General framework and model building in the class of Hidden Mixture Transition Distribution models

Danilo Bolano^{a,b,*}, André Berchtold^{a,c}

^a*National Center of Competence in Research LIVES, Switzerland*

^b*Institute of Demographic and Life Course Studies, University of Geneva, Switzerland*

^c*Institute of Social Sciences, University of Lausanne, Switzerland*

Abstract

Modelling time series that present non-Gaussian features play as central role in many fields, including finance, seismology, psychological, and life course studies. The Hidden Mixture Transition Distribution model is an answer to the complexity of such series. The observed heterogeneity can be induced by one or several latent factors, and each level of these factors is related to a different component of the observed process. The time series is then treated as a mixture and the relation between the components is governed by a Markovian latent transition process. This framework generalizes several specifications that appear separately in related literature. Both the expectation and the standard deviation of each component are allowed to be functions of the past of the process. The latent process can be of any order, and can be modeled using a discrete Mixture Transition Distribution. The effects of covariates at the visible and hidden levels are also investigated. One of the main difficulties lies in correctly specifying the structure of the model. Therefore, we propose a hierarchical model selection procedure that exploits the multilevel structure of our approach. Finally, we illustrate the model and the model selection procedure through a real application in social science.

Keywords: Mixture Model, Model Selection, Hidden Markov Model, Mixture Transition Distribution model, BIC, Panel Data.

1. Introduction

Real data are often a combination of many different, possibly non-observed causes that lead to apparently unpredictable behaviors. For example, in the context of longitudinal data, time series may show non-homogeneous behaviors,

*Corresponding author. University of Geneva (CH), bd. du Pont d'Arve 40. Tel: +41 22 379 98 74

Email addresses: Danilo.Bolano@unige.ch (Danilo Bolano), Andre.Berchtold@unil.ch (André Berchtold)

5 can switch between alternative regimes characterized by a low or high vari-
6 ance, can contain extreme values, and the distribution of future values can take
7 complex multimodal shapes.

8 The Hidden Mixture Transition Distribution (HMTD) model considered in
9 this study is a general framework to study time series. The model can be used to
10 describe and analyze the evolution of any continuous variable observed on a set
11 of M independent sequences that can also vary in length. The model integrates
12 several refinements of the mixture model and the hidden Markov model. More
13 specifically, it can be used for different purposes, including describing observed
14 data, searching for a generalizable model, testing hypotheses, prediction, and
15 classifying time series.

16 Mixture models are a popular and efficient approach, both for cross-sectional
17 data and time series, to describe multimodal distributions that do not corre-
18 spond to any specific statistical family. Related literature provides many exam-
19 ples of the usefulness of such models since the work of Weldon and Pearson at the
20 end of the 19th century. Historically, mixture models for continuous-valued time
21 series were introduced in Le et al. (1996) as the Gaussian Mixture Transition
22 Distribution (GMTD) model, building upon the earlier work of Raftery (1985)
23 in the discrete case. See Berchtold and Raftery (2002) for a complete review
24 of the basic principles of Mixture Transition Distribution (MTD) models. The
25 general principle of all MTD-like models for count data is to combine different
26 Gaussian distributions (called *components*) using a mixture model in which the
27 mean of each distribution is a function of the past observed process. The weights
28 associated to each component are interpreted as the probability of that specific
29 component generating the next value of the process. This class of models has
30 been expanded on in several ways. For example, allowing detailed specifications
31 for the mean of each component (e.g., Wong and Li, 2001), allowing the vari-
32 ance of each component to depend on its past (Wong and Li, 2001; Berchtold,
33 2003), and replacing the fixed probabilities associated with each component by
34 a Markov chain (Bartolucci and Farcomeni, 2010). Then, covariates have been
35 included (Chariatte et al., 2008), Luo and Qiu (2009) used non-Gaussian dis-
36 tributions, and extensions to bivariate data have been suggestion (Hassan and
37 Lii, 2006; Hassan and El-Bassiouni, 2013).

38 Hidden Markov Models (HMM) form another class of stochastic processes
39 often used to represent and analyze complex time series in the presence of over-
40 dispersion. They are particularly well suited to analyzing data switching be-
41 tween several regimes. In its traditional formulation, a HMM combines a hidden
42 first-order Markov chain with different conditionally independent distributions
43 for the observed process. Here each distribution is related to one of the states
44 of the Hidden Markov chain. Many developments to the basic HMM have been
45 proposed, including using high-order Markov chains and removing the condi-
46 tional independence hypothesis. The latter change allows the observed process
47 to depend on both its past and the hidden process (Wellekens, 1987; Berchtold,
48 1999, 2002). Under the hypothesis of stationarity, the marginal distribution
49 of the observations in a HMM is a finite mixture, with the number of compo-
50 nents being equal to the number of states in the unobserved Markov process.

Therefore, we may consider the two models as one unique approach. When the transitions between components are driven by a Markov chain, the mixture transition model becomes a HMM. In addition, when the observations are allowed to follow an autoregressive process, the HMM becomes a mixture transition model. Some authors even refer to HMMs as Markov-dependent mixture models (as coined by Leroux, 1992).

It is worth noting that mixture transition and Markovian models are distributional stochastic models, in contrast to other well-known and widespread approaches that are essentially point processes. The latter group includes the ARMA and ARIMA models, which are based on autoregressive equations, and the ARCH and GARCH models, which explicitly consider the variance of the process. See, for example, Box et al. (1994), Hamilton (1994), and Bollerslev et al. (1992) for a complete review of these models, and Kon (1984) and Kim and Kon (1994) for a comparison of the different models. When trying to predict the next value of a series, the advantage of the point approach is that the model will provide a clear answer as one numerical value (associated with a confidence interval). However, the drawback is that, in most cases, the answer is either inaccurate or completely wrong. Given the high variability of many time series, the expectation of the model is not a good estimator of the value of the next observation. Even when modeling the variance of the process, there is only a small probability of accurately predicting the next value. However, in the probabilistic approach, instead of trying to determine *the* next value of the series, a non-null probability is associated with values (discrete case) or intervals (continuous case), which are then possible candidates to be *the* next data. Then, an adequate probabilistic model does not provide a single value, but rather leads to a complete representation of possible futures through a (possibly multimodal) distribution. In that sense, the answer given by this approach generally has a higher probability of helping to make the right decisions, because it shows all possibilities rather than one (probably) wrong value.

Hidden Markov models were historically used for speech recognition (Rabiner, 1989; Baum and Petrie, 1966), but many applications have since been found in other fields, including econometrics (e.g., Elliott et al., 1998; Hayashi, 2004; Netzer et al., 2008) and the biosciences (e.g., Le Strat and Carrat, 1999; Shirley et al., 2010). Mixture models for count data are also quite common in finance and biomedical studies (Sclatmann, 2009) and behavioral studies (under the name of growth mixture modeling, e.g., Muthen, 2001). However, Mixture Transition Models seem to be used quite exclusively in economics and finance (e.g., Wong and Chan, 2005; Frydman and Schuermann, 2008). In fact, despite their unanimously recognized advantages, mixture transition and hidden models are still only sparsely used in social sciences. This is unfortunate, since the current trend in this field is clearly to switch from cross-sectional to longitudinal surveys, hence the need for advanced methods for modeling longitudinal data showing non-Gaussian distributions.

In addition, even though many developments have been proposed during the past few decades on the basic MTD and HMM models, there is still a need for a more general framework that integrates these refinements. As a result, this

study has three objectives. First, we define a general framework that integrates the many extensions presented previously separately in the literature (see Section 2), and then we discuss the estimation procedure (see Section 3). Second, as noted by Rabiner (1989), the number of possibilities offered by hidden models is so large that it becomes difficult to identify an adequate model structure for a particular research question. Therefore, in Section (4), we outline a search strategy similar to that used with other multilevel models. Finally, we illustrate the model and the proposed procedure using a real dataset, the U.S. Panel Study of Income Dynamics (see Section 5).

2. The HMTD model

The Hidden Mixture Transition Distribution (HMTD) model developed in this paper combines a hidden and an observed level. It can be used to describe and analyze the evolution of any continuous variable observed on a set of M independent sequences, which may vary in length.

At the hidden level, a latent discrete variable, S , taking values in the finite set $\{1, \dots, k\}$ follows a Markov chain of order $\ell \geq 0$ ($\ell = 0$ is a special case reducing the model to a simple mixture). Each of these values represents a different hidden state of the process. In addition to its lags, the value taken by S at time t can also be influenced by one or several categorical covariates. Since S is unobserved, we never know its exact value at time t , but we can estimate its distribution and the most likely sequence of states corresponding to each sequence of observations.

At the observed level, we consider a random variable, X , taking values in $\mathbb{R}$. A Gaussian component is associated with each hidden state of the latent process, with an expectation and variance that may depend on both the lags of X and on a set of categorical and/or continuous covariates. Since we only know the distribution of the possible hidden states at time t , the visible model is a mixture of the k Gaussian components.

