Hyppolite *Journal of Statistical Distributions and Applications* (2017) 4:9
DOI 10.1186/s40488-017-0066-3

Journal of Statistical Distributions and Applications

RESEARCH Open Access

Alternative approaches for econometric modeling of panel data using mixture distributions

Judex Hyppolite

Correspondence:
jhyppoli@monmouth.edu
Department of Economics, Finance, and Real Estate, Monmouth University, 400 Cedar Avenue, 07776, West Long Branch, New Jersey, USA

Abstract

The economic researcher is sometimes confronted with panel datasets that come from a population made of a finite number of subpopulations. Within each subpopulation the individuals may also be heterogenous according to some unobserved characteristics. A good understanding of the behavior of the observed individuals may then require the ability to identify the groups to which they belong and to study their behavior across groups and within groups. This may not be a complicated exercise when a group indicator variable is available in the dataset. However, such a variable may not be included in the dataset; and as a result, the econometrician is forced to work with the marginal distribution of the observed response variable, which takes the form of a mixture distribution.

One can model a given response variable with a variety of mixture distributions. In this paper, I present several related mixture models. The most flexible one is an extension of the model by Kim et al. (2008) to the panel data setting.

I have reviewed the estimation of some of these models by the Expectation-Maximization (EM) algorithm. The intent is to exploit the nice convergence properties of this algorithm when it is difficult to find good starting values for a Newton-type algorithm. I have also discussed how to compare these models and ultimately identify the one that provides the best fit to the data set under investigation. As an application I examine the investment behavior of U.S. manufacturing firms.

Keywords: Panel data, Mixture of distributions, Hidden Markov models, Heterogeneity

Introduction

To model the heterogeneity of economic agents I present a series of panel data mixture models of increasing degree of flexibility and complexity and show how they can be used to handle at least two types of heterogeneity: heterogeneity with respect to group membership, and heterogeneity with respect to within group differences in individual characteristics. I have also reviewed the methods of estimation of some of these models via the Expectation-Maximization algorithm. The objective is to take advantage of the nice convergence properties of this algorithm when it is difficult to find good starting values for a newton-type algorithm. I have also reviewed some statistical tests that can be used to choose the best models among those discussed in this paper.

Heterogeneity is an important problem faced by the statistician or the econometrician trying to infer the behavior of economic agents from available data sets. Economic decison makers are heterogeneous in their characteristics and they usually operate in heterogeneous (different) environments. As a result, their behavior generate data whose distributions are sometimes difficult to approximate with the traditional single component econometric models. To deal with this problem, often economists divide their sample into groups using observed variables such as time (in time series) or other individual characteristics (in time series and longitudinal data). The groups obtained this way are usually static and may differ from alternative groups obtained using different observed variables.

While this strategy may allow the researchers to draw some useful conclusions, it is less attractive than the approach that uses multiple characteristics for determining group membership. It is also less flexible than the approach that allows for the possibility that an individual changes group membership depending on the evolution of his characteristics and of the conditions that he is facing. Lastly, it is much less flexible than the approach that offers a unified way (one step method) to make inference about both group membership and behavior. Mixture of distribution models offer such flexibility. These models are justified not only in theory, because they offer a nice way to model heterogeneity, but also in practice since they can be used to provide a semi-parametric approximation to the non-standard distributions of some economic variables at a reasonable cost (McLachlan and Peel 2000). Mixture of distributions are in fact at the crossroad between parametric and non parametric families of distributions. They are parametric because each component distribution usually belongs to a parametric family of distributions, and they are non-parametric because it is possible to provide a very good approximation to the distribution of some variables by increasing the number of components of the mixture (Fink 2007).

Among economic variables whose study can benefit from the applications of mixture distributions one can cite firms' investment, households consumption, money demand, household use of healthcare, etc. Finite mixture distributions are commonly used in Econometrics, mainly in cross-sectional and time series analyses. Following Hamilton (1988), some versions of the hidden Markov models have been extensively used in macroeconometrics to model business cycle fluctuations under the name of Markov Switching regression models. Nevertheless, applications of mixture of distributions in the panel data setting appear to be limited. In many cases the panel data set is treated almost the same way as a cross section. In some rare cases, as in Deb and Trivedi (2013) the dependence of the observations within each unit is modeled using individual specific effects. However, if the panel data set is viewed as a collection of time series it is not difficult to extend the hidden Markov models used in time series analysis to the panel data setting. This is the point of view adopted in this paper and also by Asea and Blomberg (1998) as well as Atman (2007) and Maruotti (2007). The most flexible models presented in this paper extends the times series model by Kim et al. (2008). I allow the Markov chains to be time-inhomogeneous and non-stationary and I introduce within group heterogeneity in the component distributions using the specification by Mundlak (1978). The models are closer to the models by Atman (2007) and Maruotti (2007). A related set of models applied to Panel Count data can also be found in Trivedi and Hyppolite (2012).

The models

Several alternative mixture distributions can be used to model the bivariate process constituted by an economic agent's decision and its group membership. In the following sections, nine such models are described going from the simplest to the most complicated. All of the models are assumed to be made of two components, but extension to more than two components is not difficult.

The models can be used to study several different economic phenomena such as households consumption under financial constraints, firms investment under financial constraints, households demand for money, household use of healthcare, etc. In what follows I will use the example of investment choices under financial constraints to motivate the specifications.

A finite mixture model with constant mixing proportions ($\mathcal{M}_1$)

Consider the vector of random variables $(Y_{it}, W_{it})'$ where Y_{it} represents agent i's decision at time t $(; t = 1, \ldots, T_i; i = 1, \ldots n)$ while W_{it} is a discrete random variable

$$w_{it} = \begin{cases} 1 \text{ if agent } i \text{ belongs to group 1 at time } t \\ 2 \text{ otherwise} \end{cases}$$

In a model about firms' investment decisions under financing constraints, Y_{it} would represent firm i's investment rate at time t, while W_{it} would be the firm's financial status at that time. Y_{it} and W_{it} are assumed to be dependent in the sense that the agent's decision depends on the group he belongs to; more precisely I assume that

$$\begin{aligned} f\left(y_{it}|w_{it}=1;\boldsymbol{\beta}_1,\sigma_1\right) &= \phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right) \\ f\left(y_{it}|w_{it}=2;\boldsymbol{\beta}_2,\sigma_2\right) &= \phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2\right), \\ \sigma_1 > 0, \sigma_2 > 0, & \end{aligned}$$

where $\phi(.)$ is the density function of a univariate normal distribution and $\boldsymbol{x}_{it}$ is a row vector of covariates including individual characteristics that influence the agent's decisions, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are column vectors of parameters. The joint density of $(y_{it}, w_{it})'$ is given by

$$f\left(y_{it}, w_{it};\boldsymbol{\beta}_\nu,\sigma_\nu\right) = p(w_{it}=\nu)\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_\nu,\sigma_\nu\right), \nu = 1, 2,$$

and the marginal density of y_{it} is

$$f\left(y_{it};\boldsymbol{\beta}_1,\sigma_1,\boldsymbol{\beta}_2,\sigma_2\right) = p(w_{it}=1)\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right) + p(w_{it}=2)\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2\right).$$

Let

$$\boldsymbol{\theta} = \left(\pi,\boldsymbol{\beta}_1,\sigma_1,\boldsymbol{\beta}_2,\sigma_2\right).$$

When W_{it} follows a Bernoulli distribution with parameter, π, the marginal density becomes

$$f\left(y_{it};\boldsymbol{\theta}\right) = \pi\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right) + (1-\pi)\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2\right).$$

This is a classical finite mixture of distributions with constant weights π and $1-\pi$.

Parameters Estimation

The parameters of the preceding model can be estimated using maximum likelihood. The complete-data likelihood is

$$L^c(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{t=1}^{T_i}\left(\pi\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right)\right)^{\mathbb{I}(w_{it}=1)}\left((1-\pi)\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2\right)\right)^{1-\mathbb{I}(w_{it}=1)},$$

while the marginal likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{t=1}^{T_i}\left(\pi\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right) + (1-\pi)\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2\right)\right).$$

Since W_{it} is missing, maximizing the marginal likelihood appears to be the most natural estimation approach. However, the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) offers a much simpler alternative. This algorithm maximizes the complete-data likelihood after augmenting the data for the missing variable W_{it} during the expectation step.

The two main steps of the algorithm are the following:

- E-Step (Expectation Step)
  During this step, an intermediate quantity

  $$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right) = \mathbb{E}_{w_{it}}\left(\log\left(L^c(\boldsymbol{\theta})|\boldsymbol{\theta}'\right)\right)$$

- M-setp (Maximization step)
  during which the following maximization problem is solved

  $$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right)$$

  Subject to:

  Various appropriate constraints.

For the model considered here

$$\begin{aligned} Q\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right) &= \mathbb{E}_{w_{it}}\left(log\left(L^c(\boldsymbol{\theta})|\boldsymbol{\theta}'\right)\right) \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(1-\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\right)\left(\ln(1-\pi)+\ln\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2\right)\right) \\ &\quad + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\left(\ln\pi+\ln\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right)\right). \end{aligned}$$

Defining

$$\sqrt{\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)}y_{it} = y_{it}^{(1)} \tag{1}$$

$$\sqrt{\left(1-\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\right)}y_{it} = y_{it}^{(2)} \tag{2}$$

$$\sqrt{\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)}\boldsymbol{x}_{it} = \boldsymbol{x}_{it}^{(1)} \tag{3}$$

$$\sqrt{\left(1-\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\right)}\boldsymbol{x}_{it} = \boldsymbol{x}_{it}^{(2)} \tag{4}$$

and,

$$\left(y_{11}^{(1)},\ldots,y_{1T_1}^{(1)},\ldots,y_{nT_n}^{(1)}\right)' = \mathbf{y^{(1)}} \tag{5}$$

$$\left(y_{11}^{(2)},\ldots,y_{1T_1}^{(2)},\ldots,y_{nT_n}^{(2)}\right)' = \mathbf{y^{(2)}} \tag{6}$$

$$\left(w_{11}^{(1)},\ldots,w_{1T_1}^{(1)},\ldots,w_{nT_n}^{(1)}\right)' = \mathbf{w^{(1)}} \tag{7}$$

$$\left(w_{11}^{(2)},\ldots,w_{1T_1}^{(2)},\ldots,w_{nT_n}^{(2)}\right)' = \mathbf{w^{(2)}} \tag{8}$$

$$\left(\boldsymbol{x}_{11}^{(1)},\ldots,\boldsymbol{x}_{1T_1}^{(1)},\ldots,\boldsymbol{x}_{nT_n}^{(1)}\right)' = \boldsymbol{x}^{(1)}, \tag{9}$$

$$\left(\boldsymbol{x}_{11}^{(2)},\ldots,\boldsymbol{x}_{1T_1}^{(2)},\ldots,\boldsymbol{x}_{nT_n}^{(2)}\right)' = \boldsymbol{x}^{(2)} \tag{10}$$

the expected complete-data log-likelihood can be written as

$$\begin{aligned}
Q\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right) = & \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(1-\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\right)\ln\left(1-\pi\right) \\
& + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right) ln\pi \\
& - \frac{\ln\sigma_2^2}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(1-\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\right) \\
& - \frac{\ln\sigma_1^2}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right) \\
& - \frac{1}{2\sigma_2^2}\left(\mathbf{y^{(2)}}-\mathbf{x^{(2)}}\boldsymbol{\beta}_2\right)'\left(\mathbf{y^{(2)}}-\mathbf{x^{(2)}}\boldsymbol{\beta}_2\right) \\
& - \frac{1}{2\sigma_1^2}\left(\mathbf{y^{(1)}}-\mathbf{x^{(1)}}\boldsymbol{\beta}_1\right)'\left(\mathbf{y^{(1)}}-\mathbf{x^{(1)}}\boldsymbol{\beta}_1\right).
\end{aligned}$$

After solving the system of equations derived from the first order conditions we get

$$\begin{aligned}
\hat{\pi} &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)}{\sum_{i=1}^{n}T_i} \\
\hat{\boldsymbol{\beta}}_1 &= \left(\left(\mathbf{x^{(1)}}\right)'\mathbf{x^{(1)}}\right)^{-1}\left(\mathbf{x^{(1)}}\right)'\mathbf{y^{(1)}} \\
\hat{\boldsymbol{\beta}}_2 &= \left(\left(\mathbf{x^{(2)}}\right)'\mathbf{x^{(2)}}\right)^{-1}\left(\mathbf{x^{(2)}}\right)'\mathbf{y^{(2)}} \\
\hat{\sigma_1^2} &= \frac{\left(\mathbf{y^{(1)}}-\mathbf{x^{(1)}}\hat{\boldsymbol{\beta}}_1\right)'\left(\mathbf{y^{(1)}}-\mathbf{x^{(1)}}\hat{\boldsymbol{\beta}}_1\right)}{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbb{E}_{w_{it}}(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}')} \\
\hat{\sigma_2^2} &= \frac{\left(\mathbf{y^{(2)}}-\mathbf{x^{(2)}}\hat{\boldsymbol{\beta}}_2\right)'\left(\mathbf{y^{(2)}}-\mathbf{x^{(2)}}\hat{\boldsymbol{\beta}}_2\right)}{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(1-\mathbb{E}_{w_{it}}(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}')\right)}.
\end{aligned}$$

Once we get an estimate for $\mathbb{E}_{w_{it}}(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}')$, computing the preceding estimators is simple. In fact,

$$\begin{aligned}\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right) &= \text{prob}\left(w_{it}=1|y_{it};\boldsymbol{\theta}'\right)\\ &= \frac{\text{prob}(w_{it}=1)\times f\left(y_{it}|w_{it}=1;\boldsymbol{\theta}'\right)}{f\left(y_{it};\boldsymbol{\theta}'\right)}\\ &= \frac{\pi\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right)}{\pi\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right)+(1-\pi)\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2\right)}.\end{aligned}$$

So, if we know π, $\left(\boldsymbol{\beta}_2,\sigma_2\right)$ and $\left(\boldsymbol{\beta}_1,\sigma_1\right)$ we can find an estimate for $\mathbb{E}\left(w_{it}|y_{it};\boldsymbol{\theta}'\right)$. The EM algorithm for this model can be summarized as follows:

1. Choose initial values $\boldsymbol{\theta}^0=\left(\pi^0,\boldsymbol{\beta}_1^0,\sigma_1^0,\boldsymbol{\beta}_2^0,\sigma_2^0\right)$
2. Compute $E\left(w_{it}|y_{it};\boldsymbol{\theta}^0\right)$ for each observation
3. Substitute $E\left(w_{it}|y_{it};\boldsymbol{\theta}^0\right)$ in the complete-data log-likelihood
4. Find new values for the parameters $\boldsymbol{\theta}^1=\left(\pi^1,\boldsymbol{\beta}_1^1,\sigma_1^1,\boldsymbol{\beta}_2^1,\sigma_2^1\right)$ by maximizing the complete-data likelihood
5. Compute error $=\frac{|L(\boldsymbol{\theta}^1)-L(\boldsymbol{\theta}^0)|}{|L(\boldsymbol{\theta}^0)|}$
6. If error is higher than a chosen tolerance level, repeat step 2 with the last estimates for the parameters
7. Otherwise, stop; the last estimates are the maximum likelihood estimates.

