ORIGINAL PAPER

S-estimation of hidden Markov models

Alessio Farcomeni · Luca Greco

Abstract A method for robust estimation of dynamic mixtures of multivariate distributions is proposed. The EM algorithm is modified by replacing the classical M-step with high breakdown S-estimation of location and scatter, performed by using the bisquare multivariate S-estimator. Estimates are obtained by solving a system of estimating equations that are characterized by component specific sets of weights, based on robust Mahalanobis-type distances. Convergence of the resulting algorithm is proved and its finite sample behavior is investigated by means of a brief simulation study and n application to a multivariate time series of daily returns for seven stock markets.

Keywords Bisquare · EM · HMM · Mahalanobis distance · Mixture · Robust distance · S-estimation

1 Introduction

A hidden Markov model (HMM) is a flexible and general model for time series with applications in many fields (see, e.g., MacDonald and Zucchini 1997). Multivariate HMMs can also be used to monitor the simultaneous evolution of many dependent and independent time series. In that context the model is often referred to as Latent Markov model (Bartolucci et al. 2013). The basic formulation of a HMM relies on two assumptions: first, the observation at time t is independent of previous observa-

A. Farcomeni
Dipartimento di Sanitá Pubblica e Malattie Infettive, Sapienza - University of Rome,
Rome, Italy

L. Greco (✉)
DEMM Department, University of Sannio, Benevento, Italy
e-mail: luca.greco@unisannio.it

tions conditionally on the value of an unobserved discrete random variable. Secondly, the latent discrete random process evolves according to a homogeneous first-order Markov chain. There are many generalizations of this framework, possibly including constraints, covariates, or relaxing some of the basic assumptions [for instance, Altman (2007), Yu (2010), Bartolucci and Farcomeni (2010), Bartolucci et al. (2013)]. HMMs can be seen as discrete mixture models in which the mixing proportions evolve over time. There are, then, k sampling distributions, identified by the levels of the latent process, often assumed to be multivariate normals with class-specific mean vector and covariance matrix.

The latter model specification may in practice be unsatisfactory since it can happen that a small fraction of data follows a different random mechanism, exhibits a different pattern or no pattern at all. These atypical values are called *outliers*. Outliers can lead to unreliable inference if they are not taken properly into account: observations located far from the bulk of the data may break down component specific parameter estimates, bridge points (e.g., points between two components) may force genuinely separate components to be artificially merged, with a consequent bias in the estimate of location and an inflation in the estimate of scatter. Outliers can be isolated *and/or* clustered in one or more spurious additional components, which may lead to over estimate the complexity of the underlying discrete latent variable.

The need for robust procedures in the estimation of mixture models has been first addressed in Cambpell (1984) who suggested to replace standard maximum likelihood with M-estimation, but in the case of a static mixture model. In the same framework of static mixtures, recent contributions can be found in Markatou (2000), who applied the weighted likelihood methodology, Neykov et al. (2007), who used the trimmed likelihood estimator, as well as Cuesta-Albertos et al. (2008) (see also Farcomeni 2013, 2014), and Gallegos and Ritter (2009) where two step procedures are proposed based on trimming and censoring. There also is a different approach in the literature, characterized by the use of flexible models. The Gaussian mixture assumption is relaxed by embedding it in a *supermodel*: important contributions can be found in McLachlan and Peel (2000), who introduced a mixture of Student's t distributions in place of the commonly used Gaussian components, Fraley and Raftery (1998), who considered an additional component modeled as a Poisson process to handle noisy data, Hennig (2004), who considered the addition of an improper uniform mixture component to improve breakdown point properties of maximum likelihood estimators. The use of an enlarged model is not always the best solution, because it may only be able to deal with outliers in some direction, therefore focusing on a possibly thin subspace of the possible departures from the assumed sampling model (Ronchetti 1997; Huber and Ronchetti 2009). In particular, the strategy of simply introducing an additional latent class may not solve the problem, as outliers may be scattered among underlying centroids, and very far apart. In the worst case, each outlier should be included in a separate latent class, making the model not estimable and overly complex.

Despite probably the first instance of the EM algorithm was developed for estimation in HMMs (Baum and Petrie 1966; Baum et al. 1970), there are not many papers dealing with robustness issues in HMMs. Moreover, up to our knowledge, all available approaches to date provide robust solutions based on a *supermodel* that includes the dynamic Gaussian mixture as a special case; with the single exception

of Maruotti (2014) in the context of hidden Markov *univariate regression* models. Among the proposed solutions, we can mention those based on the use of the mixture of multivariate Student's *t* distribution (Humburg et al. 2008; Bulla 2011). Another interesting proposal has been suggested by Shah et al. (2006) in one dimension. Shah et al. (2006) split each conditional density of the mixture into two components, one of which is aimed at handling outliers. They obtain inference in a Bayesian context and a multivariate extension of their method is beyond the scope of our paper.

In this paper we propose a formal robust strategy for estimation in HMMs, which does not rely on flexible modeling, but on the use of robust S-estimators at the M-step of the EM algorithm. This technique provides estimates that are close to maximum likelihood estimates (MLE) under normality both in terms of accuracy and precision and that are insensitive to small departures from the model assumptions.