The model can be estimated by simultaneously using as many independent sequences of observations as required. Each sequence typically corresponds to the observation of a separate subject. Each hidden state and its associated visible component can instead be interpreted as one possible behavior of the subjects under investigation and multiplying the number of components, we allow the subjects to follow different behaviors. This enables us to capture both the complexity of the overall population and the evolution of each individual over time.

The following subsections describe the two levels of the model in more detail.

2.1. The visible level

Let $\{X_t, t \in \mathbb{N}^*\}$ be a sequence of random variables taking values in $\mathbb{R}$. Let X_{t-a}^{t-1} denote the past observations between time $t-a$ and $t-1$. A frequently used and convenient hypothesis assumes that the probability of X_t , given its past, follows a Gaussian distribution, but this hypothesis is generally too simplistic

139 to account for the complexity of real data. A better solution is to assume that
 140 the observed time series was generated by k different sub-models, each model
 141 being used for one or several parts of the overall time series, and to write the
 142 resulting model as a mixture (McLachlan and Peel, 2000). We then have

$$F(x_t|x_1^{t-1}) = \sum_{g=1}^k \lambda_g(t) G_g(x_t|x_{t-r_g}^{t-1}, C(t)).$$

143 where $F(x_t|x_1^{t-1})$ is the cumulative distribution function of x_t , given its past,
 144 $G_g(x_t|x_{t-r_g}^{t-1})$ is a cumulative distribution function of x_t , given a part of its past
 145 (from x_{t-r_g} to x_{t-1} , $r_g \geq 1$), $C(t)$ represents a set of covariates available at
 146 time t , and $\lambda_g(t)$ is the weight of the g -th component at time t , with

$$\sum_{g=1}^k \lambda_g(t) = 1, \quad \lambda_g(t) > 0, \quad \forall g, t.$$

147 Different specifications can be used for the G s. However, we only consider
 148 the Gaussian case here because, as pointed out by Rabiner (1989), Gaussian
 149 mixtures can approximate almost all continuous density functions as closely as
 150 necessary. The g -th component is then written as

$$G_g(x_t|x_{t-r_g}^{t-1}) = \Phi\left(\frac{x_t - \mu_{g,t}}{\sigma_{g,t}}\right).$$

151 In order to explicitly incorporate the dependence between successive ob-
 152 servations into the model, the expectation and the standard deviation of each
 153 Gaussian component are written as functions of the past. The expectation of
 154 the g -th component is specified by

$$\mu_{g,t} = \varphi_{g,0} + \sum_{i=1}^{p_g} \varphi_{g,i} x_{t-i} + \sum_{j=1}^{c_g} \delta_{g,j} c_j(t), \quad p_g \geq 0, c_g \geq 0. \quad (2.1)$$

155 The first part of the equation is an autoregressive model, with the $\varphi_{g,i}$
 156 coefficient associated with the i -th lag of X_t . The second part represents the
 157 influence of the covariates: $c_j(t)$ is the j -th covariate observed at time t , and
 158 $\delta_{g,j}$ is its associated coefficient. Covariates can be continuous or categorical. As
 159 usual, in order to facilitate the interpretation of the results, we suggest recoding
 160 the categorical covariates as 0-1 dummy variables. When the number of lags is
 161 fixed as zero ($p_g = 0$) and there are no covariates ($c_g = 0$), the expectation of
 162 the component becomes a constant (Equation 2.1).

163 Different specifications can be chosen to model the standard deviation of the
 164 g -th component:

$$\sigma_{g,t} = \sqrt{\theta_{g,0} + \sum_{j=1}^{q_g} \theta_{g,j} x_{t-j}^2}, \quad q_g \geq 0, \quad (2.2)$$

$$\sigma_{g,t} = \sqrt{\theta_{g,0} + \sum_{j=1}^{q_g} \theta_{g,j} (x_{t-j} - \bar{x}_{t-q_g}^{t-1})^2}, \quad q_g \geq 2, \quad (2.3)$$

$$\sigma_{g,t} = \sqrt{\theta_{g,0} + \sum_{j=1}^{q_g} \theta_{g,j} (x_{t-j} - \mu_{g,t})^2}, \quad q_g \geq 1, \quad (2.4)$$

$$\sigma_{g,t} = \sqrt{\theta_{g,0} + \sum_{j=1}^{q_g} \theta_{g,j} e_{g,t-j}^2}, \quad q_g \geq 1, \quad (2.5)$$

165 where

$$\begin{aligned} \bar{x}_{t-q_g}^{t-1} &= \frac{1}{q_g} \sum_{j=1}^{q_g} x_{t-j}, \\ e_{g,t-j} &= x_{t-j} - \mu_{g,t-j}. \end{aligned}$$

166 and

$$\begin{aligned} \theta_{g,0} &> 0, \quad \forall g, \\ \theta_{g,j} &\geq 0, \quad \forall g, \quad j = 1, \dots, q_g. \end{aligned}$$

167 The first three specifications were introduced in Berchtold (2003). The case
 168 of a constant standard deviation is included in the first specification by fixing
 169 $q_g = 0$. Since this first specification uses only the past squared observations,
 170 it should be used only on datasets in which a substantial part of the data
 171 is homoscedastic. Indeed, if we compare two time series, S_t^a and S_t^b , such that
 172 $S_t^b = S_t^a + c$, the two series have the same standard deviation, but the parameters
 173 of Equation (2.2) would take different values. The next two specifications of
 174 the standard deviation do not suffer from this problem because they are both
 175 derived from the usual standard deviation formula. The difference lies in the
 176 reference point used. Equation (2.3) directly compares each lag to the empirical
 177 mean. The only restriction here is that the number of lags have to be greater
 178 than or equal to 2. Equation (2.4) compares each lag to the expectation of the
 179 component. As noted by Wong and Li (2000), the latter approach is somewhat
 180 curious, since $\mu_{g,t}$ is not always a good predictor of the next observation when
 181 the series is highly variable. However, in practice, this specification has the
 182 advantage of ensuring consistency between the two time-dependent elements
 183 ($\mu_{g,t}$ and $\sigma_{g,t}$) of each Gaussian component. Finally, Equation (2.5) is the
 184 ARCH specification proposed by Wong and Li (2001).

185 With regard to modeling the expectation, it is also possible to let the stan-
 186 dard deviation of a component depend on external factors, and to introduce
 187 covariates into equations (2.2) to (2.5). Furthermore, when the variability of
 188 the studied process is related to its magnitude, using specification (2.4), with

189 the covariates included in the modeling of $\mu_{g,t}$, could prove particularly useful.

190 In the above equations, p_g denotes the number of lags used when modeling
 191 the expectation of the g -th component, and q_g denotes the corresponding num-
 192 ber of lags used in the modeling of the variance. The larger time dependence
 193 for each component is then $r_g = \max(p_g, q_g)$. Note that, even if $p_g > 1$ or
 194 $q_g > 1$, it is not mandatory to use all lags between 1 and p_g or q_g . In practice,
 195 it can be interesting to associate some lags with only some components. For
 196 instance, in the original GMTD model (Le et al., 1996), the expectation of the
 197 g -th component used only the g -th lag of the observed variable, X . Another
 198 possibility is to constrain the sum of several parameters to take a particular
 199 value. For instance, Le et al. (1996) proposed imposing the following constraint
 200 on a component g :

$$\sum_{i=1}^{p_g} \varphi_{g,i} = 1.$$

201 2.2. The hidden level

202 In the previous subsection, we wrote the weight of the g -th component as a
 203 function of time: $\lambda_g(t)$. Of course, it is possible to remove this dependence and
 204 consider

$$\lambda_g(t) = \lambda_g, \quad \forall t.$$

In this case, the HMTD model reduces to a pure mixture model. On the other
 hand, following Weigend and Shi (2000), there are at least two different pos-
 sibilities for specifying varying weights. The first one, called “gated experts”,
 allows each weight to depend on a set of covariates (Weigend et al., 1995). We
 then have

$$\lambda_t = (\lambda_1(c_1(t)), \lambda_2(c_2(t)), \dots, \lambda_k(c_k(t))).$$

205 Since covariates are supposed to evolve over time, the component weights do
 206 too.

The second possibility is to let the weights at time t depend on the past
 through a Markov chain and, in this way, to implement a HMM at the hidden
 level. We then write

$$\lambda_g(t) = P(S_t = g | S_{t-1}, \dots, S_{t-\ell}).$$

207 where S_t represents the component chosen at time t , and ℓ is the order of the
 208 Markov chain. When the order of the hidden Markov chain is set to zero, the
 209 model once again reduces to a pure mixture.