This algorithm is attractive not only because it provides an intuitive interpretation of the estimation, but also because of its monotone and global convergence properties. It has been proved (McLachlan and Krishman 1997) that the log-likelihood is non-decreasing at each consecutive iteration. This property is very useful for detecting programming errors. Moreover, the global convergence property allows for more flexibility in the choice of starting values than is possible with a Newton-type algorithm.

However, the EM Algorithm is criticized not only because it converges at a low rate, but also because it does not supply automatically an estimate of the covariance matrix of the parameters (McLachlan and Krishman 1997). The Hessian necessary to obtain an estimate of the information matrix in the maximum likelihood setting is not used in the computations. There have been several solutions proposed in the literature to solve this problem. The most notable one is provided by Louis (1982).

Note that according to this model the probability that an economic agent belongs to a certain group remains the same every period. In a dynamic economic environment this assumption is too restrictive. For example, the financial status of a firm cannot be determined by flipping a coin; it is more likely to be dependent on the firm's performance, its characteristics and on the economic conditions it is facing. Thus, several observed variables should help in determining group membership. So, a more realistic model should allow for covariates dependent mixing proportions.

A finite mixture model with smoothly varying mixing proportions ($\mathcal{M}_2$)

Suppose

$$w_{it}=\begin{cases}1 \text{ if } w_{it}^*>0\\ 2 \text{ otherwise}\end{cases}$$

where

$$w_{it}^{*} = \mathbf{z}_{it}\boldsymbol{\gamma} - \epsilon_{it}, \epsilon_{it} \sim N(0,1) \tag{11}$$

$\mathbf{z}_{it}$ is a row vector of covariates that impact the probability for an agent i to belong to a certain group and $\boldsymbol{\gamma}$ is a column vector of parameters.

The group membership equation (Eq. 11) could be modeled with the logistic distribution. Since I want to compare all the models, I would also need to model endogeneity in the same setting and this is not straightforward. It is then better to use the normal distribution and take advantage of the nice properties of the conditional distributions of a partitioned normal random vector. Let

$$\boldsymbol{\theta} = \left(\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \sigma_1, \boldsymbol{\beta}_2, \sigma_2\right).$$

The joint density of (Y_{it}, W_{it}) is

$$\begin{aligned} f(y_{it}, w_{it}; \boldsymbol{\theta}) &= \begin{cases} p(w_{it}=1)\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_1, \sigma_1) \\ p(w_{it}=2)\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_2, \sigma_2) \end{cases} \\ &= \begin{cases} p(\epsilon_{it} < \mathbf{z}_{it}\boldsymbol{\gamma})\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_1, \sigma_1) \\ p(\epsilon_{it} \geq \mathbf{z}_{it}\boldsymbol{\gamma})\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_2, \sigma_2) \end{cases} \\ &= \begin{cases} \Phi(\mathbf{z}_{it}\boldsymbol{\gamma})\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_1, \sigma_1) \\ (1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}))\,\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_2, \sigma_2) \end{cases}, \end{aligned}$$

and the marginal density is

$$f(y_{it}; \boldsymbol{\theta}) = \Phi(\mathbf{z}_{it}\boldsymbol{\gamma})\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_1, \sigma_1) + (1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}))\,\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_2, \sigma_2),$$

where $\Phi(.)$ is the univariate cumulative distribution function of a standard normal random variable.

Parameters Estimation

This model can also be estimated using the EM algorithm. The complete-data likelihood is

$$L^{c}(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{t=1}^{T_i} \left(\Phi(\mathbf{z}_{it}\boldsymbol{\gamma})\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_1, \sigma_1)\right)^{\mathbb{I}(w_{it}=1)} \left((1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}))\,\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_2, \sigma_2)\right)^{1-\mathbb{I}(w_{it}=1)}$$

and the marginal likelihood

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{t=1}^{T_i} \left(\Phi(\mathbf{z}_{it}\boldsymbol{\gamma})\phi\left(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_1, \sigma_1\right) + (1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}))\,\phi\left(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_2, \sigma_2\right)\right).$$

The intermediate EM quantity is

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}') &= E_{w_{it}}(log(L^{c}(\boldsymbol{\theta})|\boldsymbol{\theta}')) \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T_i} \left(1 - \mathbb{E}_{w_{it}}(\mathbb{I}(w_{it}=1)|y_{it}; \boldsymbol{\theta}')\right) \left(\ln\left(1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}\right) + \ln\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_2, \sigma_2)\right) \\ &\quad + \sum_{i=1}^{n}\sum_{t=1}^{T_i} \mathbb{E}_{w_{it}}(\mathbb{I}(w_{it}=1)|y_{it}; \boldsymbol{\theta}') \left(\ln\Phi(\mathbf{z}_{it}\boldsymbol{\gamma}) + \ln\phi(y_{it}; \boldsymbol{x}_{it}\boldsymbol{\beta}_1, \sigma_1)\right). \end{aligned}$$

Using Eqs. (1) - (10), the intermediate EM quantity can be rewritten as

$$\begin{aligned}
Q(\boldsymbol{\theta},\boldsymbol{\theta}') = & \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(1-\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\right)\ln\left(1-\Phi(\mathbf{z}_{it}\boldsymbol{\gamma}\right) \\
& + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\ln\Phi(\mathbf{z}_{it}\boldsymbol{\gamma}) \\
& - \frac{\ln\sigma_2^2}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(1-\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\right) \\
& - \frac{\ln\sigma_1^2}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right) \\
& - \frac{1}{2\sigma_2^2}\left(\mathbf{y^{(2)}}-\mathbf{x}^{(2)}\boldsymbol{\beta}_2\right)'\left(\mathbf{y^{(2)}}-\mathbf{x}^{(2)}\boldsymbol{\beta}_2\right) \\
& - \frac{1}{2\sigma_1^2}\left(\mathbf{y^{(1)}}-\mathbf{x^{(1)}}\boldsymbol{\beta}_1\right)'\left(\mathbf{y^{(1)}}-\mathbf{x^{(1)}}\boldsymbol{\beta}_1\right).
\end{aligned}$$

The first order conditions for the maximization of $Q(\boldsymbol{\theta},\boldsymbol{\theta}')$ will not produce a closed form solution for $\boldsymbol{\gamma}$, but the estimators for $\left(\boldsymbol{\beta}_1,\sigma_1,\boldsymbol{\beta}_2,\sigma_2\right)$ are the same as before. The intermediate EM quantity being separable in the different group of parameters, $\hat{\boldsymbol{\gamma}}$ can be obtained separately using the Newton-type method:

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}} \Bigg(& \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(1-\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\right)\ln\left(1-\Phi(\mathbf{z}_{it}\boldsymbol{\gamma}\right) \\
& + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right)\ln\Phi(\mathbf{z}_{it}\boldsymbol{\gamma})\Bigg)
\end{aligned}$$

Also

$$\mathbb{E}_{w_{it}}\left(\mathbb{I}(w_{it}=1)|y_{it};\boldsymbol{\theta}'\right) = \frac{\Phi(\mathbf{z}_{it}\boldsymbol{\gamma})\phi(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1)}{\Phi(\mathbf{z}_{it}\boldsymbol{\gamma})\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1\right)+(1-\Phi(\mathbf{z}_{it}\boldsymbol{\gamma}))\phi\left(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2\right)}$$

The EM algorithm can be implemented exactly as before.

An Endogenous Switching Regression Model ($\mathcal{M}_3$)

The preceding models are not as common in economics as they are in statistics. The more general model known as switching regression model seems to be preferred. The latter model has been used in several papers in the literature about firms investment. One reason why this model is more popular in econometrics may be, as signaled by Kim et al. (2008), is that the authors were mainly interested in modeling limited dependent variables. The main difference between the preceding models and the switching regression model is that in the case of the switching regression model the distributions of the components error terms are defined on the whole population while in the case of the

mixture models they are defined only on the corresponding sub-populations (Maddala 1999). The model can be presented as follows:

$$w_{it} = \begin{cases} 1 \text{ if } w_{it}^* > 0 \\ 2 \text{ otherwise} \end{cases}$$

$$w_{it}^* = \mathbf{z}_{it}\boldsymbol{\gamma} - \epsilon_{it}$$

$$y_{it} = \begin{cases} y_{it1} = \boldsymbol{x}_{it}\boldsymbol{\beta}_1 + u_{1it}, \text{ if } w_{it} = 1 \\ y_{it2} = \boldsymbol{x}_{it}\boldsymbol{\beta}_2 + u_{2it}, \text{ if } w_{it} = 2 \end{cases}$$

$$\begin{bmatrix} u_{1it} \\ u_{2it} \\ \epsilon_{it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1\epsilon} \\ \sigma_{12} & \sigma_2^2 & \sigma_{2\epsilon} \\ \sigma_{\epsilon 1} & \sigma_{\epsilon 2} & 1 \end{bmatrix}\right)$$

Let

$$\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \sigma_1, \boldsymbol{\beta}_2, \sigma_2, \sigma_{\epsilon 1}, \sigma_{\epsilon 2}).$$

σ_{12} will not enter the density function and is then not estimable.

The joint density for (Y_{it}, W_{it}) is given by

$$\begin{aligned} f(y_{it}, w_{it}; \boldsymbol{\theta}) &= \begin{cases} p(w_{it} = 1) f(y_{it} | w_{it} = 1; \boldsymbol{\theta}) \\ p(w_{it} = 2) f(y_{it} | w_{it} = 2; \boldsymbol{\theta}) \end{cases} \\ &= \begin{cases} p(\epsilon_{it} < \mathbf{z}_{it}\boldsymbol{\gamma}) f(y_{it} | w_{it}^* > 0; \boldsymbol{\theta}) \\ p(\epsilon_{it} \geq \mathbf{z}_{it}\boldsymbol{\gamma}) f(y_{it} | w_{it}^* \leq 0; \boldsymbol{\theta}) \end{cases} \\ &= \begin{cases} \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}) f(y_{it} | w_{it}^* > 0; \boldsymbol{\theta}) \\ (1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma})) f(y_{it} | w_{it}^* \leq 0; \boldsymbol{\theta}) \end{cases} \\ f(y_{it} | w_{it}^* > 0; \boldsymbol{\theta}) &= \frac{f(y_{it}, w_{it}^* > 0)}{p(w_{it}^* > 0)} = \frac{f(y_{it1}) p(w_{it}^* > 0 | y_{it1})}{p(w_{it}^* > 0)} \\ &= \frac{f(y_{it1}) p(\epsilon_{it} < \mathbf{z}_{it}\boldsymbol{\gamma} | u_{1it})}{p(w_{it}^* > 0)} \\ &= \frac{\left(\int_{-\infty}^{\mathbf{z}_{it}\boldsymbol{\gamma}} f(\epsilon_{it} | u_{1it})\right) f(y_{it1})}{p(w_{it}^* > 0)} \\ &= \frac{\left(\int_{-\infty}^{\mathbf{z}_{it}\boldsymbol{\gamma}} f(\epsilon_{it} | u_{1it})\right) f(y_{it1})}{\Phi(\mathbf{z}_{it}\boldsymbol{\gamma})}. \end{aligned}$$

Similarly

$$\begin{aligned} f(y_{it} | w_{it}^* \leq 0; \boldsymbol{\theta}) &= \frac{f\left(y_{it}, w_{it}^* \leq 0; \boldsymbol{\theta}\right)}{p(w_{it}^* \leq 0)} = \frac{f(y_{it1}) p(w_{it}^* \leq 0 | y_{it2})}{p(w_{it}^* \leq 0)} \\ &= \frac{f(y_{it2}) p\left(\epsilon_{it} > \mathbf{z}_{it}\boldsymbol{\gamma} | u_{2it}\right)}{p(w_{it}^* \leq 0)} \\ &= \frac{\left(\int_{\mathbf{z}_{it}\boldsymbol{\gamma}}^{\infty} f\left(\epsilon_{it} | u_{2it}; \boldsymbol{\theta}\right)\right) f(y_{it2})}{p(w_{it}^* \leq 0} \\ &= \frac{\left(\int_{\mathbf{z}_{it}\boldsymbol{\gamma}}^{\infty} f\left(\epsilon_{it} | u_{2it}; \boldsymbol{\theta}\right)\right) f(y_{it2})}{1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma})}. \end{aligned}$$

The joint density becomes

$$f(y_{it}, w_{it}; \boldsymbol{\theta}) = \begin{cases} \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}) f(y_{it} | w_{it}^* > 0; \boldsymbol{\theta}) = \left(\int_{-\infty}^{\mathbf{z}_{it}\boldsymbol{\gamma}} f(\epsilon_{it} | u_{1it}; \boldsymbol{\theta})\right) f(y_{it1}) \\ (1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma})) f\left(y_{it} | w_{it}^* \leq 0; \boldsymbol{\theta}\right) = \left(\int_{\mathbf{z}_{it}\boldsymbol{\gamma}}^{\infty} f\left(\epsilon_{it} | u_{2it}; \boldsymbol{\theta}\right)\right) f(y_{it2}) \end{cases}.$$

The complete-data likelihood is

$$\left[\left(\int_{-\infty}^{\mathbf{z}_{it}\boldsymbol{\gamma}} f(\epsilon_{it}|u_{1it};\boldsymbol{\theta})\right)f(y_{it1})\right]^{\mathbb{I}(w_{it}=1)}\left[\left(\int_{\mathbf{z}_{it}\boldsymbol{\gamma}}^{\infty} f(\epsilon_{it}|u_{2it};\boldsymbol{\theta})\right)f(y_{it2})\right]^{1-\mathbb{I}(w_{it}=1)},$$

and the marginal likelihood is

$$f(y_{it};\boldsymbol{\theta}) = \left(\int_{-\infty}^{\mathbf{z}_{it}\boldsymbol{\gamma}} f(\epsilon_{it}|u_{1it};\boldsymbol{\theta})\right)f(y_{it1}) + \left(\int_{\mathbf{z}_{it}\boldsymbol{\gamma}}^{\infty} f(\epsilon_{it}|u_{2it};\boldsymbol{\theta})\right)f(y_{it2}), \tag{12}$$

which takes the form of a mixture of two distributions. However, since ϵ_{it} is dependent on u_{2it} and u_{1it} and since $u_{2it} \neq u_{1it}$, the weights do not necessarily add up to one which is another difference between the latter model and the regular finite mixture of two normal distributions. Note that

$$\begin{bmatrix} \epsilon_{it} \\ u_{1it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{\epsilon 1} \\ \sigma_{\epsilon 1} & \sigma_1^2 \end{bmatrix}\right)$$
$$\begin{bmatrix} \epsilon_{it} \\ u_{2it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{\epsilon 0} \\ \sigma_{\epsilon 0} & \sigma_2^2 \end{bmatrix}\right).$$