The idea of employing S-estimators in the EM algorithm has already been considered by Bashir and Carter (2005, 2007) in the context of discriminant analysis under the assumption of a common covariance matrix. Here, we generalize their approach to any finite mixture model, without assumptions on the covariance matrices, with a specific interest on S-estimation of HMMs. A sample non-optimized R code can be found at http://afarcome.altervista.org/ES.r. The rest of the paper is organized as follows. Necessary background is reviewed in Sect. 2. In Sect. 3 we illustrate an Expectation S-estimation (ES) algorithm for robust fitting of HMMs and show its convergence. Some numerical studies are given in Sect. 4 and a real data example is discussed in Sect. 5. We give a brief discussion in Sect. 6.

2 Set up and background

In this section we briefly review multivariate S-estimation of location and covariance matrix (see Maronna et al. (2006) and references therein for a more detailed account), with particular attention to the case where data arise from multivariate normal distributions. Let $y = (y_1, \ldots, y_n)$ be a matrix of i.i.d. observations from a family of multivariate normal distributions

$$\mathcal{F} = \left\{ f(y; \theta) = \frac{h(d(y; \mu, \Sigma))}{\sqrt{|\Sigma|}}, \theta = (\mu, \Sigma), \mu \in \mathbf{R}^p, \Sigma \in PDS(p), p > 1 \right\}, \tag{1}$$

where $h(d) = (2\pi)^{-p/2} \exp\left(-\frac{d^2}{2}\right)$, $d(y; \mu, \Sigma) = [(y - \mu)^T \Sigma^{-1} (y - \mu)]^{1/2}$ is the Mahalanobis distance and $PDS(p)$ is the set of all positive definite symmetric $p \times p$ matrices. Any different choice of a positive function $h(\cdot)$, with strictly negative derivative and such that $f(y; \theta)$ integrates to unity leads to an elliptically symmetric family of distributions. A relevant example of non normal elliptical distribution is the multivariate Student's T distribution.

The MLE of $\theta = (\mu, \Sigma)$ can be found by maximizing the log-likelihood function

$$\ell(\theta) = -\frac{n}{2} \log |\Sigma| - \sum_{i=1}^{n} \rho(d_i), \tag{2}$$

with $d_i = d(y_i; \mu, \Sigma)$, $\rho(d) = -\log h(d)$. The MLE can equivalently be obtained by the solution to the system of estimating equations

$$\Psi_\mu(y; \theta) = \sum_{i=1}^n (y_i - \mu) w_i = 0$$

$$\Psi_\Sigma(y; \theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \mu)(y_i - \mu)^T w_i - \Sigma = 0, \tag{3}$$

with $w = w(d)$ given by $(\partial\rho(d)/\partial d)/d$. In the case of the normal distribution, we have that $w_i = 1$. The system of estimating Eq. (3) defines an M-estimator of location and scatter. See "Appendix 1" for a derivation of (3).

S-estimates of location and scatter (Davies 1987; Lopuhaa 1989) are defined as the minimizer of the determinant $|\Sigma|$ subject to the bound

$$\frac{1}{n} \sum_{i=1}^n \rho(d_i; c_0) = \delta, \ 0 < \delta < \sup \rho(d; c_0). \tag{4}$$

A popular choice is the Tukey's bisquare function

$$\rho(d; c_0) = \frac{c_0^2}{6} \min \left\{ 1, 1 - \left[1 - \left(\frac{d}{c_0} \right)^2 \right]^3 \right\}. \tag{5}$$

The solution to the constrained minimization above also satisfies the system of M-type estimating equations of the form

$$\Psi(y; \theta) = (\Psi_\mu(y; \theta), \Psi_\Sigma(y; \theta))^T = \mathbf{0}$$

with

$$\Psi_\mu(y; \theta) = \sum_{i=1}^n (y_i - \mu) w(d_i) = 0$$

$$\Psi_\Sigma(y; \theta) = \sum_{i=1}^n \left[p(y_i - \mu)(y_i - \mu)^T w(d_i) - v(d_i)\Sigma \right] = 0, \tag{6}$$

and

$$w(z) = \frac{1}{z} \frac{\partial\rho(z; c_0)}{\partial z} = \left(1 - z^2 \right)^2 I(z \le c_0),$$

$$v(z) = w(z)z^2. \tag{7}$$

A detailed derivation of (6) is given in "Appendix 2". The constant c_0 in (5) is fixed according to the following considerations. Define the asymptotic breakdown point

(BP) of an estimate as the minimal proportion of outliers the data may contain before that its bias becomes unbounded. In order to attain a given asymptotic BP $b \leq 0.5$ one should set c_0 as the solution to

$$E_F[\rho(Z; c_0)] = \delta = b \sup \rho = b \frac{c_0^2}{6}. \tag{8}$$

Some loss of efficiency with respect to classical M-estimates is inevitable, and the higher the BP, the lower the efficiency.