210 The last two approaches can be combined to let the weights of the compo-
 211 nents depend on both the past of the latent process and on the covariates. This
 212 is accomplished by adding additional terms to the previous equation:

$$\lambda_g(t) = P(S_t = g | S_{t-1}, \dots, S_{t-\ell}, C_1(t), \dots, C_{c_g}(t)). \quad (2.6)$$

Equation (2.6), even if still describing transition probabilities, no longer corresponds to a Markov chain. Moreover, the number of parameters to be estimated can become prohibitive, especially when there is more than one lag. A solution is to model the hidden process through a discrete MTD model (Berchtold and Raftery, 2002), as follows:

$$\lambda_g(t) = \sum_{j=1}^{\ell} \psi_j q_{s_j, g} + \sum_{h=1}^{c_g} f_h(c_h),$$

where $q_{s_j, g}$ denotes the transition probability from state value s_j at time $t - j$ to state value g at time t , and ψ_j is the weight of lag j . Then, the set of all transition probabilities can be written as a transition matrix, Q :

$$Q = \begin{pmatrix} q_{1,1} & \cdots & q_{1,k} \\ \vdots & & \vdots \\ q_{k,1} & \cdots & q_{k,k} \end{pmatrix}$$

Then, $C = \{C_1, \dots, C_k\}$ is a set of covariates and $f_1, \dots, f_k$ are transformation functions. When the covariates are categorical, we can write

$$\sum_{h=1}^{c_g} f_h(c_h) = \sum_{h=1}^{c_g} \gamma_h q_{c_h, g}.$$

where $q_{c_h, g}$ denotes the probability of the transition from the value of the h -th covariate to the state value g at time t , and γ_h is the corresponding coefficient. Of course, the preceding equations have to be constrained in order for the results to be probabilities.

By allowing all possible transitions between hidden states to occur, we can represent the behavior of the observed time-series as precisely as needed. However, it is also sometimes useful to constrain the transition matrix of the hidden level. For instance, consider representing the transitions between successive hidden states using a first-order Markov chain, with no covariates. With four components, the matrix is written as

$$A = \begin{pmatrix} q_{1,1} & q_{1,2} & q_{1,3} & q_{1,4} \\ q_{2,1} & q_{2,2} & q_{2,3} & q_{2,4} \\ q_{3,1} & q_{3,2} & q_{3,3} & q_{3,4} \\ q_{4,1} & q_{4,2} & q_{4,3} & q_{4,4} \end{pmatrix}$$

An interesting case occurs when the different states are hierarchical. In other words, when a subject enters in state s_i , he/she can only stay in this state or move to state s_j , such that $s_j > s_i$ (see the left-hand side of Table 1). For example, in behavioral studies, we observe this kind of situation when the phenomenon evolves with the age of the subjects. Young subjects start in the first hidden state and then switch to higher states as they grow older. Hearing capabilities are a good example. Young people are supposed to have maximal capabilities, which then decline with age, passing through different stages.

Table 1: Transition matrices for hierarchical states

$$A = \begin{pmatrix} q_{1,1} & q_{1,2} & q_{1,3} & q_{1,4} \\ 0 & q_{2,2} & q_{2,3} & q_{2,4} \\ 0 & 0 & q_{3,3} & q_{3,4} \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} q_{1,1} & q_{1,2} & 0 & 0 \\ 0 & q_{2,2} & q_{2,3} & 0 \\ 0 & 0 & q_{3,3} & q_{3,4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Here, the first state would represent people with full hearing capabilities, and the last state could represent those who are deaf. Of course, the probability of remaining in the last state is one. If we want to impose a strict order between the states (one can switch only to the next state), then the transition matrix can be represented as in shown on the right-hand side of Table 1. Another useful specification is to set A to the identity matrix. Each state is then absorbing, meaning that it is impossible to leave a state. This implies that each independently observed subject is associated with one, and only one state during its observation period. The model then classifies the subjects into k mutually exclusive groups based on the sequence of data observed for each subject. Other approaches that use HMMs to classify sequence data are presented in Helske et al. (2010).

From a practical point of view, the constraints discussed above are very easy to set, because the Expectation-Maximization (EM) algorithm used to estimate the parameters of the HMTD model has the property that parameters initialized to zero will not be updated during the optimization process. Hence, it is sufficient to initialize the required parameters to zero.

3. Estimation

Belonging to the class of hidden Markov models, the HMTD model can be estimated using the general framework proposed by Rabiner (1989), in which three different problems have to be considered:

1. The computation of the (log-)likelihood of the sequence of observed data, given the current model.
2. The identification of the optimal sequence of hidden states, given the current model and the sequence of observed data.
3. The estimation of the optimal model parameters, given the sequence of observed data—that is, the parameters that maximize the log-likelihood of the model.

The first problem is solved through an iterative computation known as the Forward-Backward procedure (Rabiner, 1989). The second problem, sometimes called the *decoding problem*, is solved using the Viterbi algorithm (Viterbi, 1967). Lastly, the third problem is solved using a version of the Expectation-Maximization (EM) algorithm especially designed for the Hidden Markov Model, known as the Baum-Welch algorithm (Baum et al., 1970). Each of these three

275 problems has been studied extensively and the interested reader can find de-
 276 tailed discussions in Baum et al. (1970); Dempster et al. (1977); Rabiner (1989);
 277 McLachlan and Krishnan (1996); Berchtold (2002, 2004). Even though the
 278 framework presented in this study incorporates many extensions to the basic
 279 HMM and MTD models, the estimation procedure remains essentially the same.
 280 Only minor adjustments are necessary in order to explicitly incorporate the re-
 281 lationships between successive observations and the effect of the covariates.

282 Although the maximization of the log-likelihood is not difficult from a purely
 283 computational point of view, it is much more difficult to assess the quality of
 284 the solution. Different problems can occur during the estimation phase. For
 285 instance, even though it has been demonstrated that EM algorithms converge
 286 to a maximum of the log-likelihood, there is no way of ensuring that this is the
 287 global maximum, rather than a local optimum. Even for quite simple HMTD
 288 models with few components and lags, the solution space can be very complex
 289 with many local optima. Different solutions have been proposed to overcome
 290 this issue, ranging from several independent runs of the EM algorithm and vari-
 291 ants of the basic algorithm (e.g., the Classification EM and the Stochastic EM
 292 algorithm) to a combination of the EM algorithm and a gradient-type or ge-
 293 netic algorithm. More detailed can be found in Biernacki et al. (2000); Böhning
 294 (2001); Berchtold (2004).

295 Another well-known situation is the possible degeneracy of the log-likelihood,
 296 with the log-likelihood taking values larger than one. This issue is easy to detect
 297 and then to avoid by imposing constraints on the model parameters, especially
 298 on the coefficients of the standard deviation. Finally, when a component is fitted
 299 to very small subsets of the data, the degeneracy in the maximum likelihood
 300 estimation might be due to particular features of those subsets, preventing a
 301 good generalization of the results. To avoid these fallacious situations, we again
 302 constrain some of the parameters, for instance, by limiting the ratio between the
 303 largest and smallest weights of the different components. For further details,
 304 refer to Berchtold (2003).

305 Other difficulties related to mixture models are formulating standard errors
 306 and t-statistics, in order to perform a diagnostic of the parameters estimated
 307 from the EM algorithm, and computing confidence intervals. The simple use of
 308 the expectation is typically not possible, and obtaining the Hessian matrix is not
 309 straightforward. In the literature, two classes of methods have been proposed to
 310 estimate the variance matrix of maximum likelihood estimators: information-
 311 based methods and resampling methods. Information-based methods propose
 312 estimating the variance matrix as the inverse of the information matrix. One
 313 way is to approximate the information matrix from the complete(-data) likeli-
 314 hood (e.g., Louis, 1982) so the mixture membership is treated as known. On
 315 the other hand, Dietz and Böhning (1996) proposed approximating the Fisher
 316 information using the original data (i.e., using the incomplete-data likelihood).
 317 These methods are computationally efficient. However, in both cases, we only
 318 have an asymptotic approximation of the covariance matrix, so the sample size
 319 must be large enough to guarantee that the asymptotic theory of maximum like-
 320 lihood holds. More recently, Boldea and Magnus (2009) derived the analytical

forms of the score vector and Hessian matrix for multivariate Gaussian mixture models, which are used to estimate the observed information matrix directly.

The bootstrap method is a more popular technique for performing model diagnostics in mixture modeling. In this class of methods, we can distinguish three main approaches: the parametric bootstrap (Basford et al., 1997; McLachlan and Peel, 2000), the non-parametric bootstrap (Efron, 1979), and a modified version of the non-parametric bootstrap, proposed by Newton and Raftery (1994), known as the weighted bootstrap. In the latter approach, the data are weighted proportionally to the number of times that a sample values occurs in the bootstrap re-sample. The parametric and non-parametric bootstrap methods differ in the way the re-sampling is performed. In the first case, the repeated draws are made from the fitted mixture (i.e., the mixture with parameters fixed at their estimated values). In the non-parametric approach, the bootstrap samples are drawn directly from the sampling distribution of the original data. Whatever resampling technique is chosen, the bootstrap approach estimates the model in each bootstrap sample, then computes the in-sample bootstrap standard errors of the corresponding parameter estimates using a Monte Carlo approximation (McLachlan and Peel, 2000). Efron and Tibshirani (1994) showed that 50–100 bootstrap replications are sufficient for reliable standard error estimation.