Thus

$$\epsilon_{it}|u_{1it} \sim N\left(E(\epsilon_{it}) + \frac{\sigma_{\epsilon 1}}{\sigma_1^2}u_{1it}, 1 - \frac{\sigma_{\epsilon 1}^2}{\sigma_1^2}\right)$$
$$\epsilon_{it}|u_{2it} \sim N\left(E(\epsilon_{it}) + \frac{\sigma_{\epsilon 2}}{\sigma_2^2}u_{2it}, 1 - \frac{\sigma_{\epsilon 2}^2}{\sigma_2^2}\right)$$

or

$$\begin{aligned} \epsilon_{it}|u_{1it} &\sim N\left(\frac{\sigma_{\epsilon 1}}{\sigma_1^2}\left(y_{it1} - \boldsymbol{x}_{it}\boldsymbol{\beta}_1\right), 1 - \frac{\sigma_{\epsilon 1}^2}{\sigma_{11}^2}\right) \\ \epsilon_{it}|u_{2it} &\sim N\left(\frac{\sigma_{\epsilon 2}}{\sigma_2^2}\left(y_{it2} - \boldsymbol{x}_{it}\boldsymbol{\beta}_2\right), 1 - \frac{\sigma_{\epsilon 2}^2}{\sigma_2^2}\right). \end{aligned} \tag{13}$$

Thus,

$$\begin{aligned} \int_{\mathbf{z}_{it}\boldsymbol{\gamma}}^{\infty} f(\epsilon_{it}|u_{2it};\boldsymbol{\theta})d\epsilon_{it} &= p(\epsilon_{it}|u_{2it} > \mathbf{z}_{it}\boldsymbol{\gamma}) \\ &= p\left(\frac{\epsilon_{it}|u_{2it} - \frac{\sigma_{\epsilon 2}}{\sigma_2^2}\left(y_{it2} - \boldsymbol{x}_{it}\boldsymbol{\beta}_2\right)}{\sqrt{1 - \frac{\sigma_{\epsilon 2}^2}{\sigma_2^2}}} > \frac{\mathbf{z}_{it}\boldsymbol{\gamma} - \frac{\sigma_{\epsilon 2}}{\sigma_2^2}\left(y_{it2} - \boldsymbol{x}_{it}\boldsymbol{\beta}_2\right)}{\sqrt{1 - \frac{\sigma_{\epsilon 2}^2}{\sigma_2^2}}}\right) \\ &= 1 - \Phi\left(\frac{\mathbf{z}_{it}\boldsymbol{\gamma} - \frac{\sigma_{\epsilon 2}}{\sigma_2^2}\left(y_{it2} - \boldsymbol{x}_{it}\boldsymbol{\beta}_2\right)}{\sqrt{1 - \frac{\sigma_{\epsilon 2}^2}{\sigma_2^2}}}\right). \end{aligned} \tag{14}$$

Similarly,

$$\begin{aligned}\int_{-\infty}^{\mathbf{z}_{it}\boldsymbol{\gamma}} f(\epsilon_{it}|u_{1it};\boldsymbol{\theta})d\epsilon_{it} &= p(\epsilon_{it}|u_{1it} \leq \mathbf{z}_{it}\boldsymbol{\gamma}) \\ &= p\left(\frac{\epsilon_{it}|u_{1it} - \frac{\sigma_{\epsilon 1}}{\sigma_1^2}\left(y_{it1} - \boldsymbol{x}_{it}\boldsymbol{\beta}_1\right)}{\sqrt{1-\frac{\sigma_{\epsilon 1}^2}{\sigma_1^2}}} \leq \frac{\mathbf{z}_{it}\boldsymbol{\gamma} - \frac{\sigma_{\epsilon 1}}{\sigma_1^2}\left(y_{it1} - \boldsymbol{x}_{it}\boldsymbol{\beta}_1\right)}{\sqrt{1-\frac{\sigma_{\epsilon 1}^2}{\sigma_1^2}}}\right) \\ &= \Phi\left(\frac{\mathbf{z}_{it}\boldsymbol{\gamma} - \frac{\sigma_{\epsilon 1}}{\sigma_1^2}\left(y_{it1} - \boldsymbol{x}_{it}\boldsymbol{\beta}_1\right)}{\sqrt{1-\frac{\sigma_{\epsilon 1}^2}{\sigma_1^2}}}\right).\end{aligned} \tag{15}$$

By plugging Eqs. (14) and (15) in Eq. (12), the marginal likelihood becomes

$$\begin{aligned}f(y_{it};\boldsymbol{\theta}) = &\Phi\left(\frac{\mathbf{z}_{it}\boldsymbol{\gamma} - \frac{\sigma_{\epsilon 1}}{\sigma_1^2}\left(y_{it1} - \boldsymbol{x}_{it}\boldsymbol{\beta}_1\right)}{\sqrt{1-\frac{\sigma_{\epsilon 1}^2}{\sigma_1^2}}}\right)\phi(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_1,\sigma_1) \\ &+\left(1-\Phi\left(\frac{\mathbf{z}_{it}\boldsymbol{\gamma} - \frac{\sigma_{\epsilon 2}}{\sigma_2^2}\left(y_{it2} - \boldsymbol{x}_{it}\boldsymbol{\beta}_2\right)}{\sqrt{1-\frac{\sigma_{\epsilon 2}^2}{\sigma_2^2}}}\right)\right)\phi(y_{it};\boldsymbol{x}_{it}\boldsymbol{\beta}_2,\sigma_2).\end{aligned}$$

When $\sigma_{\epsilon 1} = \sigma_{\epsilon 2} = 0$, the preceding likelihood is the same as in the previous model and the weights would add up to one. Xiaoqiang and Schiantarelli (1998), Hovakimian and Titman (2006) and Almeida and Campello (2007) use classical econometric methods to estimate the preceding endogeneous switching regression model with fixed effects.

This model can also be estimated with the EM algorithm, but the intermediate EM quantity is no longer separable in the parameters which makes this less appealing than the direct maximization of the log of the marginal likelihood. Maximizing $Q(\boldsymbol{\theta},\boldsymbol{\theta}')$ at each iteration is potentially as computationally involved as the one-step maximization of the marginal likelihood. However, if one has difficulty finding good starting values for a Newton-type algorithm, one can still benefit from the nice convergence properties of the EM algorithm via the simpler model $\mathcal{M}_2$. As indicated before, if the correlations between the components and the group membership equation are zero $\mathcal{M}_3$ is identical to $\mathcal{M}_2$ and as a result the latter will provide very good starting values for the former. One just has to apply the EM algorithm to $\mathcal{M}_2$ and use the solution as starting value for $\mathcal{M}_3$.

An Endogenous Switching Regression Model with Random Effect ($\mathcal{M}_4$)

The endogenous switching regression can be extended by adding random effects in the components to capture within group heterogeneity, which is a very important issue in the panel data setting considered in this paper. Let

$$w_{it} = \begin{cases} 1 \text{ if } w_{it}^* > 0, i = 1,\ldots,N, t = 1,\ldots,T_i \\ 2 \text{ otherwise} \end{cases}$$

$$w_{it}^* = \mathbf{z}_{it}\boldsymbol{\gamma} - \epsilon_{it}$$

$$y_{it} = \begin{cases} y_{it1} = \boldsymbol{x}_{it}\boldsymbol{\beta}_1 + \alpha_{i1} + u_{it1}, \text{ if } w_{it} = 1 \\ y_{it2} = \boldsymbol{x}_{it}\boldsymbol{\beta}_2 + \alpha_{i2} + u_{it2}, \text{ if } w_{it} = 2 \end{cases}.$$

Following Mundlak (1978) I assume

$$\begin{cases} \alpha_{1i} = \bar{\boldsymbol{x}}_{.i}\boldsymbol{\zeta}_1 + \xi_{i1} \\ \alpha_{2i} = \bar{\boldsymbol{x}}_{.i}\boldsymbol{\zeta}_2 + \xi_{i2} \end{cases}$$

α_{i1} and α_{i2} capture within group heterogeneity which is decomposed into two parts: a fixed-effect part ($\mathbf{x}_{.i}\boldsymbol{\zeta_0}$ and $\mathbf{x}_{.i}\boldsymbol{\zeta_1}$) and a random effect part (ξ_{i0} and ξ_{i1}) uncorrelated with the exogenous variables, where

$$\bar{\mathbf{x}_{\mathbf{i}}} = \frac{\sum_{t=1}^{T_i} \mathbf{x_{it}}}{T_i}$$

$$\begin{bmatrix} u_{it1} \\ u_{it2} \\ \epsilon_{it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1\epsilon} \\ \sigma_{12} & \sigma_2^2 & \sigma_{2\epsilon} \\ \sigma_{\epsilon 1} & \sigma_{\epsilon 2} & 1 \end{bmatrix} \right)$$

$$\begin{pmatrix} \xi_{1i} \\ \xi_{2i} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}).$$

This specification of the firm-specific effect is interesting because in practice one expects that some of the exogenous variables will be correlated with the agent's unobserved characteristics, which may also contain a random component. Moreover, the use of two different random effects for each component distribution allows the data to dictate whether or not those agents who fall more often in a given group have the same unobserved specific characteristics as those who fall most of the time in the other group. When $\boldsymbol{\zeta_1}$ and $\boldsymbol{\zeta_2}$ equal zero one obtains the usual random effect specification.

Let

$$\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \sigma_1, \boldsymbol{\beta}_2, \sigma_2, \sigma_{\epsilon 1}, \sigma_{\epsilon 2}, \boldsymbol{\Sigma}).$$

Then

$$\begin{aligned} f(y_{it}|\xi_{i1}, \xi_{i2}; \boldsymbol{\theta}) = \Phi & \left(\frac{\mathbf{z}_{it}\boldsymbol{\gamma} - \frac{\sigma_{\epsilon 1}}{\sigma_1^2}\left(y_{it1} - \boldsymbol{x}_{it}\boldsymbol{\beta}_1 - \bar{\boldsymbol{x}}_{.i}\boldsymbol{\zeta}_1 - \xi_{i1}\right)}{\sqrt{1 - \frac{\sigma_{\epsilon 1}^2}{\sigma_1^2}}} \right) \phi(\boldsymbol{x}_{it}\boldsymbol{\beta}_1 + \bar{\boldsymbol{x}}_{.i}\boldsymbol{\zeta}_1 + \xi_{i1}, \sigma_1) \\ & + \left(1 - \Phi \left(\frac{\mathbf{z}_{it}\boldsymbol{\gamma} - \frac{\sigma_{\epsilon 2}}{\sigma_2^2}\left(y_{it2} - \boldsymbol{x}_{it}\boldsymbol{\beta}_2 - \bar{\boldsymbol{x}}_{.i}\boldsymbol{\zeta}_2 - \xi_{i2}\right)}{\sqrt{1 - \frac{\sigma_{\epsilon 2}^2}{\sigma_2^2}}} \right) \right) \\ & \times \phi(\boldsymbol{x}_{it}\boldsymbol{\beta}_2 + \bar{\boldsymbol{x}}_{.i}\boldsymbol{\zeta}_2 + \xi_{i2}, \sigma_2). \end{aligned}$$

Assuming that the response variable, y_{it}, is independent, conditional on the random effects, the unconditional likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} f(y_{it}|\xi_{i2}, \xi_{i1}; \boldsymbol{\theta}) g(\xi_{i1}, \xi_{i2}) d\xi_{i1} d\xi_{i2} \right).$$

Since the random effects are assumed to follow a normal distribution, the double integral is computed using Gauss-Hermite Quadrature. To put the integral in the convenient form I first need to write the vector of correlated normally distributed random effects as a linear function of a vector of standard normal random variables. This is done using the spectral decomposition of $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{-1},$$

where $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{\Sigma}$ while $\mathbf{S}$ is the corresponding matrix of eigenvectors. Let

$$\begin{pmatrix} \xi_{0i} \\ \xi_{1i} \end{pmatrix} = \mathbf{S}\sqrt{\mathbf{\Lambda}} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

If I write

$$\mathbf{S}\sqrt{\mathbf{\Lambda}} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

I then have:

$$\xi_{i0} = az_1 + bz_2$$
$$\xi_{i1} = cz_1 + dz_2,$$

where z_1 and z_2 are independent univariate standard normal random variables. The integral can then be approximated as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{t=1}^{T_i} f(y_{it}|\xi_{i1},\xi_{i2};\boldsymbol{\theta})g(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2} \approx \frac{1}{\pi}\sum_{r=1}^{R}\sum_{l=1}^{R} w_r w_l \prod_{t=1}^{T_i} \times f(y_{it}|az_{1r}+bz_{2l}, cz_{1r}+dz_{2l}),$$

using an R-point one-dimensional Gauss-Hermite weight w_r and nodes z_r, $r = 1,\ldots,R$.

One can alternatively use a Cholesky decomposition, but as noted by Jäckel (2005), the spectral decomposition provides a better rotation of the sampling points, which makes the evaluation of the integral potentially more robust. Another issue is the waste of computation time. The two-dimensional standard normal density for example has circular level curves centered at the origin. Its mass is concentrated within circles of rays less than or equal to 3. However, the set of sampling points obtained by taking the cartesian product of one-dimensional sets of sampling points is a square in two dimensions. The mass at the points located at the extremities of the axes of the square is almost zero and does not contribute to the integral, which wastes computation time. One way to deal with this issue is to use what is called "pruning" (Jäckel 2005) which is a way of eliminating these non-important points. This can be done by rewriting the integral approximation as:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{t=1}^{T_i} f(y_{it}|\xi_{i1},\xi_{i2})g(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2} \approx \frac{1}{\pi}\sum_{r=1}^{R}\sum_{l=1}^{R} \mathbb{I}_{\{w_r w_l > \theta_R\}} w_r w_l \prod_{t=1}^{T_i} \times f(y_{it}|az_{1r}+bz_{2l}, cz_{1r}+dz_{2l})$$

where

$$\theta_R = \frac{w_1 w_{\left[\frac{R+1}{2}\right]}}{R}.$$

Using the MATLAB function *mherzo.m* written by Zhang and Jin (1996) I have generated 9-point one-dimensional Gauss-Hermite weights, w_r and nodes z_r (r=1,..,9).

The cartesian product of the nodes with and without pruning are shown in Fig. 1.