Existence, consistency, asymptotic normality and breakdown properties of S-estimators have been investigated by Davies (1987). Some recent developments can be found in Riani et al. (2014). The influence function has been derived by Lopuhaa (1989). The system (3) may have multiple solutions, because the first derivative of $\rho(\cdot)$ is a re-descending function (see Maronna et al. 2006), and one of them is the S-estimate that solves the minimization problem (4) (Lopuhaa 1989).

In the following we will focus on the multivariate normal case, where $F = \Phi$, but the S-estimation procedure, and therefore our ES algorithm, is still valid for any F within the elliptically symmetric family.

3 Robust estimation in HMM

Let now $y = (y_1, y_2, \ldots, y_n)$ be a multivariate time series of repeated observations at n consecutive occasions. We assume $y_i \in \mathbb{R}^p$ and that the marginal density of y_i is an element of the family (1).

Hidden Markov modeling proceeds by assuming that at each time occasion there is a binary latent random vector $z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$, with $\sum_j z_{ij} = 1$, which affects the distribution of y. This discrete latent variable is unobserved and evolves over time according to a first order time homogeneous Markov chain. More precisely,

$$p(y_i | \theta, z_{ij} = 1) = f(y_i, \mu_j, \Sigma_j).$$

By denoting $\pi_{ij} = \Pr(Z_{ij} = 1)$, we can set the model in the form of a finite mixture of k distributions as

$$p(y_i | \tau_i) = \sum_{j=1}^{k} \pi_{ij} f(y_i; \mu_j, \Sigma_j), \tag{9}$$

where $\tau_i = (\theta_1, \ldots, \theta_k, \pi_{i1}, \ldots, \pi_{ik})$. The main feature of expression (9) is that the mixing probabilities π_{ij} are specific to time occasion i and not held fixed for each component, as it happens in the context of static mixtures. In HMM these probabilities are not assumed to evolve freely over time, but to follow those connected with a first order homogeneous Markov chain. The number of free parameters is therefore drastically reduced to the $k(k-1)$ transition probabilities $\Pr(Z_{ij_2} = 1 | Z_{i-1, j_1} = 1) = \pi_{j_1 j_2}, j_1, j_2 = 1, \ldots, k$, collected in the hidden transition probability matrix Π. Identifiability is guaranteed as long as the latent state at the first occasion is arbitrarily fixed. Conventionally, this is the first latent state. The reason behind this constraint

resides in the need of conditioning on a baseline measurement in first order Markovian time series.

The vector π_i collecting the k probabilities at the i-th occasion is computed as

$$\pi_i = \eta \Pi^{i-1},$$

where η is a vector of zeros with a one indicating the (fixed and arbitrarily chosen) latent state at the first occasion. For an element of the family (1) defined by a positive function $h(\cdot)$, the complete data log-likelihood for the proposed HMM can be written as

$$
\begin{aligned}
\ell^c(\tau) &= \sum_{i=2}^{n} \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} z_{ij_2} z_{i-1,j_1} \log \pi_{j_1 j_2} + \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \log f(y_i; \mu_j, \Sigma_j) \\
&= \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \log \pi_{j_1 j_2} \sum_{i=2}^{n} z_{ij_2} z_{i-1,j_1} - \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \left\{ \frac{1}{2} \log |\Sigma_j| + \rho(d_{ij}) \right\}.
\end{aligned}
\tag{10}
$$

The complete data log-likelihood is the basis for the commonly used EM algorithm, which is used to find the MLE. Robust estimation of τ is achieved by iterating the classical E step and a robust version of the standard M step involving S-estimation, that we call S-step. In the E-step conditional expected values for z_{ij} and $z_{ij_2} z_{i-1,j_1}$ are computed through appropriate recursions. These are the classical HMM recursions, adapted from Baum and Petrie (1966); Baum et al. (1970). First, one applies a forward recursion initialized with

$$\alpha_{1j} = f(y_1; \mu_j, \Sigma_j)\eta_j,$$

and iterated as

$$\alpha_{ij} = f(y_i; \mu_j, \Sigma_j) \sum_{h=1}^{k} \alpha_{i-1,h} \pi_{hj}, \quad i = 2, \ldots, n.$$

After that, a backward recursion is applied by setting $\beta_{nj} = 1$ and then

$$\beta_{ij} = \sum_{h=1}^{k} f(y_{i+1}; \mu_h, \Sigma_h)\beta_{i+1,h}\pi_{jh}, \quad i = n-1, \ldots, 1$$

It can be shown that

$$E[z_{ij}|y, \tau] = m_{ij} = \frac{\alpha_{ij}\beta_{ij}}{\sum_{j=1}^{k} \alpha_{nj}}.
\tag{11}$$

and

$$E[z_{ij_2} z_{i-1 j_1}|y, \tau] = m_{ij_1 j_2} = \frac{\pi_{j1j2}\alpha_{i-1,j_1} f(y_i; \mu_j, \Sigma_j)\beta_{ij_2}}{\sum_{j=1}^{k} \alpha_{nj}}, \quad i = 2, \ldots, n.
\tag{12}$$