In the case of resampling-based methods, because of the invariance of the likelihood under permutations of the mixture components, a label-switching problem may arise (Redner and Walker, 1984). However, according to McLachlan and Peel (2000), if we use the parameters estimated from the original data as initial values in the EM algorithm in each bootstrap sample, this problem should remain rare. Nevertheless, the label-switching problem has been extensively studied in the literature, and different solutions have been proposed (Stephens, 2000).

Longitudinal analyses in the social sciences generally focus more on panel data than on univariate time series. Therefore, when employing a bootstrap method, two further issues have to be addressed. First, we have to choose how to perform the sampling. Several methods of resampling panel data have been proposed in the literature: temporal sampling, where the resampling is performed at each time point; the block bootstrap, which re-samples blocks of consecutive observations; and individual or cross-sectional sampling, which analyzes individual sequences as unique whole unit in order to preserve intra-individual dependence (Kapetanios, 2008).

The second issue is related to longitudinal panel weights. The resampling must account for the sample weights to avoid a biased estimation, in a way similar to the weighted bootstrap proposed by Newton and Raftery (1994). In survey methodologies, resampling methods from stratified populations have been introduced (e.g., Kovar et al., 1988), but these still need to account for complex sampling designs, as well as deal with non-response and longitudinal panel weights.

364 4. Model selection

365 The great flexibility of the HMTD model can sometimes also be a drawback.
 366 For example, it is difficult to decide on the structure for the “best” model to fit
 367 the data. Should we increase the number or lags, or the number of components?
 368 Should we use more covariates? Should we use a non-constant variance? These
 369 are just some of the options we need to consider. However, before addressing
 370 these questions, it is necessary to consider the concept of a statistical model and
 371 the notion of a “best” model.

372 Modeling a time series can be motivated by different objectives, for instance:

- 373 1. Obtaining the most precise representation of the data under investigation.
- 374 2. Obtaining a model able to represent times series similar to the observed
 375 data.
- 376 3. Building a model that provides the best possible predictions.
- 377 4. Classifying the data into a finite set of mutually exclusive categories.

378 These different objectives are, of course, related. However, even the notion of a
 379 “best” model can differ. In general, we consider that model A describes a given
 380 time series better than model B does when model A has the greater probability
 381 of generating the data. This is equivalent to saying that the (log-)likelihood of
 382 model A , given the data, is larger than the (log-)likelihood of model B . On the
 383 other hand, since it is almost always possible to increase the (log-)likelihood of
 384 a model by simply adding more parameters, the model representing the data in
 385 the most accurate way can be very complex in terms of free parameters that need
 386 to be estimated. In such a case, this model is perhaps not the most desirable for
 387 the purpose of analysis and decision-making. This is why we generally prefer
 388 to select a model based on a penalized criterion, considering both the precision
 389 of the modeling and the complexity. The Bayesian Information Criterion (BIC,
 390 see Schwarz, 1978; Kass and Raftery, 1995; Raftery, 1995) is an example of
 391 such a criterion. Moreover, Leroux (1992) proves the consistency of maximum
 392 (penalized) likelihood estimators for a mixing distribution obtained from the
 393 number of components selected using the AIC or BIC.

394 For a given model, the BIC is defined as

$$BIC = -2LL + P \log(N).$$

395 where LL is the log-likelihood of the model, P is the number of independent
 396 parameters, and N is the number of data used to compute the log-likelihood.
 397 The model with the smallest BIC is chosen. For the log-likelihood, we can
 398 choose between the observed and the complete log-likelihood. In the first case,
 399 we consider the contribution of the visible part of the model, given the hidden
 400 part, and the BIC explains how well the model fits the data. However, by using
 401 the complete log-likelihood, we have the information from both the visible and

402 the hidden part of the model. Since we are interested in assessing the quality
403 of both levels of the HMTD, we use the latter approach.

404 When we are mainly interested in using the model to predict the probability
405 of future events, we have to consider that the future behavior of the series will,
406 generally, not be the same as its past behavior. Consequently, a model that
407 too precisely describes a particular, but rarely observed feature of the data can
408 be disadvantageous. In that case, using the BIC can also be interesting, since
409 the precise description of a particular aspect of a time series generally involves
410 many additional parameters.

411 Another use of mixture models is in classifying data. Here, we assume that
412 each component of the model generates a different kind of data, and we try to
413 allocate each observed data element to one of the components in order to obtain
414 homogeneous clusters. In the context of time series, we usually try to obtain
415 a representation in which groups of several successive data are identified. We
416 then need a model that is sufficiently rough, which will not be too influenced by
417 relatively small variations in the data. In practice, the mixture model is used to
418 associate, a posteriori, each observation to one component (e.g., McLachlan and
419 Basford, 1988), and in particular, to the component with the largest expectation.

420 Whatever the final goal of the modeling, we can adopt either a deductive or
421 an inductive approach. In the deductive case, we start from a clear theory, which
422 the statistical model should reflect. For instance, if the theory tells us that three
423 different behaviors can be adopted by the subjects under study, so we should
424 choose a three component model, one to represent each separate behavior. Next,
425 the theory tells us that the first behavior is memoryless, while the other two
426 can be explained by their past behavior. The first component should then have
427 a constant mean, while the other two should use one or several lags, and so on.
428 If the theory is sufficiently detailed, then we can use it to completely define the
429 structure of the model.

430 Once computed, a theory-driven model can be analyzed and its parameters
431 interpreted. Hypotheses can be tested by computing alternative models that
432 modify aspects of the initial model. These alternative models are then compared
433 to the initial model based on either likelihood ratio tests (Giudici et al.,
434 2000; Dannemann and Holzmann, 2008) or the BIC. If alternative models prove
435 to be superior to the first theory-driven model, then the theory may be ques-
436 tioned. The alternative is to perform an empirical search adopting an inductive
437 approach. In this case, we are not driven by a theory, but instead search for the
438 model that best explains the data under investigation. Given the large number
439 of different models that could possibly be specified from the same dataset, the
440 question then arises of how to define an optimal search strategy. A possibility
441 is to rely on the fact that the HMTD belongs to the family of multilevel models
442 and to adopt an ad hoc strategy, such as the one proposed by Hox (1995). The
443 idea is to work hierarchically, starting from a very basic model, then adding
444 more advanced elements one at a time. After each change, the new model is
445 compared to the previous one using a measure such as the BIC. In the context
446 of a HMTD, this can be summarized as follows:

- 447 1. Begin with a two-component model, a first-order hidden transition matrix,
448 constant expectations and variance for both components, and no covari-
449 ates.
- 450 2. Try to find the optimal number of components by computing additional
451 models with 3, 4, 5, ... components. Note that since no other explanatory
452 elements are used in the model, the number of components could become
453 very large. In this situation, a limit (for instance 5 or 10) should be
454 imposed.
- 455 3. Add lags to the modeling of the expectation of the visible components.
- 456 4. Allow the variance of each component to be non-constant using one of the
457 specifications proposed in Section 2.1.
- 458 5. Replace the first-order transition matrix with a high-order transition ma-
459 trix.
- 460 6. Introduce covariates at the visible level.
- 461 7. Introduce covariates at the hidden level.

462 This strategy calls for several remarks:

- 463 • Since we start with a very simple model in which the only explanatory
464 element lies in the number of visible components, this number could be
465 very large at first. Even if it is restricted to a maximal value in step
466 2, it is likely that introducing additional explanatory elements will later
467 render some of the components useless. A good idea is then to try, at least
468 at the end of each of the aforementioned steps, to reduce the number of
469 components.
- 470 • Beginning with a refinement of the visible model before trying the hidden
471 model is based on the fact that we know the structure of the data at the
472 visible level. However, the structure of the hidden level is known only in
473 distribution. Therefore, we have better control of the visible level, making
474 it more useful to begin with a precise model of the observed level.
- 475 • Covariates are introduced in the model in the last steps of this strategy. In
476 general, longitudinal data are, at least partially, self-explaining. Therefore,
477 we should begin by using the information contained in the data, as much as
478 possible, before using external information. Of course, if the phenomenon
479 under study is known to be strongly related to a particular external factor,
480 for example gender or age, then this factor can be introduced at an earlier
481 stage.

482 Lastly, always keep in mind that a theory-driven model allows more gen-
483 eralization than does an empirical model. In that sense, and whatever the
484 approach used, one should avoid trying models without a real practical inter-
485 pretation. The overall “best” model might not be the model with the optimal

486 BIC value or the one that optimally fits a sample of data. Instead, it is the
 487 model that allows to best understand the phenomenon under study.

488 5. An empirical application

489 In this section, we illustrate an application of the HMTD and the selection
 490 procedure described in Section 4 on unbalanced panel data. The data are from
 491 the core representative sample of the Panel Study of Income Dynamics (PSID).
 492 The PSID is the longest running nationally representative household panel study
 493 and allows unique opportunities to conduct life course research and generational
 494 studies over four decades. The PSID began collecting socio-economic and de-
 495 mographic data in 1968 on a national sample of U.S. households (about 4,800
 496 families) with the objective of studying the dynamics of income and poverty in
 497 the United States (Hill, 1991). In order to allow the sample to remain repre-
 498 sentative of the U.S. population, the PSID follows each family member of the
 499 original 1968 sample, even if they move out of the original household and create
 500 their own new families. This occurred yearly until 1997, then changed to every
 501 two years. As of 2009, 8,690 households had been interviewed, incorporating
 502 24,385 individuals.