One should note that Gaussian quadrature, or numerical integration methods in general, suffer from the curse of dimensionality. The number of function evaluations required to approximate the integral to a certain degree of accuracy increases exponentially with the dimension of the integral. Monte Carlo Integration or a monomial rule may be less costly. González et al. (2006) show that Monte Carlo and Quasi-Monte Carlo methods can

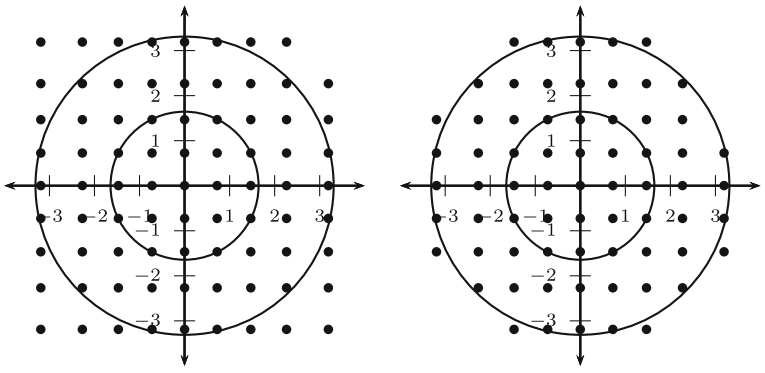

Fig. 1 An example showing the effect of pruning. The number of points is reduced from 81 on the left to the most heavily weighted 69 points on the right

not only reduce computation time but also provide better accuracy in the case of logistic regressions.

Alternatively, one can use the *h-likelihood* method by Lee and Nelder (1996) and bypass the computation of the integral. In this case the random effects are treated as additional parameters that are estimated with the other parameters. For panel data with a large number of units this method increases significantly the number of parameters to be estimated.

A Hidden Markov Model ($\mathcal{M}_5$)

One problem with the models already presented is that they do not allow the group membership at time t to be dependent on the group membership at time t-1. When the groups are made of firms having the same financial status, one should note that several of the variables used in the literature to determine the presence or the absence of financial constraints such as the firm's size, the fraction of its assets that can be used as collateral, are likely to be time-dependent and as a result, the firm's financial status at time t is potentially dependent on its status at time t-1. One way to capture this time dependence is to make the following assumption

$$p(w_{it} = 1) \neq p(w_{it} = 1|w_{it-1} = j), j = 1, 2.$$

Let

$$p(w_{it} = l|w_{it-1} = k) = {}_iP_{kl}, k = 1, 2; l = 1, 2.$$

I assume that w_{it} is an unobserved variable following a first order Markov chain on a discrete state-space. The bivariate discrete-time process (Y_{it}, W_{it}) where $Y_{it}|\boldsymbol{w}_{it}$ is independent, is a hidden Markov model (Cappé et al. 2005). Thus, the joint density for (Y_{it}, W_{it}) is given by

$$f(y_{it}, w_{it}) = \begin{cases} p(w_{it} = 1|\Im_{i(t-1)})f(y_{it}|w_{it} = 1) \\ p(w_{it} = 2|\Im_{i(t-1)})f(y_{it}|w_{it} = 2) \end{cases}$$

where $\Im_{i(t-1)}$ means information about firm i available up to time *t-1*.

If, for a given firm i, the path of the chain is:$\{w_{i1} = j_1, w_{i2} = j_2, \ldots, w_{iT_i} = j_{T_i}\}$, the joint density for this firm would be

$$\begin{aligned} f\left((y_{i1}, \ldots, y_{iT_i}), (w_{i1} = j_1, \ldots, w_{iT_i} = j_{T_i})\right) &= p(w_{i1} = j_1, \ldots, w_{iT_i} = j_{T_i}) \\ &\quad \times f\left((y_{i1}, \ldots, y_{iT_i}) | w_{i1} = j_1, w_{i2} = j_2, \ldots, w_{iT_i}\right) \\ &= p(w_{i1} = j_1) p(w_{i2} = j_2 | w_{i1} = j_1) \times \ldots \\ &\quad \times p(w_{iT_i} = j_{T_i} | w_{i(T_i-1)} = j_{T_i-1}) \\ &\quad \times f(y_{i1} | w_{i1} = j_1) \times f(y_{iT_i} | w_{iT_i} = j_{T_i}). \end{aligned}$$

Note that the total number of possible paths is 2^{T_i} for firm i. Suppose that the initial probability vector for firm i is

$${}_i\pi = ({}_i\pi_1, {}_i\pi_2).$$

The joint density can be rewritten as

$$\begin{aligned} &{}_i\pi_{j_1} \times {}_iP_{j_1 j_2} \times \ldots \times {}_iP_{j_{T_i-1} j_{T_i}} \times f(y_{i1} | w_{i1} = j_1) \times \ldots \times f\left(y_{iT_i} | w_{iT_i} = j_{T_i}\right) \\ &= {}_i\pi_{j_1} f(y_{i1} | w_{i1} = j_1) \prod_{t=2}^{T_i} {}_iP_{j_{t-1} j_t} f(y_{it} | w_{it} = j_t). \end{aligned}$$

The preceding is true if we know a priori the full path of the state variable, w_{it}. If we don't, the joint density can be written as

$$\prod_{j=1}^{2} \left({}_i\pi_j f(y_{i1} | w_{it} = j)\right)^{\mathbb{I}(w_{i1}=j)} \prod_{t=2}^{T_i} \prod_{k=1}^{2} \prod_{l=1}^{2} \left({}_iP_{kl} f(y_{it} | w_{it} = l)\right)^{\mathbb{I}(w_{it-1}=k, w_{it}=l)}.$$

The marginal density for firm i for the observed data is then

$$\sum_{j_1=1}^{2} \cdots \sum_{j_{T_i}=1}^{2} {}_i\pi_{j_1} {}_iP_{j_1 j_2} \times \ldots \times {}_iP_{j_{T_i-1} j_{T_i}} \times f(y_{i1} | w_{i1} = j_1) \times \ldots \times f\left(y_{iT_i} | w_{iT_i} = j_{T_i}\right).$$

If

$$\lambda(y_{it}) = \begin{bmatrix} f(y_{it} | w_{it} = 1) & 0 \\ 0 & f(y_{it} | w_{it} = 2) \end{bmatrix}, \boldsymbol{\gamma}_{it} = \begin{bmatrix} {}_iP_{11} & {}_iP_{12} \\ {}_iP_{21} & {}_iP_{22} \end{bmatrix},$$

then the marginal density can be rewritten using vector-matrix operations (MacDonald and Zucchini 1997)

$${}_i\pi \lambda(y_{i1}) \boldsymbol{\gamma}_{i2} \lambda(y_{i2}) \times \ldots \times \boldsymbol{\gamma}_{i_{T_i}} \lambda(y_{i_{T_i}}) \mathbf{1}' = {}_i\pi \lambda(y_{i1}) \left(\prod_{t=2}^{T} \boldsymbol{\gamma}_{it} \lambda(y_{it}) \right) \mathbf{1}', \tag{16}$$

where $\mathbf{1}'$ is a column vector of ones.

Parameters Estimation (EM Algorithm)

Let $\boldsymbol{\theta} = \left({}_i\pi_1, {}_iP_{11}, {}_iP_{22}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$ be the vector of parameters of the model. With N firms, the dimension of the vector $\boldsymbol{\theta}$ is

$$\dim(\boldsymbol{\theta}) = 3N + 2\dim(\boldsymbol{\beta}_i)$$

which is a large number of parameters. To reduce the number of parameters to be estimated, I assume that

$${}_i\pi_1 = \pi_1, {}_iP_{11} = P_{11}, {}_iP_{22} = P_{22},$$

then,

$$\dim(\boldsymbol{\theta}) = 3 + 2\dim(\boldsymbol{\beta}_i).$$

The *complete-data likelihood* is given by

$$L^c(\boldsymbol{\theta}) = \prod_{i=1}^{N}\left(\prod_{j=1}^{2}\left(\pi_j f(y_{i1}|w_{it}=j)\right)^{\mathbb{I}(w_{i1}=j)}\prod_{t=2}^{T_i}\prod_{k=1}^{2}\prod_{l=1}^{2}\left(P_{kl}f(y_{it}|w_{it}=l)\right)^{\mathbb{I}(w_{it-1}=k,w_{it}=l)}\right), \tag{17}$$

and the *complete-data log-likelihood* is

$$\begin{aligned} l^c(\boldsymbol{\theta}) = \log\left(L^c(\boldsymbol{\theta})\right) = \sum_{i=1}^{n}\Bigg(&\sum_{j=1}^{2}\mathbb{I}(w_{i1}=j)\log\left(\pi_j f(y_{it}|w_{i1}=j)\right) \\ &+\sum_{t=2}^{T_i}\sum_{k=0}^{1}\sum_{l=1}^{2}\mathbb{I}(w_{i(t-1)}=k,w_{it}=l)\log\left(P_{kl}f(y_{it}|w_{it}=l)\right)\Bigg) \end{aligned}$$

or

$$\begin{aligned} l^c(\boldsymbol{\theta}) = \log\left(L^c(\boldsymbol{\theta})\right) = \sum_{i=1}^{n}\Bigg(&\sum_{j=1}^{2}\mathbb{I}(w_{i1}=j)\log\left(\pi_j f(y_{it}|w_{i1}=j)\right) \\ &+\sum_{t=2}^{T_i}\sum_{k=1}^{2}\sum_{l=1}^{2}\mathbb{I}(w_{i(t-1)}=k,w_{it}=l)\log\left(P_{kl}\right) \\ &+\sum_{t=2}^{T_i}\sum_{k=1}^{2}\sum_{l=1}^{2}\mathbb{I}(w_{it}=l)\log f(y_{it}|w_{it}=l)\Bigg). \end{aligned}$$

The intermediate quantity of EM is

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') &= E_{\boldsymbol{\theta}'}(l^c(\boldsymbol{\theta})|\mathfrak{F}_T) \\ &= \sum_{i=1}^{n}\Bigg(\sum_{j=1}^{2}E_{\boldsymbol{\theta}'}\left[\mathbb{I}(w_{i1}=j)\log\left(\pi_j f(y_{it}|w_{i1}=j)\right)|\mathfrak{F}_{iT_i}\right] \\ &\quad+\sum_{t=2}^{T_i}\sum_{k=1}^{2}\sum_{l=1}^{2}E_{\boldsymbol{\theta}'}\left[\mathbb{I}(w_{i(t-1)}=k,w_{it}=l)\log\left(P_{kl}\right)\right] \\ &\quad+\sum_{t=2}^{T_i}\sum_{k=1}^{2}\sum_{l=1}^{2}E_{\boldsymbol{\theta}'}\left[\mathbb{I}(w_{it}=l)\log\left(f(y_{it}|w_{it}=l)\right)|\mathfrak{F}_{iT_i}\right]\Bigg). \end{aligned}$$

To get the preceding expectation, only $E_{\boldsymbol{\theta}'}(\mathbb{I}(w_{it}=j|\mathfrak{F}_{iT_i}))$ and $E_{\boldsymbol{\theta}'}(\mathbb{I}(w_{i(t-1)}=k,w_{it}=l|\mathfrak{F}_{iT_i}))$ need to be evaluated. Note that

$$\begin{aligned} E_{\boldsymbol{\theta}'}\left(\mathbb{I}(w_{it}=j|\mathfrak{F}_{iT_i})\right) &= p\left(w_{it}=j|\mathfrak{F}_{iT_i},\boldsymbol{\theta}'\right) \\ E_{\boldsymbol{\theta}'}\left(\mathbb{I}(w_{i(t-1)}=k,w_{it}=l|\mathfrak{F}_{iT_i})\right) &= p\left(w_{i(t-1)}=k,w_{it}=l|\mathfrak{F}_{iT_i},\boldsymbol{\theta}'\right) \end{aligned}$$

$$
\begin{aligned}
&p(w_{i(t-1)}=k,w_{it}=l|\Im_{iT_i},\boldsymbol{\theta}')\\
&=\frac{p\left(w_{i(t-1)}=k,w_{it}=l,\Im_{iT_i};\boldsymbol{\theta}'\right)}{f(\Im_{iT_i})}\\
&=\frac{p\left(w_{i(t-1)}=k,w_{it}=l,y_{i1},\ldots,y_{i(t-1)},y_{it},..,y_{iT_i};\boldsymbol{\theta}'\right)}{f(y_{i1},\ldots,y_{i(t-1)},y_{it},..,y_{iT_i})}\\
&=\frac{p\left(y_{i1},\ldots,y_{i(t-1)},w_{i(t-1)}=k;\boldsymbol{\theta}'\right)\times p(y_{it},..,y_{iT_i},w_{it}=l|w_{i(t-1)}=k,y_{i1},\ldots,y_{i(t-1)};\boldsymbol{\theta}')}{f(y_{i1},\ldots,sy_{i(t-1)},y_{it},..,y_{iT_i})}\\
&=\frac{\alpha_{i(t-1)}(k)\times p\left(w_{it}=l|w_{i(t-1)}=k;\boldsymbol{\theta}'\right)\times f\left(y_{it},..,y_{iT_i}|w_{it}=l,w_{i(t-1)}=k;\boldsymbol{\theta}'\right)}{f(y_{i1},\ldots,y_{i(t-1)},y_{it},..,y_{iT_i})}\\
&=\frac{\alpha_{i(t-1)}(k)\times_i P_{kl}f\left(y_{it}|w_{it}=l\right)\times f\left(y_{i(t+1)},..,y_{iT_i}|w_{it}=l,w_{i(t-1)}=k;\boldsymbol{\theta}'\right)}{f(y_{i1},\ldots,y_{i(t-1)},y_{it},..,y_{iT_i})}\\
&=\frac{\alpha_{i(t-1)}(k)\times_i P_{kl}f(y_{it}|w_{it}=l)\times\breve{\beta}_{it}(l)}{f(y_{i1},\ldots,y_{i(t-1)},y_{it},..,y_{iT_i})}
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_{it}(k)&=p(y_{i1},\ldots,y_{it},w_{it}=k)\\
\breve{\beta}_{it}(k)&=f(y_{i(t+1)},\ldots,y_{iT_i}|w_{it}=k).
\end{aligned}
$$

The EM algorithm for this model proceeds as follows:

1. Choose initial values $\boldsymbol{\theta}_0$ and,
2. Compute $p(w_{it}=j|\Im_{iT_i};\boldsymbol{\theta}_0)$ and $p(w_{i(t-1)}=k,w_{it}=l|\Im_{iT_i};\boldsymbol{\theta}_0)$ for each observation,
3. Substitute the computed probability in the intermediate EM quantity $(Q)(\boldsymbol{\theta}_1,\boldsymbol{\theta}_0)$,
4. Solve

$$
\begin{aligned}
&\boldsymbol{\theta}_1=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\\
&\text{subject to:}\\
&\sum_{j=1}^{2}\pi_j=1\\
&\sum_{l=1}^{2}P_{kl}=1;k=1,2\\
&0\leq\pi_j\leq 1,j=1,2.
\end{aligned}
$$

5. Repeat step 2 after replacing $\boldsymbol{\theta}_0$ by $\boldsymbol{\theta}_1$,
6. Keep going until convergence.

It should be noted that the forward-backward algorithm used to obtain $\alpha_{it}(k)$ and $\breve{\beta}_{it}(k)$ is subject to numerical underflow. To avoid this problem the FORTRAN codes used for this algorithm apply the scaling method proposed by Rabiner (1989). The version of the EM algorithm just presented is also known as the Baum-Welch algorithm. Step 4 is called the M-step or maximization step. The Lagrangian for the problem is

$$
\mathcal{L}(\boldsymbol{\theta},\lambda,\lambda_k;\boldsymbol{\theta}')=\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}')+\lambda\left(1-\sum_{j=1}^{2}\pi_j\right)+\sum_{k=1}^{2}\lambda_k\left(1-\sum_{l=1}^{2}P_{kl}\right).
$$