To avoid numerical issues, the strategy in appendix of Farcomeni (2012) has been used. For a detailed rationale behind the forward and backward recursions and expressions (11) and (12), see for instance MacDonald and Zucchini (1997); Bartolucci et al. (2013). It can be noted that all these follow from the conditional independence assumptions of y_i given z_i, and of z_i and $z_1, \ldots, z_{i-2}$ given z_{i-1}

Given that its expression is linear in z_{ij} and $z_{ij_2} z_{i-1,j_1}$, the conditional expectation of (10) is obtained by plugging-in the conditional expected values (11) and (12), i.e.

$$
E\left[\ell^c(x; \tau)|y, \tau^{(s)}\right] = \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \log \pi_{j_1 j_2} \sum_{i=2}^{n} m_{ij_1 j_2}^{(s)}
$$
$$
- \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{(s)} \left\{ \frac{1}{2} \log |\Sigma_j| + \rho(d_{ij}) \right\}, \qquad (13)
$$

where the superscript (s) denotes the estimate at the sth iteration. In the S-step, the components of $\theta_j = (\mu_j, \Sigma_j)$ are estimated by multivariate S-estimators, while performing the M-step for π_1 and Π in a standard fashion. In detail, the transition probabilities are updated as $\pi_{j_1 j_2} \propto \sum_{i=2}^{n} m_{ij_1 j_2}$, whereas the estimates of μ_j and Σ_j are obtained by solving k minimization problems, one for each component. Within each class it is aimed at minimizing $|\Sigma_j|$ subject to the bound

$$
\frac{1}{n_j} \sum_{i=1}^{n} \rho(d_{ij}; c_0) m_{ij} = \delta, \qquad (14)
$$

with $n_j = \sum_{i=1}^{n} m_{ij}$. The S-step can be summarized as maximizing

$$
Q(\tau | \tau^{(s)}) = \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \log \pi_{j_1 j_2} \sum_{i=2}^{n} m_{ij_1 j_2}^{(s)}
$$
$$
- \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{(s)} \left\{ \frac{1}{2} \log |\Sigma_j| - \lambda_j \left[\rho(d_{ij}; c_0) - \delta \right] \right\}
$$
$$
= \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \log \pi_{j_1 j_2} \sum_{i=2}^{n} m_{ij_1 j_2}^{(s)} - \sum_{j=1}^{k} L_{n_j^{(s)}}(\theta_j, \lambda_j). \qquad (15)
$$

It is worth noting that we stated the S-estimators as in (15), rather than in a more classical form (Lopuhaä 1989), in order to stress the connection with expression (13). The resulting estimating equations for $\theta_j = (\mu_j, \Sigma_j)$ are similar to (6):

$$
\Psi_{\mu_j}(y; \theta_j) = \sum_{i=1}^{n} (y - \mu_j) w_{ij}^{(s)} m_{ij}^{(s)} = 0
$$

$$\Psi_{\Sigma_j}(y;\theta_j) = \sum_{i=1}^{n} \left[p(y_i - \mu_j)(y_i - \mu_j)^T w_{ij}^{(s)} - v_{ij}^{(s)} \Sigma_j \right] m_{ij}^{(s)} = 0, \quad (16)$$

with $w_{ij}^{(s)} = w(d_{ij}^{(s)})$, $v_{ij}^{(s)} = v(d_{ij}^{(s)})$, $d_{ij}^{(s)} = d(y_i; \mu_j^{(s)}, \Sigma_j^{(s)})$, $j = 1, 2, \ldots, k$, as defined in (7).

For details on the derivation of (16), refer to "Appendix 3".

The ES algorithm requires that k distinct S-estimation problems are solved. These depend not only on the estimated m_{ij} but also on k component-wise sets of weights w_{ij}.

The resulting ES algorithm can be summarized as follows:

Initialization. $\tau^{(s)} = (\theta_1^{(s)}, \ldots, \theta_k^{(s)}, \pi_1^{(s)}, \pi_k^{(s)})$

E-step. By forward and backward recursions find $m_{ij}^{(s)}$ and $m_{ij_1 j_2}^{(s)}$ as in (11) and (12), respectively.

S-step. By using the current distances $d_{ij}^{(s)}$ maximize (15): first, obtain the robustness weights defined in (7), then, update estimates:

$$w_{ij}^{(s)} = \left[1 - \left(d_{ij}^{(s)} \right)^2 \right]^2 I_{\{d_{ij} < c_0\}}$$

$$v_{ij}^{(s)} = w_{ij}^{(s)} \left(d_{ij}^{(s)} \right)^2$$

$$\pi_j^{(s+1)} = \frac{\sum_{i=1}^{n} m_{ij}^{(s)}}{n}$$

$$\mu_j^{(s+1)} = \frac{\sum_{i=1}^{n} y_i m_{ij}^{(s)} w_{ij}^{(s)}}{\sum_{i=1}^{n} m_{ij}^{(s)} w_{ij}^{(s)}}$$

$$\Sigma_j^{(s+1)} = \frac{p}{\sum_{i=1}^{n} m_{ij}^{(s)} v_{ij}^{(s)}} \sum_{i=1}^{n} \left(y_i - \mu_j^{(s+1)} \right) \left(y_i - \mu_j^{(s+1)} \right)^T m_{ij}^{(s)} w_{ij}^{(s)}.$$

Anomalous observations can be identified by examining the weights at convergence. A data point will be flagged as an outlier when all the corresponding weights, one for each component of the mixture, are small enough. For a good data point at least one weight is expected to be close to one. Moreover, standard tools based on the inspection and display of the robust Mahalanobis distances can be used to detect outliers and even to verify goodness of fit [e.g., Cerioli and Farcomeni (2011), Cerioli et al. (2013), and references therein].