503 Our application concerns household income from 1968 to 2009 in the United
 504 States. The family income used in this example reflects income from any source
 505 (labor, assets, transfers, and so on) and from all persons living in the family
 506 unit. Since 1979, the value of family income was topcoded at \$99,999. In 1980,
 507 this increased to \$999,999, and since 1981, has been \$9,999,999. To correctly
 508 compare the evolution over four decades, the data used in this study are adjusted
 509 for inflation in 2008 dollars (Consumer Price Index–U series).

510 During a life course, it is highly likely that a household will suffer a series
 511 of independent economic shocks due to the loss of a job, national economic
 512 downturns, changes in the family structure, and so on. Analyzing the household
 513 income dynamics over almost four decades, the impact of these shocks should
 514 be relevant. Thus, using an appropriate model becomes crucial. For instance,
 515 if we represent the income dynamics with a linear stationary (autoregressive)
 516 process, the time taken to recover from an economic shock is assumed to be the
 517 same for poor and wealthier households. In the literature, one of the alternatives
 518 proposed is to use a nonlinear model in the classic way of quadratic or cubic
 519 functions of lagged income (Lokshin and Ravallion, 2001).

520 The HMTD model, with different possible specifications for the expecta-
 521 tion and the standard deviation of each Gaussian component (see Section 2.1),
 522 constitutes a valid alternative to this classic approach. In addition, using the
 523 HMTD, we can also simultaneously consider the income stratification over time
 524 in terms of the serially inter-temporal correlation between successive income
 525 levels (the level of income today is naturally a good predictor of its value in sub-
 526 sequent years), as well as the problem of population heterogeneity. Interpreting
 527 the hidden variable, S , as different discrete income regimes, issues related to
 528 the heterogeneity in income dynamics can be solved in two ways. The first is
 529 to compute the most likely sequence of hidden wealth states directly, and the

530 transition probabilities between them. The second is to constrain the hidden
 531 transition matrix to the identity matrix, and then to use the HMTD to classify
 532 the households into mutually exclusive homogeneous groups according to their
 533 income trajectories.

534 In order to capture the effects of demographic events and relevant changes in
 535 household composition on the earning dynamics process, we include four socio-
 536 demographic factors in the model: the size of the household, the age and gender
 537 of the head of the household, and his/her educational level. In this illustrative
 538 example, we focus only on the effect of certain socio-demographic characteristics
 539 on the evolution of household income over time, even though other aspects, such
 540 as job history and job sector, may be relevant. The variables are all considered to
 541 be time varying to capture the effect of changes in the household composition
 542 on the earning dynamic process. We assume that a household may suffer a
 543 negative shock when a change occurs involving the head of the household. For
 544 instance, owing to the gender gap in earnings, having a man or a woman as the
 545 head of the household probably affects the overall income of the family.

546 In order to show the relevance of the model using medium- and long-term
 547 panel data, we only include households that have been interviewed at least 10
 548 times over the four decades covered by the PSID.

549 *5.1. Model selection procedure*

550 As underlined in the previous section, a crucial point is the model selection
 551 procedure we use to find the “best” possible model. For illustration purposes,
 552 we use an inductive data-driven approach here. Considering the large amount of
 553 data available in the PSID database and the large number of models that could
 554 be estimated, and to avoid an unnecessary large computation time, we consider
 555 a random subsample of 1,000 U.S. households from the core representative sam-
 556 ple. The dataset used is made up of 1,000 multivariate series of different lengths
 557 (from 10 to 36 observations for each household) representing the (log transfor-
 558 mation of the) household income. To compare the specifications, all models are
 559 computed using the same number of observations with a total of 22,496 data
 560 points.

561 In the following tables, we summarize the model specifications and in order
 562 to assess the quality of the model, we report the complete log-likelihood and
 563 the BIC. Following Raftery (1995), a BIC difference of more than five points
 564 suggests “strong” evidence in favor of the model with the lower BIC.

565 As shown in Equation (2.1), it is possible to assign a different number of lags
 566 when modeling the expectation of each component. This allows components to
 567 have a shorter or longer memory, according to a specific research question, but
 568 for convenience, we set the same number of lags for all components in this
 569 application.

570 In applied research, in particular in social sciences, researchers might be
 571 more interested in an interpretable and generalizable model than in the nu-
 572 merical optimal one. A possibility is to set some constraints on the number
 573 of components and lags in order to have an interpretable model and also to

574 avoid to compute too many different models with different specifications saving
575 computational time. These constraints can be based for instance on previous
576 investigations or on theoretical assumptions.

577 In this illustration, we assume that there are no more than four types of
578 income regimes (i.e. the maximum number of components is four) representing
579 for instance four type of income shocks: idiosyncratic or aggregate, transitory
580 or permanent.

581 We also make the assumption that, from an economic point of view, the
582 number of lags when modelling the expectation for each component should be
583 between two and four. Considering only a time dependence of order one, the
584 current income may be over-influenced by negative short-term shocks. On the
585 other hand, due to business cycle, we assume that income shocks are not likely
586 to persist after four years (e.g. Hamilton, 1989).

587 The idea of the proposed model selection procedure is to work hierarchically,
588 starting from the simplest possible model. So, we try first to find the optimal
589 number of components (i.e. income regimes) without explanatory variables
590 using a first-order hidden transition matrix and a constant variance.

Table 2: Step 1. Finding the optimal number of components

No. Components	p	q	Order HC	No. of parameters	Complete log-likelihood	BIC
2	2	0	1	10	-62,532.2	125,165
3	2	0	1	18	-48,134.8	96,449.1
4	2	0	1	28	-42,105.6	84,491.7

Note: p represents the dependence order of each expectation, and q denotes the number of lags for each standard deviation.

591 Looking at the quality of the models shown in Table (2), and using a hidden
592 chain of order one and two lags for the expectation, the BIC always improves
593 when increasing the number of components. Then, considering the constraint
594 discussed above, the HMTD with the lowest BIC is the one with four components
595 (i.e., a BIC of 84,491.7).

596 By increasing the number of independent parameters through additional
597 components, the log-likelihood always improves, but these additional compo-
598 nents may not bring useful information to the model, resulting in a very marginal
599 change in the log-likelihood. Since at the beginning of the selection procedure
600 the only explanatory element is the number of components, the number of free
601 parameters and then the subsequent penalization term in the BIC are relatively
602 low, making it easier to improve the BIC. For this reason and to avoid un-
603 necessary complicated models, it is a good idea to set some constraints on the
604 components. It might be a limit on the maximum number of components, as in
605 the example, or fixing a threshold in the membership probabilities. In such a
606 way, an additional component that represents only few cases will not be included
607 in the model selection procedure.

608 In the second step, after having fixed the number of components at four, we

Table 3: Step 2. Add lags for the expectation

No. Components	p	q	Order HC	No. of parameters	Complete log-likelihood	BIC
4	2	0	1	28	-42,105.6	84,491.7
4	3	0	1	32	-35,065.2	70,451.1
4	4	0	1	36	-27,871.1	56,103

then increase the time dependence for the expectation of each component. As mentioned before, we also make an assumption on the maximum number of lags for the expectation. From an economic perspective, we assume the existence of a business cycle such that economic shocks should not persist after a cycle of four years. So the maximum number of lags we consider for the expectation is four. Moreover, from the 1997, the PSID data are collected every two years so considering four time points for the expectation we actually cover a period of four to eight years.

According to the BIC reported in Table (3), the most adequate model uses a time dependence of order four for the expectation (i.e., a BIC = 56,103).

Once we have found the best combination of number of components and lags for the expectation, we introduce the dependence between successive observations in the specification of the standard deviation. Then, we increase the order of the hidden transition matrix. For the specification of the standard deviation, we use Equation (2.3), comparing each lag to the empirical mean of the component. As a result of the nonlinearity, during the estimation of the standard deviations, the log-likelihood sometimes diverges owing to the difficulty of identifying good initial conditions for the optimization procedure. Therefore, we include at most two lags. Table (4) shows that the best solution is to include only one lag (i.e., a BIC = 53,685.2).

Table 4: Step 3. Add lags to the modeling of the standard deviation

No. Components	p	q	Order HC	No. of parameters	Complete log-likelihood	BIC
4	4	0	1	36	-27,871	56,103
4	4	1	1	40	-26,642.2	53,685.2
4	4	2	1	44	-26,635.1	53,711.1

Since this step-by-step procedure does not compare all possible models, it might not lead directly to the best possible solution. Before including a higher dependence at the hidden level, it can be interesting to try to reduce the complexity of the model by reducing the number of components and/or the number of lags. In this particular case, any simplification of the model leads to a worse solution, as shown in Table (5), which summarizes different models in reverse order of complexity.