Assume, as before, that $f(y_{it}|w_{it}=l)$ is the density function of the normal distribution. Then,

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}\log f(y_{it}|w_{it}=l) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\ln\pi - \sum_{i=1}^{n}\sum_{t=1}^{T_i}\ln\sigma_l - \frac{1}{2\sigma_l^2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it}-\boldsymbol{x}_i\boldsymbol{\beta}_l)^2.$$

Let

$$\begin{aligned}
\sqrt{p(w_{it}=l|\Im_{iT_i};\boldsymbol{\theta})}y_{it} &= y_{it}^{(l)}\\
\sqrt{p(w_{it}=l|\Im_{iT_i};\boldsymbol{\theta})}\boldsymbol{x}_{it} &= \boldsymbol{x}_{it}^{(l)}\\
\left(y_{11}^{(l)},\ldots,y_{1T_1}^{(l)},y_{21}^{(l)},\ldots,y_{nT_n}^{(l)}\right)' &= \mathbf{y^{(l)}}\\
\left(\boldsymbol{x}_{11}^{(l)},\ldots,\boldsymbol{x}_{1T_1}^{(l)},\boldsymbol{x}_{21}^{(l)},\ldots,\boldsymbol{x}_{nT_n}^{(l)}\right)' &= \boldsymbol{x^{(l)}},
\end{aligned}$$

then

$$\begin{aligned}
p(w_{it}=l|\Im_{iT_i};\boldsymbol{\theta})\log f(y_{it}|w_{it}=l) = &-\frac{1}{2}\sum_{i=1}^{n}\sum_{t=2}^{T_i}p(w_{it}=l|\Im_{iT_i};\boldsymbol{\theta})\ln(2\pi)\\
&-\sum_{i=1}^{n}\sum_{t=2}^{T_i}p(w_{it}=l|\Im_{iT_i};\boldsymbol{\theta})\ln\sigma_l\\
&-\frac{1}{2\sigma_l^2}\left(\mathbf{y^{(l)}}-\boldsymbol{x^{(l)}}\boldsymbol{\beta}_l\right)'\left(\mathbf{y^{(l)}}-\boldsymbol{x^{(l)}}\boldsymbol{\beta}_l\right).
\end{aligned}$$

The first order conditions for the maximization problem are the following:

$$wrt:\pi_j \qquad \frac{\sum_{i=1}^{n}p\left(w_{i1}=j|\Im_{iT_i};\boldsymbol{\theta}'\right)}{\pi_j}=\lambda \tag{18}$$

$$wrt:P_{kl} \quad \frac{\sum_{i=1}^{n}\sum_{t=2}^{T_i}p(w_{i(t-1)}=k,w_{it}=l|\Im_{iT_i};\boldsymbol{\theta}')}{P_{kl}}=\lambda_k, k=1,2; l=1,2 \tag{19}$$

$$wrt:\boldsymbol{\beta}_l \qquad \left(\boldsymbol{x^{(l)}}\right)'\boldsymbol{x^{(l)}}\hat{\boldsymbol{\beta}}_l=\left(\boldsymbol{x^{(l)}}\right)'\mathbf{y^{(l)}} \tag{20}$$

$$wrt:\sigma^2 \qquad -\frac{1}{2}\sum_{i=1}^{n}\sum_{t=2}^{T_i}\frac{p\left(w_{it}=l|\Im_{iT_i};\boldsymbol{\theta}'\right)}{\sigma_l^2}=\frac{1}{2\sigma_l^4}\left(\mathbf{y^{(l)}}-\boldsymbol{x^{(l)}}\boldsymbol{\beta}_l\right)'\left(\mathbf{y^{(l)}}-\boldsymbol{x^{(l)}}\boldsymbol{\beta}_l\right) \tag{21}$$

$$wrt:\lambda \qquad \sum_{j=1}^{2}\pi_j=1 \tag{22}$$

$$wrt:\lambda_k \qquad \sum_{l=1}^{2}P_{kl}=1. \tag{23}$$

Combining Eqs. (18) and (22) one gets

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{2}p(w_{i1}=j|\Im_{iT_i};\boldsymbol{\theta}')=\sum_{j=0}^{n}\hat{\lambda}\hat{\pi}_j\\
&\Rightarrow\sum_{i=1}^{n}1=\hat{\lambda}\\
&\Rightarrow\hat{\lambda}=n.
\end{aligned}$$

Thus,

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} p(w_{i1} = j|\Im_{iT_i};\boldsymbol{\theta}'). \tag{24}$$

Combining Eqs. (19) and (23)

$$\sum_{i=1}^{n}\sum_{t=2}^{T_i}\sum_{l=1}^{2} p\left(w_{i(t-1)} = k, w_{it} = l|\Im_{iT_i};\boldsymbol{\theta}'\right) = \hat{\lambda}_k \sum_{l=1}^{2}\hat{P}_{kl}$$
$$\Rightarrow \sum_{i=1}^{n}\sum_{t=2}^{T_i} p\left(w_{i(t-1)} = k|\Im_{iT_i};\boldsymbol{\theta}'\right) = \hat{\lambda}_k,$$

which implies

$$\hat{P}_{kl} = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T_i} p\left(w_{i(t-1)} = k, w_{it} = l|\Im_{iT_i};\boldsymbol{\theta}'\right)}{\sum_{i=1}^{n}\sum_{t=2}^{T_i} p\left(w_{i(t-1)} = k|\Im_{iT_i};\boldsymbol{\theta}'\right)}. \tag{25}$$

From Eq. (20) one gets

$$\hat{\boldsymbol{\beta}}_l = \left(\left(\boldsymbol{x}^{(\mathbf{l})}\right)'\boldsymbol{x}^{(\mathbf{l})}\right)^{-1}\left(\boldsymbol{x}^{(\mathbf{l})}\right)'\mathbf{y}^{(\mathbf{l})}. \tag{26}$$

From Eq. (21) one obtains

$$\hat{\sigma^2} = \frac{\left(\mathbf{y}^{(\mathbf{l})} - \boldsymbol{x}^{(\mathbf{l})}\boldsymbol{\beta}_l\right)'\left(\mathbf{y}^{(\mathbf{l})} - \boldsymbol{x}^{(\mathbf{l})}\hat{\boldsymbol{\beta}}_l\right)}{\sum_{i=1}^{n}\sum_{t=2}^{T_i} p(w_{it} = l|\Im_{iT_i};\boldsymbol{\theta}')}. \tag{27}$$

One main drawback with the HMM model with constant transition matrix is that the probability for a firm to move from one state to another does not depend on any observable, which is unrealistic for reasons considered in the case of the first model.

HMM Model with Time dependent Transition Matrix ($\mathcal{M}_6$)

To relax the constraint imposed on the preceding model by the constant transition probabilities, a transition matrix whose components are functions of some observables can be used. Suppose

$$w_{it} = \begin{cases} 1 \text{ if } w_{it}^* > 0, t = 1,\ldots,T_i; i = 1,\ldots,n \\ 2 \text{ otherwise} \end{cases} \tag{28}$$

where

$$w_{it}^* = \mathbf{z}_{it}\boldsymbol{\gamma} + \lambda(w_{i(t-1)} - 1) - \epsilon_{it}; \epsilon_{it} \sim N(0,1) \tag{29}$$

The preceding equation means that it is possible to predict the financial situation of firm i at time t using its situation at time t-1 and some exogenous variables $\mathbf{z}_{it}$. Thus,

$$p(w_{it} = 1|w_{i(t-1)} = 1) = p(\epsilon_{it} \geq \mathbf{z}_{it}\boldsymbol{\gamma}) = \Phi(\mathbf{z}_{it}\boldsymbol{\gamma})$$
$$p(w_{it} = 2|w_{i(t-1)} = 2) = p(\epsilon_{it} < \mathbf{z}_{it}\boldsymbol{\gamma} + \lambda) = 1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma} + \lambda),$$

the transition matrix is then

$$\begin{bmatrix} \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}) & 1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}) \\ \Phi(\mathbf{z}_{it}\boldsymbol{\gamma} + \lambda) & 1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma} + \lambda) \end{bmatrix}.$$

This is a time heterogeneous transition matrix. This matrix is different from the specifications in Asea and Blomberg (1998), Atman (2007) and Maruotti (2007). It is also

possible to use a probit or logit model for each row of the transition matrix. In fact, when the Markov chain has more than two states a multinomial probit or logit model would be the most convenient choice. However, for a chain with two states, the current specification appears to be better since it involves a smaller number of parameters and offers a nice way to test for time dependence by testing the hypothesis $\lambda = 0$.

Parameters Estimation

The complete-data log-likelihood function looks the same as in the previous section. The only difference is that the transition probabilities depend now on the parameters $\boldsymbol{\gamma}$ and λ. As a result, instead of estimating the transition matrix, I will have to estimate $\boldsymbol{\gamma}$ and λ. Note that there are no closed form solutions for the first order conditions with respect to $\boldsymbol{\gamma}$ and λ. So, the M-step of the EM algorithm will include a Newton-Rapthon maximization step.

$$(\hat{\boldsymbol{\gamma}}, \hat{\lambda}) = \arg\max_{(\boldsymbol{\gamma},\lambda)} \sum_{i=1}^{n} \sum_{t=2}^{T_i} \sum_{k=1}^{2} \sum_{l=1}^{2} \mathbb{E}_{w_{it}} \left[\mathbb{I}(w_{i(t-1)} = k, w_{it} = l) \log ({}_iP_{kl}); \boldsymbol{\theta}' \right],$$

where ${}_iP_{kl}, k = 1, 2; l = 1, 2; i = 1, .., n$ are given in the preceding transition matrix. The HMM model presented in this section does not account for within group heterogeneity which opens the door for a possible extension.

Hidden Markov Model with Time Varying Transition Matrix and Random Effects ($\mathcal{M}_7$)

Even though the groups are homogeneous with respect to the financial characteristics used to form them, there are still some unobserved characteristics with respect to which the firms within a given group can be considered to be heterogeneous. One such characteristic is the difference in management. To take account of this additional source of heterogeneity, I introduce an unobserved firm specific variable in each of the two components. Let

$$\begin{cases} \alpha_{1i} = \bar{\boldsymbol{x}}_{.i}\boldsymbol{\zeta}_1 + \xi_{i1} \\ \alpha_{2i} = \bar{\boldsymbol{x}}_{.i}\boldsymbol{\zeta}_2 + \xi_{i2} \end{cases} \tag{30}$$

$$\begin{pmatrix} \xi_{1i} \\ \xi_{2i} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{31}$$

ξ_{ji} (j=1,2) are random effects that are uncorrelated with $\boldsymbol{x}_{it}$ and $\bar{\boldsymbol{x}}_{.i}$. The conditional expectations will be modeled such that

$$E(y_{it}|\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_{it}, \xi_{ji}, w_{it} = j) = \boldsymbol{x}_{it}\boldsymbol{\beta}_j + \bar{\boldsymbol{x}}_{it}\boldsymbol{\zeta}_j + \xi_{ji}, j = 1, 2. \tag{32}$$

I also assume that the random effect is independent of the firm's financial situation captured with the variable w_{it} and that conditional on the random effects and $\{w_{it}\}_1^{T_i}$, investment is independent. The complete-data likelihood can be written as

$$L^c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\prod_{j=1}^{2} \left(\pi_j f(y_{i1}|w_{it} = j, \xi_{ij})\right)^{\mathbb{I}(w_{i1}=j)} \times \prod_{t=2}^{T_i} \prod_{k=1}^{2} \prod_{l=1}^{2} \left({}_iP_{kl} f(y_{it}|w_{it} = l, \xi_{ij})\right)^{\mathbb{I}(w_{it-1}=k, w_{it}=l)} h(\xi_{1i}, \xi_{2i}) \right).$$

The complete-data log-likelihood is then

$$l^c(\boldsymbol{\theta}) = \log\left(L^c(\boldsymbol{\theta})\right) = \sum_{i=1}^{n}\left(\sum_{j=1}^{2}\mathbb{I}(w_{i1}=j)\log\left(\pi_j f(y_{it}|w_{i1}=j,\xi_{ij})\right)\right.$$
$$+\sum_{t=2}^{T_i}\sum_{k=1}^{2}\sum_{l=1}^{2}\mathbb{I}(w_{i(t-1)}=k,w_{it}=l)\log\left({}_iP_{kl}\right)$$
$$+\sum_{t=2}^{T_i}\sum_{k=1}^{2}\sum_{l=1}^{2}\mathbb{I}(w_{it}=l)\log f(y_{it}|w_{it}=l,\xi_{ij})$$
$$\left.+\log h(\xi_{i1},\xi_{i2})\right).$$

The intermediate EM quantity is given by

$$\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}') = \mathbb{E}_{\xi}\left[\mathbb{E}_{w_{it}}(l^c(\boldsymbol{\theta})|\Im_{iT_i};\boldsymbol{\theta}')\right]$$
$$= \sum_{i=1}^{n}\left(\sum_{j=1}^{2}\mathbb{E}_{w_{it}}\left[\mathbb{I}(w_{i1}=j)\log\left(\pi_j\right)\right)|\Im_{iT_i};\boldsymbol{\theta}'\right] + \sum_{t=2}^{T_i}\sum_{k=1}^{2}\sum_{l=1}^{2}$$
$$\times\,\mathbb{E}_{w_{it}}\left[\mathbb{I}(w_{i(t-1)}=k,w_{it}=l)\log\left({}_iP_{kl}\right)|\Im_{iT_i};\boldsymbol{\theta}'\right] + \int\int\sum_{t=1}^{T_i}\sum_{k=1}^{2}\sum_{l=1}^{2}$$
$$\times\,\mathbb{E}_{w_{it}}\left[\mathbb{I}(w_{it}=l)\log\left(f(y_{it}|w_{it}=l,\xi_{il})\right)|\Im_{iT_i};\boldsymbol{\theta}'\right]h(\xi_{i0},\xi_{i1}|\Im_{iT_i})d\xi_{i0}d\xi_{i1}$$
$$\left.+\iint\log\left(h(\xi_{i0},\xi_{i1})\right)h(\xi_{i1},\xi_{i2}|\Im_{iT_i})d\xi_{i0}d\xi_{i1}\right).$$

Closed-form solution for the maximization of the intermediate EM quantity $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}')$ exists only for the first component. The other three components have to be maximized using a Newton-type method. Let the Lagrangian for the first component be

$$L(\pi_j,\boldsymbol{\zeta}) = \sum_{i=1}^{n}\left(\sum_{j=1}^{2}\mathbb{E}_{w_{it}}\left[\mathbb{I}(w_{i1}=j)\log\left(\pi_j\right)\right)|\Im_{iT};\boldsymbol{\theta}'\right]\right) + \boldsymbol{\zeta}\left(1-\sum_{j=1}^{2}\pi_j\right).$$