A final comment concerns the choice of the number of latent states. As a general result, the incomplete data log-likelihood associated with the current values of the parameters is exactly equal to $\sum_{j=1}^{k} \alpha_{nj}$. Here, the likelihood is still derived under the assumed mixture of normal components but it is evaluated at the robust estimates. Note that the likelihood at convergence can be employed in the choice of the number of latent states using for instance the Bayesian Information Criterion (BIC), see McLachlan and Peel (2000) and Bartolucci et al. (2013).

3.1 Properties

We now prove that at each iteration of the ES algorithm an increase of (15) corresponds to an increase in the incomplete data likelihood function, in parallel with the EM algorithm. To this end, it is enough to show that

$$\tilde{Q}\left(\tau^{(s+1)}|\tau^{(s)}\right) - \tilde{Q}\left(\tau^{(s)}|\tau^{(s)}\right) \geq 0, \tag{17}$$

where $\tilde{Q}\left(\tau|\tau^{(s)}\right) = E\left[\ell^c(x;\tau)|y,\tau^{(s)}\right]$, as the fact that the likelihood increases is a direct consequence of (17) (see e.g. Dempster et al. 1977). To assess that relation (17) holds, first note that (13) coincides with (15) for what concerns the first term. Consequently, (17) holds in a standard fashion at the E-step. Then, to see that the inequality (17) also holds after the S-step, it is sufficient to demonstrate that

$$\sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij}^{(s)}\left\{\rho\left(d_{ij}^{(s+1)}\right) + \frac{1}{2}\log|\Sigma_j^{(s+1)}|\right\}$$

$$\leq \sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij}^{(s)}\left\{\rho\left(d_{ij}^{(s)}\right) + \frac{1}{2}\log|\Sigma_j^{(s)}|\right\}, \tag{18}$$

where $\tau^{(s+1)}$ is the maximizer of (15) obtained after the S-step. The proof is given in "Appendix 4".

Since the estimating equations in (16) may have multiple roots corresponding to local maxima, in order to increase the chances of ending up in the global maximum it is recommended to initialize the algorithm from few different starting values. There is no general optimum strategy. A possibility is to initialize the parameters for the manifest distribution as those estimated using robust techniques for *static* finite mixtures, such as the PAM or `tclust` algorithm, and to initialize the hidden transition matrix as being close to diagonal. See for instance Bartolucci et al. (2013) on this. Other initial solutions can be obtained by randomly perturbing the deterministic starting solution and/or the final one obtained from it. In this work we have obtained a total of 20 initial solutions in the real data application. For the numerical studies that follow, we have found that the deterministic solution usually leads to the largest optimum for the likelihood. Let $\theta_0 = (\mu_{0j}, \Sigma_{0j})$, $j = 1, 2, \ldots, p$ be the ES estimates of location and covariance obtained from using the PAM estimates as initial values. Multiple initial solutions are then obtained by randomly perturbing μ_{0j} to get new initial values μ_{1j}. We tried two situations: in the first we ran the ES algorithm for 50 initial values of the form $\theta^* = (\mu_{1j}, \Sigma_{0j})$, in the second we only used μ_{1j} and set the starting estimates for the covariance matrices equal to unit matrices of dimension p. We ran some numerical studies and found that in the first case the deterministic solution always leads to a global maximum, whereas in the second case this happens at an average rate larger than 85 %. Based on these findings, we conclude that in our problem the deterministic solution may be believed to lead to a global optimum and that 10–50 random starts are enough to guarantee it. Furthermore, we have also experienced the ES algorithm to be less dependent of the initial solution than the standard EM when data are contaminated.

It is natural now to wonder about the characteristics of the estimate obtained at convergence of the ES algorithm. What we can intuitively argue is that in absence of contamination the ES and EM algorithm give approximately the same estimates. This is a consequence of consistency of the S-estimates at the normal model, and will be illustrated in simulation below.

3.2 Classification

The classification after the last iteration of the ES algorithm proceeds with the prediction of the most likely hidden state z_i. It would be tempting to set $z_i = j$ in correspondence of the state which is a-posteriori most likely, as in the context of static mixtures. This would maximize the expected number of correct individual states, but would not take into proper account the joint probability of the resulting sequence. For instance, even if a transition between two states is impossible they may still be the most likely at neighboring time points marginally, so producing inconsistent estimates. Hence, one must estimate the hidden states jointly, producing the most likely *sequence* of hidden states. This is performed through the Viterbi algorithm (Viterbi 1967; Juang and Rabiner 1991).