Turning now to the hidden level (Table 6), we notice first that the HMTD performs better than a pure mixture model. The model with a hidden chain

Table 5: Recursive Step: Complexity reduction

No. Components	p	q	Order HC	No. of parameters	Complete log-likelihood	BIC
4	4	1	1	40	-26,642.2	53,685.2
4	3	1	1	36	-27,570	55,501
4	3	0	1	32	-35,065.2	70,451.1
3	4	1	1	27	-42,348.1	84,966.8
3	4	0	1	24	-42,426.9	85,094.4

of order zero has a BIC almost double that of the observed level for a HMTD
of order one. On the other hand, by increasing the hidden order from one to
two, we obtain a significantly better model (i.e., a BIC = 49,403.4), however
the using a third-order model is useless.

Table 6: Step 4. Order of the Hidden Chain

No. Components	p	q	Order HC	No. of parameters	Complete Log-Likelihood	BIC
4	4	1	0	27	-47,669.8	95,610.3
4	4	1	1	40	-26,642.2	53,685.2
4	4	1	2	76	-24,320.9	49,403.4
4	4	1	3	220	-24,258.1	50,720.8

Once we have fixed at two the order of the hidden process and before includ-
ing the effects of external factors, we can again try to reduce the complexity of
the model (see Table 7). However, as before, this step did not lead to a better
BIC, so we stay with the previous model.

Table 7: Recursive Step: Complexity reduction

No. Components	p	q	Order HC	No. of parameters	Complete log-likelihood	BIC
4	4	1	2	76	-24,320.9	49,403.4
4	4	0	2	72	-24,470.8	49,663.1
4	3	1	2	72	-24,619.2	49,959.8
4	3	0	2	68	-25,388.1	51,457.6
3	4	1	2	39	-29,075	58,540.8
3	4	0	2	36	-30,488.1	61,337

The last step of the suggested procedure is to introduce the effect of ex-
ternal covariates. The covariates may be included at the visible level, in the
specification of the mean and standard deviation, and/or at the hidden level
by influencing the transition probabilities between hidden states. As evident
from Equation (2.1), it is possible to analyze the effect of a covariate on one
single component instead of on the overall mixture. This is useful to overcome
problems of multi-collinearity, because using the same covariate on multiple

components might give redundant information to the model. Observing the effect of the covariate on only one component might also be useful when we use the mixture models to classify data or when the analysis is theory-driven, for instance, when we search for the effect of external factors on a specific group. In our example, we include only one covariate at a time when modeling the expectation of each component. We consider four covariates. Two are categorical: having a higher level of education (H Edu) and the gender of the head of the household being male (Male). The other two are continuous: the age of the respondent (Age) and the size of the family (H Size). Table (8) reports the quality of the more relevant models. The best model is that using all four covariates, each covariate being associated with one of the four components (i.e., a BIC = 44,946.8).

Table 8: Step 5. Introducing external covariates

k	p	q	Order HC	No. of parameters	Covariate	Complete log-likelihood	BIC
4	4	1	2	76		-24,320.9	49,403.4
4	4	1	2	77	[1]	-24,126.8	49,025.2
4	4	1	2	77	[2]	-24,167.5	49,106.7
4	4	1	2	77	[4]	-24,310.8	49,393.1
4	4	1	2	78	[1]-[3]	-22,285.1	45,351.6
4	4	1	2	78	[1]-[4]	-22,320.8	45,423.3
4	4	1	2	79	[1]-[3]-[4]	-22,161.1	45,113.8
4	4	1	2	80	[1]-[2]-[3]-[4]	-22,072.5	44,946.8

Note: k is number of components; p the dependence order of each expectation; q the number of lags for each standard deviation.

Covariate: [1] *Male*: Man as head of the household; [2] *H Edu*: Higher level of Education; [3] *Age*: Age of head of the household; [4] *H Size*: Size of Household.

5.2. Model interpretation

In this section, we present an approach to interpret the results of a HMTD. To sum up, for the visible level, we have a mixture of four components, with four lags for the expectation of each component, four covariates, and a time dependence of order one for the standard deviation. The hidden level is driven by a second-order transition matrix. The dependent variable is the (log transformation of the) income levels of a random sub-sample of 1,000 U.S. households from the core representative sample of the PSID database. Four covariates are considered: the size of the household, and the age, gender, and educational level of the head of the household.

The transition matrix between hidden states is as follows (in reduced form):

		S_t			
S_{t-2}	S_{t-1}	1	2	3	4
1	1	0.8200	0.0246	0.1403	0.0152
2	1	0.7594	0.0010	0.2198	0.0198
3	1	0.8661	0.0190	0.0917	0.0233
4	1	0.9195	0.0440	0.0149	0.0216
1	2	0.6566	0.0072	0.3148	0.0214
2	2	0.3368	0.0000	0.6437	0.0195
3	2	0.3987	0.0069	0.5685	0.0259
4	2	0.7504	0.0016	0.2261	0.0219
1	3	0.8629	0.0067	0.1029	0.0275
2	3	0.8542	0.0007	0.1029	0.0422
3	3	0.8240	0.0173	0.1318	0.0269
4	3	0.9031	0.0053	0.0726	0.0190
1	4	0.8563	0.0037	0.0271	0.1128
2	4	0.8692	0.0000	0.0462	0.0846
3	4	0.8620	0.0026	0.0239	0.1115
4	4	0.9482	0.0049	0.0191	0.0278

Here, Q represents the general second-order transition process between hidden states, given the value assumed by the hidden variable in the two previous periods (S_{t-2} , S_{t-1}). The distribution of the first hidden state (π_1) and the second hidden state, conditional on the first ($\pi_{2|1}$), are

$$\pi_1 = (0.7805; 0.0091; 0.1716; 0.0388) \quad \pi_{2|1} = \begin{pmatrix} 0.8412 & 0.0221 & 0.1167 & 0.0200 \\ 0.5356 & 0.0039 & 0.4383 & 0.0222 \\ 0.8611 & 0.0075 & 0.1025 & 0.0289 \\ 0.8839 & 0.0028 & 0.0291 & 0.0842 \end{pmatrix}$$

Both the transition matrix and the distributions of the hidden states indicate the predominance of the first state (and, thus, of the first component at the visible level of the model). The process starts in the first hidden state in 78%. Considering the overall series, the first hidden state is used 93.44% of the time, the second hidden state 0.67%, the third 4.27%, and the fourth 1.61%. For a graphical representation of the most common transitions, see Figure 1 (lower, right corner). With the support of the R package TraMineR (Gabadinho et al., 2011), we can graphically represent the most likely sequence of hidden states and their distributions.

The cross-sectional entropy plotted in Figure 1 (lower, left corner) shows an increasing variability in the distribution of the hidden states among the households over time, and two periods of significant change. The first occurs at the beginning of the series, with the 1970s oil crisis, and the second in the early 1990s, with the economic downturns (1989–1992). Another change in the entropy is observed with the 2008 financial crisis. For the latter two periods, the observed increase in variability may also be due to a high level of attrition.

Figure 1: Graphical representation of hidden states

For the last time point, we have only 383 sequences left of the original 1,000.

The estimated parameters of Equations (2.1) and (2.3) are shown in Table 9 along with the bootstrapped standard errors and the p-values for statistical significance using 1,000 bootstrap replications (Efron and Tibshirani, 1994).

The modelling of the expectation of the first two components is dominated by the first lag, while the second lag dominates the third component, and the third and fourth lags are the most important in the last component, but with opposite effects (positive for the third lag and negative for the fourth lag).

The most common income regime, the first component that represents more than 90% of cases, shows a general slightly positive trend in income dynamics given the positive and relatively high coefficients for the intercept ($\hat{\varphi}_{1,0} = 3.0965$) and first and last lag ($\hat{\varphi}_{1,1} = 0.6789$ and $\hat{\varphi}_{1,4} = 0.2298$). As mentioned before, the second state is however quite rare (it is used only 0.67% of the time) and it represents a fluctuated regime. This component is associated both with an increasing trend in the short run (e.g. $\hat{\varphi}_{2,1} = 3.3473$) and with negative shocks. The third and four components finally implies lower income levels and they are generally associated with short run negative shocks.