The first order conditions are

$$\sum_{i=1}^{n}\mathbb{E}_{w_{it}}\left[\mathbb{I}(w_{i1}=j)|\Im_{iT};\boldsymbol{\theta}'\right] = \hat{\boldsymbol{\zeta}}\hat{\pi}_j; j=1,2 \tag{33}$$

$$\sum_{j=1}^{2}\hat{\pi}_j = 1. \tag{34}$$

Thus,

$$\sum_{j=0}^{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{I}(w_{i1}=j)|\Im_{iT};\boldsymbol{\theta}'\right] = \hat{\boldsymbol{\zeta}}\sum_{j=1}^{2}\hat{\pi}_j. \tag{35}$$

Using Eq. (34) in Eq. (35), I get

$$\hat{\boldsymbol{\zeta}} = \sum_{j=0}^{n}\sum_{i=1}^{n}\mathbb{E}_{w_{it}}\left[\mathbb{I}(w_{i1}=j)|\Im_{iT};\boldsymbol{\theta}'\right] = n \tag{36}$$

since

$$\sum_{j=1}^{2}\sum_{i=1}^{n}\mathbb{E}_{w_{it}}\left[\mathbb{I}(w_{i1}=j)|\mathfrak{F}_{iT};\boldsymbol{\theta}'\right]=\sum_{i=1}^{n}\sum_{j=1}^{2}p(w_{i1}=j|\mathfrak{F}_{iT})=\sum_{i=1}^{n}1=n.$$

Thus

$$\begin{aligned}\hat{\pi}_j&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{w_{it}}\left[\mathbb{I}(w_{i1}=j)|\mathfrak{F}_{iT};\boldsymbol{\theta}'\right]\\&=\frac{1}{n}\sum_{i=1}^{n}p(w_{i1}=j|\mathfrak{F}_{iT})\\&=\frac{1}{n}\sum_{i=1}^{n}\frac{p(w_{i1}=j,\mathfrak{F}_{iT})}{f(\mathfrak{F}_{iT})}=\frac{1}{n}\sum_{i=1}^{n}\frac{\int\int p(w_{i1}=j,\mathfrak{F}_{iT}|\xi_{i1},\xi_{i2})h(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2}}{\int\int f(\mathfrak{F}_{iT}|\xi_{i1},\xi_{i2})h(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2}}\\&=\frac{1}{n}\sum_{i=1}^{n}\frac{\int\int p(w_{i1}=j,y_{i1}|\xi_{i1},\xi_{i2}))\times f(y_{i2},\cdots,y_{iT_i}|w_{i1}=j,\xi_{i1},\xi_{i2})h(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2}}{\int\int f(y_{i1},\cdots,y_{iT_i}|\xi_{i1},\xi_{i2})h(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2}}\\&=\frac{1}{n}\sum_{i=1}^{n}\frac{\int\int \nu_{it}(j|\xi_{i1},\xi_{i2})\breve{\beta}_{it}(j|\xi_{i1},\xi_{i2})h(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2}}{\sum_{j=1}^{2}\int\int \nu_{it}(j|\xi_{i1},\xi_{i2})\breve{\beta}_{it}(j|\xi_{i1},\xi_{i2})h(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2}},\end{aligned}$$

where

$$\nu_{i1}(j|\xi_{i1},\xi_{i2})=p(w_{i1}=j,y_{i1}|\xi_{i1},\xi_{i2})=p(w_{i1}=j|\xi_{i1},\xi_{i2})f(y_{i1}|w_{i1}=j,\xi_{i1},\xi_{i2})$$

$$\begin{aligned}\breve{\beta}_{i1}(j|\xi_{i1},\xi_{i2})&=f(y_{i2},\cdots,y_{iT_i}|w_{i1}=j,\xi_{i1},\xi_{i2})\\&=\sum_{k=1}^{2}\frac{f(y_{i2},\cdots,y_{iT_i},w_{i1}=j,w_{i2}=k|\xi_{i1},\xi_{i2})}{p(w_{i1}=j|\xi_{i1},\xi_{i2})}\\&=\sum_{k=1}^{2}\left[\frac{p(w_{i1}=j|\xi_{i1},\xi_{i2})p(w_{i2}=k|w_{i1}=j,\xi_{i1},\xi_{i2})}{p(w_{i1}=j|\xi_{i1},\xi_{i2})}\right.\\&\quad\left.\times f(y_{i2},\cdots,y_{iT_i}|w_{i1}=j,w_{i2}=k,\xi_{i1},\xi_{i2})\right]\\&=\sum_{k=1}^{2}\left[p(w_{i2}=k|w_{i1}=j,\xi_{i1},\xi_{i2})\right.\\&\quad\left.\times f(y_{i2}|w_{i2}=k,\xi_{i1},\xi_{i2})f(y_{i3},\cdots,y_{iT_i}|w_{i2}=k,\xi_{i1},\xi_{i2})\right]\\&=\sum_{k=1}^{2}\left({}_iP_{jk}f(y_{i2}|w_{i2}=k,\xi_{i1},\xi_{i2})\breve{\beta}_{i2}(k|\xi_{i1},\xi_{i2})\right).\end{aligned}$$

Let

$$\begin{aligned}\nu_{it}(j|\xi_{i1},\xi_{i1}&=p(w_{it}=j,y_{it}|\xi_{i1},\xi_{i1})\\\breve{\beta}_{it}(j|\xi_{i1},\xi_{i1})&=\sum_{k=1}^{2}\left({}_iP_{jk}f(y_{i(t+1)}|w_{i(t+1)}=k,\xi_{i1},\xi_{i2})\breve{\beta}_{i(t+1)}(k)\right),\end{aligned}$$

then

$$\begin{aligned}\nu_{it}(j)&=\int\int(\nu_{it}(j|\xi_{i1},\xi_{i2})h(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2}\\\breve{\beta}_{it}(j)&=\int\int\breve{\beta}_{it}(j|\xi_{i1},\xi_{i2})h(\xi_{i1},\xi_{i2})d\xi_{i1}d\xi_{i2}.\end{aligned}$$

The integrals are computed using Gauss-Hermite quadrature.

Hidden Markov Model with Time Varying Transition Matrix and endogeneity ($\mathcal{M}_8$)

An alternative way of extending model $\mathcal{M}_6$ is to assume that the states of the Markov chain and the response variable are dependent. More precisely, we can assume

$$y_{it} = \begin{cases} y_{it1} = \boldsymbol{x}_{it}\boldsymbol{\beta}_1 + u_{1it}, \text{ if } w_{it} = 1 \\ y_{it2} = \boldsymbol{x}_{it}\boldsymbol{\beta}_2 + u_{2it}, \text{ if } w_{it} = 2 \end{cases},$$

together with Eqs. (28), (29) and

$$\begin{bmatrix} \epsilon_{it} \\ u_{1it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{\epsilon 1} \\ \sigma_{\epsilon 1} & \sigma_1^2 \end{bmatrix}\right)$$
$$\begin{bmatrix} \epsilon_{it} \\ u_{2it} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{\epsilon 2} \\ \sigma_{\epsilon 2} & \sigma_2^2 \end{bmatrix}\right).$$

The last distributional assumptions make the states of the Markov chains and the response variable y_{it} interdependent. The resulting model is an extension to the panel data setting of a modified version of the model by Kim et al. (2008). The transition matrix of the current model uses less parameters and the correlations between the state-indicator variable and the component distributions are allowed to be different.

Parameters estimation

Because of the interdependence between the states of the Markov chains and the response variable, during the maximization step of the EM the parameters of the transition matrix and the component distributions have to be estimated together. As a result, the EM algorithm does not have any computational advantage over a Newton-type algorithm applied to the marginal likelihood. The latter can be written as in Eq. (16) after some suitable transformation. Note that

$$f(y_{it}|w_{it} = 1) = \frac{f(y_{it}, w_{it} = 1)}{\Phi(\mathbf{z}_{it}\boldsymbol{\gamma} + \lambda(w_{i(t-1)} - 1))}.$$

Thus, to evaluate this conditional density the current state and the previous state are both needed. The computation of the likelihood will require conditional densities that depend on the current state and the previous state. Since the transition matrix has two states, four conditional densities will result. To write the likelihood as in Eq. (16), the Markov chain has to be written as a four-state chain. Let WW_{it} be the new Markov chain with state space

$$\{11, 12, 21, 22\}.$$

ww_{it} equals kl is equivalent to $w_{i(t-1)}$ equals k and w_{it} equals l. The transition matrix associated to the new chain can be written as

$$\boldsymbol{\gamma}_{it} = \begin{bmatrix} {}_{it}P_{11} & {}_{it}P_{12} & 0 & 0 \\ 0 & 0 & {}_{it}P_{21} & {}_{it}P_{22} \\ {}_{it}P_{11} & {}_{it}P_{12} & 0 & 0 \\ 0 & 0 & {}_{it}P_{21} & {}_{it}P_{22} \end{bmatrix}$$

Since the component densities now depend on the current state and the previous state, if the initial distribution of the old state-indicator variable (w_{it}) is still the distribution at time 1 the first observation of each firm will not enter the computation of the likeli-

hood. One alternative is to assume that the initial distribution is the distribution at time 0. In this case, even when the correlation between the state-indicator variable and the component distributions are zero, the likelihood of the current model will not be equal to the likelihood of model $\mathcal{M}_7$. As a result when testing for endogeneity, direct tests on the correlation coefficients may be preferred to the likelihood ratio test comparing model $\mathcal{M}_7$ and $\mathcal{M}_8$. Given the preceding assumption the initial distribution of the new state-indicator variable is

$$(\pi_1 P_{11}, \pi_1 P_{12}, \pi_2 P_{21}, \pi_2 P_{22})$$

Let

$$\boldsymbol{\lambda}(y_{it}) = \begin{bmatrix} f(y_{it}|ww_{it}=11) & 0 & 0 & 0 \\ 0 & f(y_{it}|ww_{it}=12) & 0 & 0 \\ 0 & 0 & f(y_{it}|ww_{it}=21) & 0 \\ 0 & 0 & 0 & f(y_{it}|ww_{it}=22) \end{bmatrix} \tag{37}$$

With these transformations, the marginal likelihood is given by Eq. (16). The component densities are

$$\begin{aligned} f(y_{it}|ww_{it}=1k) &= f(y_{it}|w_{i(t-1)}=1, w_{it}=k) \\ &= \frac{f(y_{it1})p(w_{i(t-1)}=k|y_{it})p(w_{it}=1|w_{i(t-1)}=k,y_{it})}{p(w_{i(t-1)}=k|y_{it})p(w_{it}=1|w_{i(t-1)}=k,y_{it})} \\ &= \frac{f(y_{it1})p(w_{it}=1|w_{i(t-1)}=k,y_{it})}{p(w_{it}=1|w_{i(t-1)}=k,y_{it1})} \\ &= \frac{f(y_{it1})p(\epsilon_{it} < \mathbf{z}_{it}\boldsymbol{\gamma} + \lambda(k-1)|u_{1it})}{p(\epsilon_{it} < \mathbf{z}_{it}\boldsymbol{\gamma} + \lambda(k-1))}, k = 1, 2. \end{aligned}$$

The conditional distribution of ϵ_{it} given u_{1it} is given in Eq. (13). Using this conditional distribution the previous expression becomes

$$f(y_{it}|ww_{it}=k1) = f(y_{it}) \frac{\Phi\left[\frac{\left(\mathbf{z}_{it}\boldsymbol{\gamma}+\lambda(k-1)-\frac{\sigma_{\epsilon 1}}{\sigma_1^2}(y_{it}-\mathbf{x}_{it}\boldsymbol{\beta}_1)\right)}{\sqrt{1-\frac{\sigma_{\epsilon 1}^2}{\sigma_1^2}}}\right]}{\Phi\left(\mathbf{z}_{it}\boldsymbol{\gamma}+\lambda(k-1)\right)}, k = 1, 2.$$

Similarly,

$$f(y_{it}|ww_{it}=k2) = f(y_{it}) \frac{\Phi\left[\frac{-\left(\mathbf{z}_{it}\boldsymbol{\gamma}+\lambda(k-1)-\frac{\sigma_{\epsilon 2}}{\sigma_2^2}(y_{it}-\mathbf{x}_{it}\boldsymbol{\beta}_2)\right)}{\sqrt{1-\frac{\sigma_{\epsilon 2}^2}{\sigma_2^2}}}\right]}{\Phi\left[-\left(\mathbf{z}_{it}\boldsymbol{\gamma}+\lambda(k-1)\right)\right]}, k = 1, 2.$$

Hidden Markov Model with Time Varying Transition Matrix, endogeneity and random effects ($\mathcal{M}_9$)

For a panel data set, a natural extension of the previous model is obtained by adding random effects in the components using the specifications in Eqs. (30) - (32). The main difference between the likelihood of the current model and that of model $\mathcal{M}_8$ is the

introduction of a double integral in the former. More formally, if for each unit the response variables are assumed to be independent conditional on the random effects and if one maintains the assumption that the units are independent, the marginal likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\int {}_i\pi \lambda(y_{i1}) \left(\prod_{t=2}^{T} \boldsymbol{\gamma}_{it} \lambda(y_{it}) \right) \mathbf{1}' f(\xi_{i1}, \xi_{i2}) d\xi_{i1} d\xi_{i2} \right].$$

Model Identification

The parameters of all the models previously presented are not automatically identified. In theory the log-likelihoods are all unbounded and a maximum likelihood estimator may not exist. Also, they all suffer from non-identification due to *label switching*. The log-likelihood is invariant under the permutation of the components which will make it difficult to dissociate the unconstrained component from the constrained component.

As suggested in the literature (Fruhwirth-Schnatter 2006), this identification problem can be solved by the use of a set of constraints. These constraints may come from economic theory. In the case of firms' physical investment one may be tempted to argue that a firm that has no trouble financing its investment activities should have a higher investment to capital ratio than when it has trouble obtaining funds, ceteris paribus. However, economic theory can only support the idea that a constrained firm is likely to choose a rate of investment below its optimal rate. Given the heterogeneity of the firms, it is possible that the majority of the constrained firms has a higher optimal rate of investment than the unconstrained ones. As a result, the previous constraint would be misleading. Thus, identification constraints should be chosen with care.

Another identification problem is associated with the use of a mixture model of too many components (overfitting). If the data set is generated by a single component, attempting to fit a mixture of two components may produce a component with a very small number of observations. In the case of a mixture with constant mixing proportions, the weight of each component will be very close to zero. As a consequence the log-likelihood will be approximately the same for any choice of parameters associated to that component.

