geqslant 1$, all $k \geqslant 1$, all $m \geqslant 0$, all sequences $y_{-m:t} \in \mathsf{Y}^{t+m+1}$ and all sequences $w_{-m:t} \in \mathsf{W}^{t+m+1}$:*

$$\sup_{\theta \in \Theta} | \ln L^\theta_{\pi,-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) - \ln L^\theta_{\pi',-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) | \leqslant \frac{\rho^{t+m-1}}{\sigma_-} , \quad (28)$$

$$\sup_{\theta \in \Theta} | \ln L^\theta_{\pi,-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) - \ln L^\theta_{\pi,-m-k:t}(y_t \mid y_{-m-k:t-1}, w_{-m-k:t}) \leqslant \frac{\rho^{t+m-1}}{\sigma_-} , \quad (29)$$

$$\sup_{\theta \in \Theta} \sup_{m \geqslant 0} L^\theta_{\pi,-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) \leqslant b^+ . \quad (30)$$

Proof. Equation (30) is a direct consequence of Equation (27). The proof of Equation (29) can be derived using Equation (28) by fixing:

$$\pi'(x_{-m}) \propto \sum_{x_{-m-k:-m-1} \in \mathsf{X}^k} \pi(x_{-m-k}) \prod_{p=-m-k}^{-m-1} f^\theta_{p|w_p,x_p}(y_p) Q^\theta_{p|w_p,w_{t+1}}(x_p, x_{p+1}) .$$

We turn to the proof of Equation (28). Using (27), Lemma 2.1 and Assumption H6, we have:

$$| \, L^\theta_{\pi,-m:t}(y_t \, | y_{-m:t-1}, w_{-m:t}) - L^\theta_{\pi',-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) \, |$$
$$\leqslant \left| \sum_{(x_{t-1},x_t) \in \mathsf{X}^2} Q^\theta_{t-1|w_{t-1},w_t}(x_{t-1}, x_t) f^\theta_{t|w_t,x_t}(y_t) \left\{ p^\theta_{\pi,-m:t}(x_{t-1}) - p^\theta_{\pi',-m:t}(x_{t-1}) \right\} \right|$$
$$\leqslant \rho^{t+m-1} \sum_{x_t \in \mathsf{X}} f^\theta_{t|w_t,x_t}(y_t) ,$$

where $p^{\theta}_{\pi,-m:t}(x_{t-1}) = \mathbb{P}^{\theta}_{\pi,-m}(X_{t-1} = x_{t-1} \mid Y_{-m:t-1} = y_{-m:t-1}, W_{-m:t} = w_{-m:t})$. Moreover,

$$L^{\theta}_{\pi,-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) \vee L^{\theta}_{\pi',-m:t}(y_t \mid y_{-m:t-1}, w_{-m:t}) \geqslant \sigma_- \sum_{x_t \in \mathsf{X}} f^{\theta}_{t|w_t,x_t}(y_t)\,.$$

Using that $|\ln u - \ln v| \leqslant |u - v| / (u \vee v)$, where $u \vee v = \max(u, v)$, completes the proof of (28). □