Among the covariates ($\hat{\delta}$), gender and educational level seem to have the largest impact on earnings dynamics (see Table 9, components one and four, respectively). In particular, the results show the presence of gender bias. Having a man as head of the household has a positive impact on the income of the family. Not surprisingly, having a high level of education is positively associated with income growth. As expected, the magnitude of the coefficients of the two

Table 9: Estimated parameters for each component

g	1	2	3	4
$\hat{\varphi}_{g,0}$	3.0965 (0.0819) [< .0001]	-3.2835 (0.3610) [.0040]	-0.7567 (0.0815) [< .0001]	0.7426 (1.2834) [.9269]
$\hat{\varphi}_{g,1}$	0.6789 (0.0148) [< .0001]	3.3473 (0.0997) [< .0001]	-0.4644 (0.0177) [.0020]	-0.1578 (0.1034) [.2162]
$\hat{\varphi}_{g,2}$	-0.0493 (0.0099) [.0020]	-0.6291 (0.0562) [< .0001]	2.3375 (0.0367) [< .0001]	1.2624 (0.379) [.0060]
$\hat{\varphi}_{g,3}$	-0.2085 (0.0166) [< .0001]	0.0160 (0.0727) [.5986]	-1.5414 (0.1703) [.0020]	3.3831 (0.4038) [.0020]
$\hat{\varphi}_{g,4}$	0.2298 (0.099) [< .0001]	-1.5077 (0.0884) [< .0001]	0.7424 (0.1776) [.0040]	-3.9031 (0.4863) [.0020]
$\hat{\delta}_{1, Male}$	0.3681 (0.0547) [< .0001]			
$\hat{\delta}_{2, H. Size}$		0.0713 (0.0071) [.2022]		
$\hat{\delta}_{3, Age}$			0.0085 (0.0115) [< .0001]	
$\hat{\delta}_{4, H. Edu}$				1.6369(0.7056) [.0400]
$\hat{\theta}_{g,0}$	0.2364 (0.1850) [< .0001]	0.4523 (0.6901) [< .0001]	0.2498 (0.0474) [< .0001]	8.9801 (2.4186) [< .0001]
$\hat{\theta}_{g,1}$	7.8083 (0.3035) [< .0001]	9.5959 (0.2929) [< .0001]	9.2314 (0.3728) [< .0001]	7.8483 (1.0098) [< .0001]

Note: Bootstrapped standard errors are reported in parentheses. P-values for statistical significance in brackets based on 1,000 bootstrap replications.

Covariate: *Male*: Man as head of the household; *Age*: Age of head of the household; *H. Size*: Size of Household; *H. Edu*: Higher level of Education.

continuous variables is lower, but also positive. Increasing the age of the head of the household increases the overall household income (second component). Note that we included the covariates through a linear additive term. Thus, we neglected the possibility that the relationship between age and earnings follows an inverted U-shape pattern, as is well known in the literature. In other words, earnings could increase in the early years, reach a peak around middle age, and then decline thereafter. These kinds of effects could be captured by using polynomial relationships to model the link between the continuous variables and the variable of interest. Finally, the size of the family seems not to have a statistical significant impact on the income dynamics.

729 6. Discussion

730 In this paper we described a general probabilistic framework for modeling
 731 continuous time-series, and we integrated simultaneously many extensions pre-
 732 viously presented separately in related literature. We also discussed searching
 733 for a useful model from among the possible solutions offered by the combination
 734 of lags and covariates.

735 Due to its flexibility, the HMTD model seems to be particularly suitable for
 736 longitudinal analyses in the social sciences and related fields. The model is able
 737 to consider the observed heterogeneity in the population and can explain the
 738 observed trajectories, making it useful when predicting the next observation in
 739 a series or for probabilistic clustering. As in the Latent Class Growth model
 740 (LCGM), each level of the discrete latent variable may represent a group or a
 741 subtype of cases. However, using the HMTD, we have two alternatives. First,
 742 we can set the transition matrix, Q , as a diagonal matrix to identify distinct
 743 subgroups following a similar pattern in the whole series, as in the LCGM.
 744 Second, we can consider the latent states to be different subpopulations, without
 745 including any constraints on the transition matrix, and allow individuals to move
 746 between latent classes at each time point.

747 The drawback of the flexibility of the HMTD is the difficulty in finding the
 748 correct specification of the model structure: the number of components and lags,
 749 the use of covariates at the hidden and/or observed levels, modeling the standard
 750 deviation of each component, and so on. Given the large number of different
 751 models that could possibly be estimated, we proposed an ad hoc hierarchical
 752 strategy. Starting from the simplest model possible, advanced elements are
 753 added, one at a time and in each step the models are compared using information
 754 criteria such as the BIC. Lastly, we illustrated the model and the suggested
 755 model selection procedure using a real dataset. Using the U.S. Panel Study
 756 of Income Dynamics, we analyzed the trajectories of household income in the
 757 United States over four decades.

758 Another issue related to the complexity of the model is the estimation pro-
 759 cedure. The EM algorithm allows an easy estimation of the parameters, but it
 760 might be quite unstable in high-dimensional settings with many local optima.
 761 Therefore, we require further investigation to test variants of EM algorithms,
 762 such as the Classification EM, Stochastic EM, or genetic algorithms.

763 Future work should empirically investigate the criteria used to identify the
 764 number of components and the model specification in more detail. Numerical
 765 criteria such as the BIC can identify an optimal model from sample data during
 766 the estimation process. However, in empirical applications, we are more inter-
 767 ested in interpretable and generalizable results than in strictly optimal ones. For
 768 instance, adding an additional component that represents only a few outliers or
 769 extreme cases is interesting from a theoretical point of view, but not necessarily
 770 from a practical perspective. These extreme cases might be due to errors in
 771 data entry, drop-out cases, or might represent small, negligible sub-populations.
 772 Further analysis can address this issue by introducing measures other than the
 773 BIC to combine the adequacy of the model with the interpretability of the re-

774 sults. Finally, further research should also consider the case of continuous time
775 data rather than panel data.

776 **Acknowledgments**

777 This publication benefited from the support of the Swiss National Centre of
778 Competence in Research LIVES Overcoming vulnerability: life course perspec-
779 tives, which is financed by the Swiss National Science Foundation. The authors
780 are grateful to the Swiss National Science Foundation for its financial assistance.
781 We also thank the AE and the two referees for their helpful comments.

782 **7. References**

- 783 Bartolucci, F., Farcomeni, A., 2010. A note on the mixture transition distri-
784 bution and hidden Markov models. *Journal of Time Series Analysis* 31 (2),
785 132–138.
- 786 Basford, K. E., Greenway, D., McLachlan, G. J., Peel, D., 1997. Standard errors
787 of fitted means under normal mixture models. *Computational Statistics* 12,
788 1–17.
- 789 Baum, L. E., Petrie, 1966. Statistical inference for probabilistic functions of
790 finite state markov chains. *The Annals of Mathematical Statistics* 37 (6),
791 1554–1563.
- 792 Baum, L. E., Petrie, T., Soules, G., Weiss, N., 1970. A Maximization Technique
793 Occurring in the Statistical Analysis of Probabilistic Functions of Markov
794 Chains. *The Annals of Mathematical Statistics* 41 (1), 164–171.
- 795 Berchtold, A., 1999. The Double Chain Markov Model. *Communications in*
796 *Statistics: Theory and Methods* 28 (11), 2569–2589.
- 797 Berchtold, A., 2002. High-Order Extensions of the Double Chain Markov Model.
798 *Stochastic Models* 18 (2), 193–227.
- 799 Berchtold, A., 2003. Mixture transition distribution (MTD) modeling of het-
800 eroscedastic time series. *Computational Statistics & Data Analysis* 41, 399–
801 411.
- 802 Berchtold, A., 2004. Optimisation of Mixture Models: Comparison of Different
803 Strategies. *Computational Statistics* 19, 385–406.
- 804 Berchtold, A., Raftery, A., 2002. The Mixture Transition Distribution Model
805 for High-Order Markov Chains and Non-Gaussian Time Series. *Statistical*
806 *Science* 17 (3), 328–359.
- 807 Biernacki, C., Celeux, G., Govaert, G., 2000. Stratégies algorithmiques pour
808 maximiser la vraisemblance dans les modèles de mélange. In: *Actes des XXXII*
809 *Journées de Statistique*.