The algorithm can be summarized as follows:

Step 1, Initialization.
$$\xi_1 = \hat{\pi}_1.$$

Step 2, Recursion.

$$\xi_{i+1,j} = \left[\max_{1 \le j_1 \le k} \xi_{ij_1} \hat{\pi}_{j_1,j}\right] f(y_{i+1}; \mu_j, \Sigma_j), \quad 1 \le j \le k, \quad 1 \le i \le n-1,$$

$$\gamma_{i+1,j} = \arg\max_{1 \le j_1 \le k} \xi_{ij_1} \hat{\pi}_{j_1,j}, \quad 1 \le j \le k, \quad 1 \le i \le n-1.$$

Step 3, Termination. After n steps we set the most likely exit state z_n as $\arg\max_{1 \le j \le k} \xi_{nj}$.

Step 4, Backtracking. Finally, the most likely hidden sequence is recursively unraveled by setting the most likely hidden state at time i as $\gamma_{i,z_{i+1}}$, for $i = n-1, \ldots, 1$.

4 Simulations

We performed numerical studies in order to investigate the performance of ES compared to EM and t-based EM. Data have been drawn from conditionally Gaussian random variables, centered on $\mu_{0j} = 5i$, $i = 1, \ldots, k$, $j = 1, \ldots, p$, with unit variance and constant correlation ρ. The initial probabilities are uniform over $\{1, \ldots, k\}$ and the cd-th element of the hidden transition matrix has been set proportional to 0.1 whenever $c \ne d$ and to 1.1 whenever $c = d$. The number of components k is assumed to be known in this section.

A time series of length n has been generated and contamination has been introduced by replacing a proportion ϵ of the n elements by a random draw from a uniform random variable, independently in each dimension, in such a way that outliers are far from the bulk of the data in each component. We used an acceptance-rejection algorithm, according to which only points having squared Mahalanobis distances from the centre of each component of the mixture larger than the $\chi^2_{p;0.975}$ quantile are considered, until reaching the chosen amount $\lfloor n\epsilon \rfloor$ of clear outliers. In the numerical studies as well as in the real data application, the tuning constant c_0 has been determined in order to achieve 50 % BP for each component.

After generating data, we fitted the classical HMM with Gaussian manifest distribution, through the EM procedure (EM in the tables); our robust Gaussian HMM using the ES procedure (ES), and an HMM with multivariate t distributions as manifest ($t - EM$).

For each setting and each procedure we report the Euclidean distance between the estimated and true μ_0 and Π_0, the log condition number of $\Sigma_0^{-1}\hat{\Sigma}$, and the modified Rand index (Hubert and Arabie 1985) for the estimated latent states. The latter is a measure of agreement between the estimated and true labels. The results are based on 1,000 replicates for each setting.

Tables 1 and 2 show the results for $k = 3$, $p = 3, 8$, $\rho = 0.1, 0.5$, $\epsilon = 0, 0.10$ and $n = 50, 300$. It can be seen that the robust method behaves reasonably well when contamination does not occur, whereas it is not affected by outliers in the contaminated scenarios. More in detail, outliers can break down EM based estimates for the mean parameters and scatter matrices whereas the ES leads to reliable estimation of them in general under contamination. The ES based estimation of the mean vectors and variance–covariance matrices is stable across the considered scenarios and consequently appears to be resistant to the presence of outliers. Some loss of efficiency is seen when outliers are not present, but the trade-off seems satisfactory. The use of the t-mixture also leads to a stable estimation of location but is not able to protect the estimation of scatter from outliers: scatter estimates are inflated as well as those from standard EM.

Anomalous values have a mild effect on the non robust estimates of the latent parameters and on the classification. In particular, the same rate of classification, measured by the Rand index, is achieved by all methods when there are not outliers in the sample at hand. Under contamination, the classification rate provided by the ES is always not smaller than that given by the EM and t-EM algorithm.

In order to assess the reliability of the ES under the general assumption of elliptical families (1), a numerical study has also been carried out in which data have been drawn from conditionally Student's t random variables, with the same setting of the previous Monte Carlo analysis. Results in Tables 3 and 4 suggests that the ES algorithm still leads to reliable results.

5 A real data example

Let us consider the stock market data analyzed also in Dias et al. (2008) and Bartolucci and Farcomeni (2010). For the markets of Argentina, Brazil, Canada, Chile, Mexico,