Another issue that makes the identification of the parameters of these models difficult is the fact that the log-likelihoods are generally multimodal. Since the optimizers that will be used to maximize the log-likelihood can only find local maxima, the parameters estimates will be highly dependent on the starting values. To deal with this problem the log-likelihood maximization will be repeated several times with different starting values and the parameters estimates will be chosen to be the vector of estimates that corresponds to the highest log-likelihood assuming that it does not have the characteristics of a spurious maximizer. Each time the starting values are generated using either the *K-means* clustering algorithm (MacQueen 1967; Fink 2007) or a random classification scheme where each observation is randomly assigned to one group by flipping a fair coin. Note that the *K-means* algorithm does not produce the same classification at each run since the initial assignments are random. With these procedures, I try to increase the probability of finding a vector of starting values that falls in the bassin of attraction of the highest log-likelihood.

Inferences

Inferences will be based on the asymptotic properties of the maximum likelihood estimator. As discussed in the previous section, the likelihood of the models presented in this paper do not have an absolute maximum. However, for model $\mathcal{M}_1$, Kiefer (1978) has showed that it is possible to find a closed set that contains the true value of the vector of parameters in which there exists a unique consistent estimator. One requirement for this set is that it does not contain $\pi = 0$, $\pi = 1$, $\sigma_2 = 0$, and $\sigma_1 = 0$. That estimator is asymptotically normal with a covariance matrix equal to the inverse of the information matrix. Choi and Zhou (2002) proved similar results for a class of models with covariate-dependent mixing proportions.

Douc and Mathias (2001) prove the consistency and the asymptotic normality of the maximum likelihood estimator of a general hidden Markov model for both stationary and non-stationary Markov chains. The asymptotic covariance is, as usual, the inverse of the information matrix.

Robust Standard errors

According to the results stated above the standard errors of the estimated parameters can be obtained by taking the square root of the diagonal of the negative inverse of Hessian of the log-likelihood. However, the target applications are panel data. Since the likelihoods of the mixture models ($\mathcal{M}_1$-$\mathcal{M}_4$) ignore the time series properties of the data, dynamic misspecification is likely to be an issue. As a result, robust standard errors should be provided. These standard errors can be estimated using the following sandwich form

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\nabla^2 L_{it}(\hat{\boldsymbol{\theta}})\right)^{-1}\hat{\mathbf{B}}\left(\sum_{i=1}^{N}\sum_{t=1}^{T_i}\nabla^2 L_{it}(\hat{\boldsymbol{\theta}})\right)^{-1},$$

where $\nabla^2 L_{it}(\hat{\boldsymbol{\theta}})$ is the Hessian of the log-likelihood for the observation associated to firm i at time t evaluated at the maximum likelihood estimator. $\hat{\mathbf{B}}$ can be computed as in Wooldridge (2002)

$$\hat{\mathbf{B}} = \sum_{i=1}^{N}\sum_{t=1}^{T_i}\left(\nabla L_{it}(\hat{\boldsymbol{\theta}})\right)'\nabla L_{it}(\hat{\boldsymbol{\theta}}) + \sum_{i=1}^{N}\sum_{r\neq s}\left(\nabla L_{ir}(\hat{\boldsymbol{\theta}})\right)'\nabla L_{is}(\hat{\boldsymbol{\theta}}),$$

where $\nabla L_{it}(\hat{\boldsymbol{\theta}})$ is a row vector containing the gradient of the log-likelihood for firm i at time t. In the preceding case the firm identification variable is used as a cluster variable.

If the sample is relatively small one can alternatively use parametric or nonparametric bootstrap. In the nonparametric case an appropriate resampling method is *Moving Blocks Bootstrap* as described in Cameron and Trivedi (2005). Nevertheless, for the hidden Markov models where the time series properties of the data are very important resampling among the units as proposed by Kapetanios (2008) may even be more appropriate.

I should note that for the models considered in this paper bootstrapping requires some care. The likelihoods being potentially multimodal the highest local maximum may not be reached at each repetition.

Statistical tests

The statistical tests that will be considered have four objectives: 1) to determine the number of components of the mixtures, 2) to choose the best mixture among the models with a given number of components, 3) to test for endogeneity and 4) to test for random effects.

As stated in McLachlan and Peel (2000) choosing the number of components for a mixture is difficult. The preceding authors provide a long discussion about this issue in their book. One important problem is that in some cases one may not be able to find evidence that favors a model of a given number of components over another model that contains more or fewer components. In such situations they advocate choosing the model with the smaller number of components.

For the applications targeted in this paper the possible number of components will be inferred from economic theory. The main issue will then be how to find the distribution of the chosen test statistic under the null hypothesis.

Let k_x and k_z be respectively the dimension of the row vector $\mathbf{x}_{it}$ and the row vector $\mathbf{z}_{it}$. Let Ω_m be the parameter space of model $\mathcal{M}_m$, m=1,...,9.

$$\Omega_1 = \left\{(\pi, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1) : (\pi, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1) \in [0,1] \times \Re^{k_x} \times \Re_+ \times \Re^{k_x} \times \Re_+ \right\} \quad (38)$$

$$\Omega_2 = \left\{(\boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1) : (\boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1) \in \Re^{k_z} \times \Re^{k_x} \times \Re_+ \times \Re^{k_x} \times \Re_+ \right\} \quad (39)$$

$$\begin{aligned}\Omega_3 = \Big\{&(\boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \rho_0, \rho_1) : (\boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \rho_0, \rho_1) \in \Re^{k_z} \times \Re^{k_x} \times \Re_+ \\ &\times \Re^{k_x} \times \Re_+ \times [-1,1] \times [-1,1] \Big\} \end{aligned} \quad (40)$$

$$\begin{aligned}\Omega_4 = \Big\{&(\boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \rho_0, \rho_1, \boldsymbol{\Sigma}) : (\boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \rho_0, \rho_1, \boldsymbol{\Sigma}) \in \Re^{k_z} \quad (41) \\ &\times \Re^{k_x} \times \Re_+ \times \Re^{k_x} \times \Re_+ \times [-1,1] \times [-1,1] \times \mathbf{P}(2) \Big\} \quad (42) \end{aligned}$$

$$\begin{aligned}\Omega_5 = \Big\{&(\pi, p_{00}, p_{11}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1) : (\pi, p_{00}, p_{11}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1) \in [0,1] \times [0,1] \times [0,1] \\ &\times \Re^{k_x} \times \Re_+ \times \Re^{k_x} \times \Re_+ \Big\} \end{aligned} \quad (43)$$

$$\Omega_6 = \left\{(\pi, \boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1) : (\pi, \boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1) \in [0,1] \times \Re^{k_z} \times \Re^{k_x} \times \Re_+ \times \Re^{k_x} \times \Re_+ \right\} \quad (44)$$

$$\begin{aligned}\Omega_7 = \Big\{&(\pi, \boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \boldsymbol{\Sigma}) : (\pi, \boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \boldsymbol{\Sigma}) \in [0,1] \times \Re^{k_z} \times \Re^{k_x} \times \Re_+ \times \Re^{k_x} \\ &\times \Re_+ \times \mathbf{P}(2) \Big\}. \end{aligned} \quad (45)$$

$$\begin{aligned}\Omega_8 = \Big\{&(\pi, \boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \sigma_{\epsilon 2}, \sigma_{\epsilon 1}) : (\pi, \boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \sigma_{\epsilon 2}, \sigma_{\epsilon 1}) \in [0,1] \times \Re^{k_z} \times \Re^{k_x} \\ &\times \Re_+ \times \Re^{k_x} \times \Re_+ \times \Re \times \Re \Big\} \end{aligned} \quad (46)$$

$$\begin{aligned}\Omega_9 = \Big\{&(\pi, \boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \sigma_{\epsilon 2}, \sigma_{\epsilon 1}, \boldsymbol{\Sigma}) : (\pi, \boldsymbol{\gamma}, \boldsymbol{\beta}_2, \sigma_2, \boldsymbol{\beta}_1, \sigma_1, \sigma_{\epsilon 2}, \sigma_{\epsilon 1}, \boldsymbol{\Sigma}) \in [0,1] \\ &\times \Re^{k_z} \times \Re^{k_x} \times \Re_+ \times \Re^{k_x} \times \Re_+ \times \Re \times \Re \times \mathbf{P}(2) \Big\} \end{aligned} \quad (47)$$

$\mathbf{P}(2)$ is the space of positive definite matrices of dimension 2.

I want to test the hypothesis of a one-component model versus a two-component model represented by any of $\mathcal{M}_1 - \mathcal{M}_9$. The null hypothesis can be stated as

$$H_0 : g = 1,$$

which means

$\pi = 0$ for $\mathcal{M}_1$

at least one element of $\boldsymbol{\gamma}$ is infinite for $\mathcal{M}_2$

at least one element of $\boldsymbol{\gamma}$ is infinite for $\mathcal{M}_3$

at least one element of $\boldsymbol{\gamma}$ is infinite for $\mathcal{M}_4$

$p_{00} = 1, p_{11} = 0$ or $p_{00} = 0, p_{11} = 1$ for $\mathcal{M}_5$

$\lambda = 0$ and at least one element of $\boldsymbol{\gamma}$ is infinite for $\mathcal{M}_6 - \mathcal{M}_9$

In all cases the null hypothesis falls on the boundary of the parameter space as can be seen from Eq. (38) to Eq. (47). As a consequence the regularity conditions used to derive the asymptotic distribution of the likelihood ratio test break down. Note also that under H_0 the parameters of the component distribution with zero mixing proportion are not identifiable. The asymptotic distribution of the likelihood ratio test is not the expected χ^2 distribution. For example, in the case of a one-component binomial distribution versus a two-component distribution Chernoff and Lander (1995) show that the distribution of twice the logarithm of the likelihood ratio is a mixture of three distributions, two of them are χ^2. Goffinet and Loisel (1992) found similar non standard results. A review of these issues can be found in McLachlan and Peel (2000).

Since the asymptotic distribution of the likelihood ratio is not standard, an interesting alternative approach is to empirically approximate the distribution of this statistic. This can be done using parametric bootstrap (McLachlan and Krishman 1997; Davidson and Hinkley 1997). This can be done as follows:

1. Compute the maximum likelihood estimator $(\boldsymbol{\beta}, \sigma)$ for the one-component model.
2. Generate a sample $y^*_{it}, t = 1, ..T_i, i = 1, \ldots, N$ from $\phi(\mathbf{x}_{it}\boldsymbol{\beta}, \sigma)$.
3. Use y^*_{it} and the other covariates to obtain $(\boldsymbol{\beta}_m, \sigma_m)$ for the one-component model and $\boldsymbol{\theta}_m$ for the alternative two-component model.
4. Use these parameters to compute the likelihood ratios t_m.
5. Repeat this process 999 times to obtain a sequence $\{t_m\}_{m=1}^{999}$.

The p-value for the test is then computed as

$$p = \frac{1 + \#\{t_m > t\}}{1000},$$

where t is the observed likelihood ratio. Note that for the random effect models $\mathcal{M}_4$ and $\mathcal{M}_7$ this procedure is likely to be time consuming because of the computation of the double integral. An alternative is to choose information criteria such as the Akaike Information criterion (AIC) and the Bayesian Information criterion (BIC).

The test for endogeneity is essentially the test of model $\mathcal{M}_3$ versus model $\mathcal{M}_2$. The null hypothesis can be stated as

$$H_0 : \rho_1 = 0 \text{ and } \rho_2 = 0.$$

The alternative hypothesis is that at least one of the coefficients of correlation is different from zero. As can be seen from Eq. (40) the boundary problem no longer exists and twice the likelihood ratio statistic has a chi-square distribution with two degrees of freedom. Alternatively, the test can also be conducted using a t-statistic.

The test for the presence or absence of random effects is also problematic. The same boundary problem discussed above is encountered. The null hypothesis of no random effect can be stated as follows:

$$H_0 : \mathbf{\Sigma} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The preceding matrix is positive semi-definite and H_0 falls on the boundary of the parameter spaces Ω_4 and Ω_7. As before the distribution of the likelihood ratio statistics is not the expected χ^2 distribution. Stram and Lee (1994) have studied this problem for one-component linear models and showed that the asymptotic distribution of the likelihood ratio statistic is a mixture of chi-square distributions.

The next important test to consider is the test of a independent mixture versus a dependent mixture (HMM). This corresponds to the test of model $\mathcal{M}_1$ versus $\mathcal{M}_2$, and $\mathcal{M}_2$ versus $\mathcal{M}_6$. In the first case the null hypothesis is

$$H_0 : \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} p_{00} & p_{01} \\ p_{00} & p_{01} \end{pmatrix} \text{ and } (\pi_0, \pi_1) = (p_{00}, p_{01}),$$

and in the second case

$$\begin{bmatrix} 1 - \Phi(\mathbf{z}_{i1}\boldsymbol{\gamma}) & \Phi(\mathbf{z}_{i1}\boldsymbol{\gamma}) \\ 1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}) & \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}) \end{bmatrix} \text{ and } (\pi_0, \pi_1) = (1 - \Phi(\mathbf{z}_{it}\boldsymbol{\gamma}), \Phi(\mathbf{z}_{it}\boldsymbol{\gamma})),$$

or

$$H_0 : \lambda = 0.$$

In both cases the rows of the transition matrices are the same under the null hypothesis. The asymptotic null distribution of the likelihood ratio is valid in theses cases. In the case where $\lambda = 0$ under the null hypothesis a t-test is also appropriate.

Application: Firms' investment and financing constraints

The basic intertemporal investment model by Hayashi (1982) assumes that a firm chooses the level of its next period capital stock by maximizing the expected discounted value of dividends. In reality, it is not always possible for certain firms to finance the level of investment that maximizes profit. This situation may arise because of the existence of information asymmetry between the firm's managers and the potential suppliers of funds. Without the ability to evaluate accurately the profitability of the firm's projects, the suppliers of funds may be unwilling to finance the firm's investment or they may be willing to supply only a fraction of the funds needed by the firm. As a result, investment may not be financed to the level that is optimal in the absence of constraints. One way of accounting for this issue is by adding a borrowing constraint to the Hayashi (1982) model (Adda and Cooper 2003). The Euler equation from the resulting model would imply two different relationships between investment and its determinants depending on whether the constraint is binding or not. If this model is a good approximation for a firm's investment behavior, at each point in time the firm will fall in one of two groups: the group of firms that are financially constrained (borrowing constraint is binding) and the group of firms that are not financially constrained. Since the observed data do not generally include any

variable that indicates group membership, this setting is well suited for the use of finite mixture models of the kinds presented in this paper.

In this application two variables are modeled: the change in firm *i*'s investment to capital ratio at time t (ΔI_{it}), and the financial status of the firm at time t (W_{it}). Given the potential interdependence of the variables, they will be modeled as a bivariate processs $(\Delta I_{it}, w_{it})'$, t=1,...,T_i,i=1,... n. Under the assumption that at any point in time, a firm can be either financially constrained or not, W_{it} is an unobserved dichotomous random variable. Assuming that the models for ΔI_{it} are obtained by taking the first difference of the models in level for I_{it}, individual-specific effects or random effects will not appear in the models for ΔI_{it}. As a result, the most appropriate models to estimate are $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, $\mathcal{M}_5$, $\mathcal{M}_6$, and $\mathcal{M}_8$. I estimated those models with data on 2263 US manufacturing firms obtained from the COMPUSTAT dataset for the period between 1974 and 2005. To compare the results with those obtained from some previous studies, I have generated the data almost exactly as stated in Hovakimian and Titman (2006). The definitions of the main variables are presented in Table 1.