- 810 Böhning, D., 2001. The Potential of Recent Developments in Nonparametric
811 Mixture Distributions. In: Proceedings of the 10th International Symposium
812 on Applied Stochastic Models and Data Analysis.
- 813 Boldea, O., Magnus, J. R., Dec. 2009. Maximum Likelihood Estimation of the
814 Multivariate Normal Mixture Model. *Journal of the American Statistical As-*
815 *sociation* 104 (488), 1539–1549.
- 816 Bollerslev, T., Chou, R. Y., Kroner, F., 1992. ARCH modeling in finance. A
817 review of the theory and empirical evidence. *Journal of Econometrics* 52, 5–59.
- 818 Box, G. E., Jenkins, G. M., C., R. G., 1994. *Time Series Analysis, Forecasting*
819 *and Control*. Prentice Hall.
- 820 Chariatte, V., Berchtold, A., Akre, C., Michaud, P.-A., Suris, J.-C., Jul. 2008.
821 Missed appointments in an outpatient clinic for adolescents, an approach to
822 predict the risk of missing. *The Journal of adolescent health* 43 (1), 38–45.
- 823 Dannemann, J., Holzmann, H., Jun. 2008. Likelihood Ratio Testing for Hid-
824 den Markov Models Under Non-standard Conditions. *Scandinavian Journal*
825 *of Statistics* 35 (2), 309–321.
- 826 Dempster, A. P., Lard, N. M., Rubin, D. B., 1977. Maximum Likelihood from
827 Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical So-*
828 *ciety. Series B* 39 (1), 1–38.
- 829 Dietz, E., Böhning, D., 1996. Statistical inference based on a general model of
830 unobserved heterogeneity. In: Fahrmeir, L., Francis, F., Gilchrist, R., Tutz,
831 G. (Eds.), *Lecture Notes in Statistics: Advances in GLIM and Statistical*
832 *Modeling*. Springer Berlin Heidelberg, pp. 75–82.
- 833 Efron, B., 1979. Bootstrap methods: another look at the jackknife. *The Annals*
834 *of Statistics* 7 (1), 1–26.
- 835 Efron, B., Tibshirani, R. J., 1994. *An Introduction to the Bootstrap*. CRC Press.
- 836 Elliott, R. J., Hunterb, W. C., Jamieson, B. M., 1998. Drift and volatility
837 estimation in discrete time. *Journal of Economics Dynamics and Control* 22,
838 209–218.
- 839 Frydman, H., Schuermann, T., 2008. Credit Rating Dynamics and Markov Mix-
840 ture Models. *Journal of banking and finance* 32, 1062–1075.
- 841 Gabadinho, A., Ritschard, G., Studer, M., 2011. Analyzing and Visualizing
842 State Sequences in R with TraMineR. *Journal of Statistical Software* 40 (4).
- 843 Giudici, P., Rydén, T., Vandekerckhove, P., 2000. Likelihood-ratio tests for hid-
844 den Markov models. *Biometrics* 56 (3), 742–7.
- 845 Hamilton, J. D., 1989. A New Approach to the Economic Analysis of Nonsta-
846 tionary Time Series and the Business Cycle. *Econometrica* 57 (2), 357–384.

- 847 Hamilton, J. D., 1994. Time Series Analysis. Princeton University Press.
- 848 Hassan, M. Y., El-Bassiouni, M. Y., 2013. Modelling Poisson marked point pro-
849 cesses using bivariate mixture transition distributions. *Journal of Statistical*
850 *Computation and Simulation* 83 (8), 1440–1452.
- 851 Hassan, M. Y., Lii, K.-S., 2006. Modeling Marked Point Processes via Bivariate
852 Mixture Transition Distribution Models. *Journal of the American Statistical*
853 *Association* 101 (475), 1241–1252.
- 854 Hayashi, T., 2004. A discrete-time model of high-frequency stock returns. *Quan-*
855 *titative finance* 4, 140–150.
- 856 Helske, J., Eerola, M., Tabus, I., 2010. Minimum description length based hid-
857 den Markov model clustering for life sequence analysis. In: *Proceedings of the*
858 *third workshop on information theoretic methods in science and engineering*.
- 859 Hill, M., 1991. *The Panel Study of Income Dynamics: a User’s Guide*. SAGE
860 Publications.
- 861 Hox, J. J., 1995. *Applied Multilevel Analysis*. Amsterdam TT-Publikaties.
- 862 Kapetanios, G., Jul. 2008. A bootstrap procedure for panel data sets with many
863 cross-sectional units. *Econometrics Journal* 11 (2), 377–395.
- 864 Kass, R. E., Raftery, A. E., 1995. Bayes Factors. *Journal of the American Sta-*
865 *tistical Association* 90 (430), 773–795.
- 866 Kim, D., Kon, S. J., 1994. Alternative Models for the Conditional Heteroscedas-
867 ticity of Stock Returns. *The Journal of Business* 67 (4), 563–598.
- 868 Kon, S. J., 1984. Models of Stock Returns: A comparison. *The Journal of*
869 *Finance* 39 (1), 147–165.
- 870 Kovar, J. G., Rao, J. N. K., Wu, C. F. J., Aug. 1988. Bootstrap and Other
871 Methods to Measure Errors in Survey Estimates. *The Canadian Journal of*
872 *Statistics / La Revue Canadienne de Statistique* 16, 25.
- 873 Le, N. D., Martin, D. R., Raftery, A. E., 1996. Modelling flat stretches, bursts,
874 and outliers in time series using mixture transition distribution models. *Jour-*
875 *nal of the American Statistical Association* 91 (436), 1504–1515.
- 876 Le Strat, Y., Carrat, F., Dec. 1999. Monitoring epidemiologic surveillance data
877 using hidden Markov models. *Statistics in medicine* 18 (24), 3463–78.
- 878 Leroux, B. Y. B. G., 1992. Estimation of a Mixing Distribution. *The Annals of*
879 *Statistics* 20 (3), 1350–1360.
- 880 Lokshin, M., Ravallion, M., 2001. Household Income Dynamics in Two Transi-
881 tion Economies. World Bank, 1–40.

- 882 Louis, T. A., 1982. Finding the Observed Information Matrix when Using the
883 EM Algorithm. *Journal of the Royal Statistical Society. Series B (Method-*
884 *ological)* 44 (2), 226–233.
- 885 Luo, J., Qiu, H.-b., Dec. 2009. Parameter estimation of the WMTD model.
886 *Applied Mathematics-A Journal of Chinese Universities* 24 (4), 379–388.
- 887 McLachlan, G., Peel, D., 2000. *Finite Mixture Models*. Wiley Series in Proba-
888 *bility and Statistics*.
- 889 McLachlan, G. J., Basford, K. E., 1988. *Mixture Models: Inference and Appli-*
890 *cations to Clustering*. Marcel Dekker Inc.
- 891 McLachlan, G. J., Krishnan, T., 1996. *The EM Algorithm and Extensions*. John
892 *Wiley & Sons*, New York.
- 893 Muthen, B. O., 2001. Second-generation structural equation modeling with com-
894 *bination of categorical and continuous latent variables: New opportunities for*
895 *latent class/latent growth modeling*. In: Collins, L. M., Sayer, A. (Eds.), *New*
896 *methods for the analysis for change*. Washington D.C.: American Psycholog-
897 *ical Association.*, pp. 291–322.
- 898 Netzer, O., Lattin, J. M., Srinivasan, V., 2008. A Hidden Markov Model of
899 *Customer Relationship Dynamics*. *Marketing Science* 27 (2), 185–204.
- 900 Newton, M. A., Raftery, A. E., 1994. Approximate Bayesian Inference with
901 *the Weighted Likelihood Bootstrap*. *Journal of the Royal Statistical Society.*
902 *Series B (Methodological)* 56 (1), 3–48.
- 903 Rabiner, L. I., 1989. A tutorial on Hidden Markov Models and Selected Appli-
904 *cations in Speech Recognition*. *Proceedings of the IEEE* 77, 257–286.
- 905 Raftery, A. E., 1985. A Model for High-order Markov Chains. *Journal of the*
906 *Royal Statistical Society. Series B* 47 (3), 528–539.
- 907 Raftery, A. E., 1995. Bayesian Model Selection in Social Research. *Sociological*
908 *Methodology* 25, 111–163.
- 909 Redner, R. A., Walker, H. F., 1984. Mixture densities, maximum likelihood and
910 *the EMalgorithm*. *SIAM Rev.* 26, 195–239.
- 911 Schwarz, G. E., 1978. Estimating the dimension of a model. *Annals of Statistics*
912 6 (2), 461–464.
- 913 Schlattmann, P., 2009. *Medical Applications of Finite Mixture Models*. Springer:
914 *Statistics for Biology and Health*.
- 915 Shirley, K. E., Small, D. S., Lynch, K. G., Maisto, S. A., Oslin, D. W., Mar.
916 2010. Hidden Markov models for alcoholism treatment trial data. *The Annals*
917 *of Applied Statistics* 4 (1), 366–395.

- 918 Stephens, M., 2000. Dealing with label switching in mixture models. Journal
919 of the Royal Statistical Society. Series B (Statistical Methodology) 62 (4),
920 795–809.
- 921 Viterbi, A. J., 1967. Error Bounds for Convolutional Codes and an Asymp-
922 totically Optimum Decoding Algorithm. IEEE Transactions on Information
923 Theory 16 (2), 260–269.
- 924 Weigend, a. S., Mangeas, M., Srivastava, a. N., 1995. Nonlinear gated experts
925 for time series: discovering regimes and avoiding overfitting. International
926 journal of neural systems 6 (4), 373–99.
- 927 Weigend S.W., Shi S., 2000. Predicting Daily Probability Distributions of
928 S&P500 Returns. Journal of Forecasting, 19, 375–392.
- 929 Wellekens, C., 1987. Explicit time correlation in Hidden Markov Models for
930 speech recognition. Proceedings ICASSP, 384–386.
- 931 Wong, C. S., Chan, W. S., 2005. Mixture Gaussian Time Series Modelling of
932 Long-Term Market Returns. North American Actuarial Journal.
- 933 Wong, C. S., Li, W. K., 2000. On a mixture autoregression model. Journal of
934 the Royal Statistical Society. Series B 62, 92–115.
- 935 Wong, C. S., Li, W. K., 2001. On a mixture autoregressive conditional het-
936 eroscedastic model. Journal of the American Statistical Association 96, 982–
937 995.

Figure

Hidden States – First 25 sequence

State distribution plot

Cross sectional entropy

Most common transitions