Table 1 HMM example

	$\|\hat{\mu} - \mu_0\|$	$\|\hat{\Pi} - \Pi_0\|$	$-\log cond(\Sigma_0^{-1}\hat{\Sigma})$	Rand inx.
$\epsilon = 0, \rho = 0.1, p = 3$				
EM	0.06	0.16	1.40	1.00
ES	0.07	0.16	1.55	1.00
t-EM	0.06	0.16	1.40	1.00
$\epsilon = 0, \rho = 0.1, p = 8$				
EM	0.06	0.15	3.72	1.00
ES	0.06	0.15	3.74	1.00
t-EM	0.06	0.15	3.76	1.00
$\epsilon = 0, \rho = 0.5, p = 3$				
EM	0.07	0.15	1.55	0.99
ES	0.07	0.15	1.69	0.99
t-EM	0.07	0.15	1.61	0.99
$\epsilon = 0, \rho = 0.5, p = 8$				
EM	0.06	0.15	4.03	1.00
ES	0.06	0.15	4.05	1.00
t-EM	0.06	0.15	4.02	1.00
$\epsilon = 0.1, \rho = 0.1, p = 3$				
EM	1.44	0.16	2.47	0.74
ES	0.06	0.15	1.43	0.81
t-EM	0.13	0.16	2.07	0.77
$\epsilon = 0.1, \rho = 0.1, p = 8$				
EM	2.58	0.15	4.69	0.81
ES	0.06	0.15	3.89	0.81
t-EM	0.19	0.15	4.42	0.80
$\epsilon = 0.1, \rho = 0.5, p = 3$				
EM	1.67	0.17	2.40	0.73
ES	0.06	0.15	1.61	0.80
t-EM	0.12	0.16	2.09	0.75
$\epsilon = 0.1, \rho = 0.5, p = 8$				
EM	2.46	0.15	4.71	0.80
ES	0.06	0.15	4.15	0.81
t-EM	0.12	0.15	4.39	0.79

Accuracy of EM, ES and t-EM estimates based on 1000 Monte Carlo trials for $p = 3, 8$, $\rho = 0.1, 0.5$, $\epsilon = 0, 0.10, n = 50$

Peru and United States, the daily closing price from July 4, 1994, to September 27, 2007, has been drawn from the Datastream database. All series are denominated in US dollars, and for each of them we model the daily rates of returns

$$y_t = \log(P_t/P_{t-1}), \quad t = 1, \ldots, 3454,$$

Table 2 HMM example

	$\|\|\hat{\mu} - \mu_0\|\|$	$\|\|\hat{\Pi} - \Pi_0\|\|$	$-\log cond(\Sigma_0^{-1}\hat{\Sigma})$	Rand inx.
$\epsilon = 0, \rho = 0.1, p = 3$				
EM	0.01	0.13	0.55	1.00
ES	0.01	0.13	0.62	1.00
t-EM	0.01	0.13	0.54	1.00
$\epsilon = 0, \rho = 0.1, p = 8$				
EM	0.01	0.13	1.38	1.00
ES	0.01	0.13	1.42	1.00
t-EM	0.01	0.14	1.37	1.00
$\epsilon = 0, \rho = 0.5, p = 3$				
EM	0.01	0.14	0.64	1.00
ES	0.01	0.14	0.72	1.00
t-EM	0.01	0.14	0.63	1.00
$\epsilon = 0, \rho = 0.5, p = 8$				
EM	0.01	0.14	1.63	1.00
ES	0.01	0.14	1.68	1.00
t-EM	0.01	0.14	1.64	1.00
$\epsilon = 0.1, \rho = 0.1, p = 3$				
EM	0.96	0.14	1.56	0.75
ES	0.01	0.14	0.61	0.81
t-EM	0.02	0.14	1.18	0.79
$\epsilon = 0.1, \rho = 0.1, p = 8$				
EM	1.74	0.14	2.44	0.79
ES	0.01	0.14	1.37	0.81
t-EM	0.03	0.14	2.14	0.79
$\epsilon = 0.1, \rho = 0.5, p = 3$				
EM	1.03	0.14	1.31	0.75
ES	0.01	0.14	0.73	0.80
t-EM	0.01	0.14	1.05	0.78
$\epsilon = 0.1, \rho = 0.5, p = 8$				
EM	2.09	0.14	2.24	0.78
ES	0.01	0.14	1.65	0.80
t-EM	0.02	0.14	1.93	0.79

Accuracy of EM, ES and t-EM estimates based on 1000 Monte Carlo trials for $p = 3, 8$, $\rho = 0.1, 0.5$, $\epsilon = 0, 0.10, n = 300$

where P_t denotes the closing price on day t. Slightly less than 2 % of the observations are not avaiable (due to e.g. to public holidays). We simply set $y_t = 0$ for those, but any other strategy would give substantially equivalent results.

Hence, we have $p = 7$ and repeatedly fit the HMM model for different values of k both with the standard and our robust algorithms. The log-likelihood and BIC obtained by each algorithm are given in Table 5.