Bootstrap likelihood ratio tests of one component versus two show a two-component distribution is favored over a one-component one in all considered cases.

Table 2 presents regular likelihood ratio tests comparing the different two-component models.

It is clear that the more flexible models are favored in all the cases. The evidence appears to be overwhelming (observed likelihood ratio are higher than 710) in all the cases except in the case of model $\mathcal{M}_2$ versus $\mathcal{M}_3$ (observed likelihood ratio is 19.5); also, Table 3 shows that only the coefficient of correlation between the second component and the choice equation is statistically significant. As a result, the addition of endogeneity may not have caused a big improvement in the fit of the data to this model. However, the correlation coefficients are both significant for model $\mathcal{M}_8$ (Table 3).

The most important result here is that the HMM model $\mathcal{M}_5$ strongly outperforms the mixture model $\mathcal{M}_1$ and the same is true for the HMM model $\mathcal{M}_6$ versus the mixture model $\mathcal{M}_2$. Moreover, the test statistic on lambda (See transition matrix of model $\mathcal{M}_6$) is quite large (31.25) which implies that this parameter is significantly different from 0 and

Table 1 Variables definitions

Variables	Definitions
INVESTMENT	Investment ÷ (beginning of period capital stock)
GROWTH OPPORTUNITIES	(Market value of the firm's asset − common equity and deferred taxes) ÷ (Book value)
CASHFLOWS	(Income before extraordinary items + Depreciation) ÷ (Beginning of period capital)
LOGBOOKASSET	Log(Value of Assets adjusted for inflation)
SHORTTERMDEBT	(Short Term Debt) ÷ (Firm's Assets)
LONGTERMDEBT	(Long Term Debt) ÷ (Firm's Assets)
FINANCIAL SLACK	Cash and short term investment ÷ previous year Assets
DUMMYDIVPAYOUT	Equal to 1 if firm pays dividend, 0 otherwise
DUMMYBONDRATING	Equal to one if firm has bond rating 0 otherwise
COVERAGE RATIO	Interests ÷ Earnings Before Interest
ASSET SALES	(Sales of property, plant and Equipment) ÷ beginning of period capital

Note: Using the COMPUSTAT Xpressfeed data items, the above variables are defined as follows: Investment=*capx*, capital=*ppent*, (Market value of Assets)=*prcc_c*csho*, (Common Equity)=*ceq*, (Book value of assets)=*at*, (Income Before Extraordinary Items) =*ib*, Depreciation=*dp*, Short Term Debt=*dlc*, Long Term Debt = *dltt*, (Cash and short term investment)=*che*,Dividends =*dv*, Interest=*xint*, Earnings Before Interest=*ebidta*, (Sales of Property, plant and equipments)=*sppe*

Table 2 Likelihood ratio tests comparing the two-component models

	$\mathcal{M}_1$ vs $\mathcal{M}_2$	$\mathcal{M}_2$ vs $\mathcal{M}_3$	$\mathcal{M}_1$ vs $\mathcal{M}_5$	$\mathcal{M}_2$ vs $\mathcal{M}_6$	$\mathcal{M}_5$ vs $\mathcal{M}_6$
Likelihood ratio	1833.337	19.504	1890.253	767.080	710.163
Degrees of freedom	7.000	2.000	9.000	2.000	7.000
pvalue	0.000	0.000	0.000	0.000	0.000

Notes: These are regular likelihood ratio tests. The likelihood ratio is 2*(L_2-L_1) where L_2 is the log-likelihood of the bigger model and L_1, the log-likelihood of the smaller model. Under the null hypothesis that the smaller model is true, this statistic has a $\chi^2(k_2 - k_1)$ distribution where k_2 is the number of parameters from the bigger model and k_1, the number of parameters from the smaller model

reinforces the idea that the firms financial states are time-dependent. Of the two hidden Markov models, the likelihood ratio test reveals that the best one is the one that allows for a covariate-dependent transition matrix.

The results of the likelihood ratio tests are also confirmed by the information criteria AIC and BIC since the most flexible Hidden Markov Model shows the lowest values. Moreover, these criteria make possible the comparison between the non-nested models $\mathcal{M}_3$ and $\mathcal{M}_6$ and models $\mathcal{M}_3$ and $\mathcal{M}_5$. Even though the hidden Markov models do not account for endogeneity they fit the data much better than the endogeneneous mixture generally used in the literature. Neglecting time-dependence is then more problematic than neglecting endogeneity.

Nevertheless, even though it is clear that a two-component distribution fits the data better, it is not obvious which component should be labeled as financially constrained. So,

Table 3 Estimates of the parameters of the components distributions for Models $\mathcal{M}_1$, $\mathcal{M}_3$, and $\mathcal{M}_8$

	CF_t[a]	CF_{t-1}	AS_{t+1}[b]	AS_t	AS_{t-1}	sigma	π	ρ
Model $\mathcal{M}_1$								
Component 1								
Estimates	0.274	0.004	0.180	0.238	0.125	0.281	0.333	___
ste	0.012	0.003	0.055	0.061	0.055	0.003	0.005	___
Component 2								
Estimates	0.113	0.016	0.023	0.042	0.027	0.073	0.667	___
ste	0.005	0.002	0.013	0.015	0.015	0.001	0.005	___
Model $\mathcal{M}_3$								
Component 1								
Estimates	0.136	0.040	0.024	0.049	0.032	0.071		0.031
ste	0.006	0.004	0.013	0.020	0.018	0.001		0.050
Component 2								
Estimates	0.231	0.006	0.143	0.156	0.145	0.281		-0.146
ste	0.010	0.003	0.055	0.052	0.051	0.003		0.050
Model $\mathcal{M}_8$								
Component 1								
Estimates	0.225	0.004	0.137	0.133	0.124	0.291	0.491	0.447
ste	0.012	0.003	0.257	0.577	0.571	0.006	0.387	0.126
Component 2								
Estimates	0.139	0.037	0.024	0.049	0.036	0.072	0.509	0.293
ste	0.006	0.005	0.022	0.033	0.033	0.001	0.387	0.093

[a]Cash flows
[b]Asset Sales at time *t+1*
Notes: For each component, π is the prior probability of belonging to the component, σ is volatility of the change in investment for firms belonging to the component, and ρ is the correlation coefficient between the change in investment and financial status. The vector of explanatory variables *does not include lags of the dependent variable*, but includes time dummies and other control variables whose coefficients are not reported to save space. The dependent variable is the first difference of investment-to-capital ratio

to interpret the results in Table 3 I first need to find some criteria to label the components. For this reason I choose the identification criteria from the literature. Financially unconstrained firms are expected to be big, old and not highly leveraged; they are expected to pay dividends regularly and to have a bond rating; and they may face lower growth opportunities and may be less interested in carrying large cash balances. The justification of these criteria is reviewed in Hovakimian and Titman (2006). Applying these criteria to models $\mathcal{M}_2$ and $\mathcal{M}_3$ one can identify the financially constrained component as the one that has the largest standard deviation. The same is true for model $\mathcal{M}_6$.

The results then suggest that investment is more responsive to cash flow and asset sales in the financially constrained state as was signaled in Hovakimian and Titman (2006). Since the standard deviation of investment is much higher for the financially constrained group (0.28 versus 0.08), one can conclude that the change in fixed capital investment is much more volatile for firms that spend a long time in the financially constrained state.

The most important results come from the HMM models. The estimated prior transition matrix and the average of the posterior transition matrices for the time homogeneous model are

$$\hat{\mathbf{P}} = \begin{pmatrix} 0.776 & 0.224 \\ 0.068 & 0.932 \end{pmatrix}, (\bar{\mathbf{P}}|\Delta I_{it}) = \begin{pmatrix} 0.573 & 0.427 \\ 0.174 & 0.826 \end{pmatrix}.$$

The financially unconstrained state appears to be quite persistent. While a firm that is currently constrained has a higher probability to stay in that state next period, it also has a significant probability (43%) to become unconstrained.

Even though the Markov chain was not assumed to be stationary, the estimated transition matrices clearly admit a stationary distribution. The stationary probability vector associated with the second transition matrix is $(p_1, p_2) = (0.29, 0.71)$, which means that in the long run a higher proportion of the observations (71%) is expected to be classified as unconstrained. This is consistent with the estimated mixing proportions for model $\mathcal{M}_1$. Similar results are obtained for the exogenous HMM model with covariate-dependent transition matrix.

Conclusion

I have presented nine alternative mixture models that may be of interest for making inference from available economic panel data sets. I have also reviewed the maximum likelihood estimation of six of them via the well known Expectation-Maximization algorithm. A series of possible tests are also discussed. These tests can be use to identify among the proposed models the one that fits the data better.

Estimation of the hidden Markov models with random effects may be time consuming because, for each unit, the log-likelihood at each point in time depends on all the previous observations of that unit; moreover, this likelihood has to be computed repeatedly for each vector of abscissae or each vector of draws of the random effects. If, however, the log-likelihood is programmed in FORTRAN or C as opposed to MATLAB or R, the computation time may be reduced significantly, but performing bootstrap tests may still require a long time. Nevertheless, the models considered in this paper are very flexible and can be used to account for several potential sources of heterogeneity in panel data.

Finally, as an application I used the models without random effects to study the differences in the investment behavior of firms when they are financially constrained and when they are not, and also to learn about the process that governs the evolution of a firm's financial status over time.

Acknowledgement
I am grateful for all the help and support received from Professor Pravin Trivedi in completing this project. I am thankful for the suggestions obtained from two anonymous referees.

Funding
I did not receive any financial support in writing this paper.

Authors' contribution
I, Judex Hyppolite, am the only author of this paper.

Competing interests
I declare that I have no competing interests in publishing this paper.

Publisher's Note

Received: 4 January 2017 Accepted: 11 July 2017
Published online: 01 August 2017

References

Adda, J, Cooper, R: Dynamic Economics: quantitative methods and applications. The MIT Press, Cambridge (2003)
Almeida, H, Campello, M: Financial constraints, asset tangibility and corporate investment. Rev. Financ. Stud. **20**(5), 1429–1460 (2007)
Asea, PK, Blomberg, B: Lending cycles. J. Econ. **83**, 89–128 (1998)
Atman, RM: Mixed hidden markov models: An extension of the hidden markov model to the longitudinal data setting. J. Am. Stat. Assoc. **102**(477), 201–210 (2007)
Cameron, AC, Trivedi, P: Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge (2005)
Cappé, O, Moulines, E, Rydén, T: Inference in Hidden Markov Models. Springer, New York (2005)
Chernoff, H, Lander, E: Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. J. Stat. Plan. Infer. **43**, 19–40 (1995)
Choi, KC, Zhou, X: Large sample properties of mixture models with covariates for competing risks. J. Multivar. Anal. **82**, 331–366 (2002)
Davidson, AC, Hinkley, DV: Bootstrap methods and their applications. Cambridge University Press, Cambridge (1997)
Deb, P, Trivedi, P: Finite mixture for panels with fixed effects. J. Econ. Methods. **2**, 31–35 (2013)
Dempster, AP, Laird, NM, Rubin, DB: Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. **39**(1), 1–38 (1977)
Douc, R, Mathias, C: Asymptotics of the maximum likelihood estimator for general hidden markov models. Bernouilli. **3**, 381–420 (2001)
Fink, GA: Markov Models for Pattern Recognition: From Theory to Applications. 1st ed. Springer-Verlag, New York (2007)
Fruhwirth-Schnatter, S: Finite Mixture and Markov Switching Models. Springer-Verlag, New York (2006)
Goffinet, B, Loisel, P: Testing in normal mixture models when the proportions are known. Biometrika. **79**, 842–846 (1992)
González, J, Tuerlinckx, F, Boeck, PD, Cools, R: Numerical integration in logistic-normal models. Comput. Stat. Data Anal. **51**, 1525–1548 (2006)
Hamilton, JD: Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates. J. Econ. Dyn. Control. **12**, 385–423 (1988)
Hayashi, F: Tobin's marginal q and average q: A neoclassical interpretation. Econometrica. **50**, 215–224 (1982)
Hovakimian, G, Titman, S: Coporate investment with financial constraints: Sensitivity of investment to funds from voluntary asset sales. J. Money Credit Bank. **38**(2), 357–374 (2006)
Jäckel, P: A note on multivariate gauss-hermite quadrature (2005). http://www.btinternet.com/pjaeckel/ANoteOnMultivariateGaussHermiteQuadrature.pdf. Accessed 14 Nov 2009
Kapetanios, G: A bootstrap procedure for panel data sets with many cross-sectional units. Econ. J. **11**, 377–395 (2008)
Kiefer, NM: Discrete parameter variation: Efficient estimation of a switching regression model. Econometrica. **46**, 427–434 (1978)
Kim, C-J, Piger, J, Startz, R: Estimation of markov regime-switching regression models with endogenous switching. J. Econ. **143**, 263–273 (2008)
Lee, Y, Nelder, JA: Hierarchical generalized linear models. J. R. Stat. Soc. **58**, 619–678 (1996)
Louis, TA: Finding the observed information matrix when using the em algorithm. J. R. Stat. Soc. **44**, 226–233 (1982)
MacDonald, IL, Zucchini, W: Hidden Markov and Other Models for Discrete-valued Times Series. Chapman & Hall, Boca Raton (1997)
MacQueen, JB: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
Maddala, GS: Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge (1999)

Maruotti, A: Hidden Markov Models for Longitudinal Data. PhD thesis. Università degli Studi di Roma (2007)
McLachlan, G, Krishman, T: The EM Algorithm and Extensions. Wiley-Interscience, New York (1997)
McLachlan, G, Peel, D: Finite Mixture Models. 1st ed. Wiley-Interscience, New York (2000)
Mundlak, Y: On the pooling of time series and cross section data. Econometrica. **46**, 69–85 (1978)
Rabiner, LR: A tutorial on hidden markov models and selected applications in speech recognition. In: Proceedings of the IEEE, vol. 77, pp. 257–286 (1989)
Stram, DO, Lee, JW: Variance components testing in the longitudinal mixed effects model. Biometrics. **50**, 1171–1177 (1994)
Trivedi, P, Hyppolite, J: Alternative approaches for econometric analysis of panel count data using dynamic latent class models (with application to doctor visits data). Health Economics. **21**, 101–128 (2012)
Wooldridge, JM: Econometric Analysis of Cross Section And Panel Data. The MIT Press, Cambridge (2002)
Xiaoqiang, H, Schiantarelli, F: Investment and capital markets imperfections: A switching regression approach using u.s. firm panel data. Rev. Econ. Stat. **80**(3), 466–479 (1998)
Zhang, S, Jin, J: Computation of special functions. Wiley-Interscience, New York (1996)