Table 3 HMM example

	$||\hat{\mu} - \mu_0||$	$||\hat{\Pi} - \Pi_0||$	$-\log cond(\Sigma_0^{-1}\hat{\Sigma})$	Rand inx.
$v = 5, \rho = 0.1, p = 3$				
EM	0.10	0.15	1.67	0.98
ES	0.07	0.15	1.68	0.97
t-EM	0.09	0.15	1.66	0.97
$v = 5, \rho = 0.1, p = 8$				
EM	0.10	0.15	4.14	0.99
ES	0.06	0.15	3.91	0.99
t-EM	0.09	0.15	4.04	0.99
$v = 5, \rho = 0.5, p = 3$				
EM	0.12	0.15	1.89	0.92
ES	0.07	0.15	1.82	0.93
t-EM	0.10	0.15	1.92	0.90
$v = 5, \rho = 0.5, p = 8$				
EM	0.10	0.15	4.52	0.95
ES	0.06	0.15	4.38	0.94
t-EM	0.09	0.14	4.54	0.92
$v = 10, \rho = 0.1, p = 3$				
EM	0.08	0.15	1.53	0.99
ES	0.07	0.15	1.63	0.99
t-EM	0.07	0.15	1.54	0.99
$v = 10, \rho = 0.1, p = 8$				
EM	0.08	0.15	3.94	1.00
ES	0.06	0.15	3.91	1.00
t-EM	0.07	0.15	3.92	1.00
$v = 10, \rho = 0.5, p = 3$				
EM	0.08	0.15	1.73	0.96
ES	0.07	0.15	1.79	0.97
t-EM	0.08	0.15	1.76	0.95
$v = 10, \rho = 0.5, p = 8$				
EM	0.08	0.15	4.28	0.98
ES	0.06	0.15	4.29	0.97
t-EM	0.08	0.15	4.30	0.97

Accuracy of EM, ES and t-EM estimates based on 1000 Monte Carlo trials for $p = 3, 8, \rho = 0.1, 0.5$, $v = 5, 10, n = 50$ under the Student distribution

We note that the non-robust algorithm leads the BIC criterion to choose $k = 6$. On the other hand, our robust algorithm leads to favor a simpler model, with $k = 5$ latent states. It is worth noting that the sequence of log-likelihood values obtained in the ES appears reasonable as they decrease with increasing number of assumed latent states and are always smaller than the global maximum reached by the proper EM algorithm.

Table 4 HMM example

	$\|\|\hat{\mu} - \mu_0\|\|$	$\|\|\hat{\Pi} - \Pi_0\|\|$	$-\log cond(\Sigma_0^{-1}\hat{\Sigma})$	Rand inx.
$v = 5, \rho = 0.1, p = 3$				
EM	0.02	0.14	0.75	0.98
ES	0.01	0.14	0.67	0.98
t-EM	0.01	0.14	0.62	0.97
$v = 5, \rho = 0.1, p = 8$				
EM	0.02	0.13	1.81	0.99
ES	0.01	0.13	1.56	0.99
t-EM	0.01	0.13	1.47	0.99
$v = 5, \rho = 0.5, p = 3$				
EM	0.02	0.13	0.91	0.94
ES	0.01	0.13	0.76	0.94
t-EM	0.01	0.13	0.75	0.94
$v = 5, \rho = 0.5, p = 8$				
EM	0.02	0.13	2.11	0.95
ES	0.01	0.13	1.83	0.95
t-EM	0.01	0.14	1.79	0.94
$v = 10, \rho = 0.1, p = 3$				
EM	0.01	0.13	0.61	0.99
ES	0.01	0.13	0.64	0.99
t-EM	0.01	0.14	0.58	0.99
$v = 10, \rho = 0.1, p = 8$				
EM	0.01	0.14	1.53	1.00
ES	0.01	0.14	1.49	1.00
t-EM	0.01	0.14	1..44	1.00
$v = 10, \rho = 0.5, p = 3$				
EM	0.01	0.14	0.72	0.97
ES	0.01	0.14	0.75	0.97
t-EM	0.01	0.14	0.69	0.97
$v = 10, \rho = 0.5, p = 8$				
EM	0.01	0.13	1.82	0.98
ES	0.01	0.13	1.76	0.98
t-EM	0.01	0.13	1.73	0.98

Accuracy of EM, ES and t-EM estimates based on 1000 Monte Carlo trials for $p = 3, 8, \rho = 0.1, 0.5,$ $v = 5, 10, n = 300$ under the Student distribution

The robust procedure reveals the occurrence of a larger rate of outliers than the EM and the employment of robust distances highlights some anomalous outcomes otherwise masked, as illustrated in Fig. 1, which displays the distances resulting from the selected HMM$_k$ models by the ES and the EM, respectively. The distances are evaluated for each point with respect to the estimated mean vector and scatter matrix of the group to which the point has been classified.

Table 5 S&P data

k	EM		ES	
	loglik	BIC	*loglik*	BIC
2	−28289.66	57108.89	−29987.12	60503.81
3	−27807.11	56404.51	−28647.10	57724.48
4	−27520.83	56107.51	−27934.12	56935.53
5	−27269.64	55899.88	−27762.59	**56885.79**
6	−27034.91	**55740.01**	−27697.31	57064.81
7	−26898.23	55792.55	−27620.51	57237.11
8	−26794.70	55927.67	−27545.78	57429.83

Log-likelihood and BIC from fitting the HMM_k model to the time series of the daily rates of return of Argentina, Brazil, Canada, Chile, Mexico, Peru and United States referred to the period from July 4, 1994, to September 27, 2007

Bold values indicate the minimum value for the BIC

A very large Mahalanobis distance around the 2000th log-return, corresponding to the crash of the market after September, 11, 2001, can be clearly seen in both panels, but it is much more evident in the first panel corresponding to the ES estimates, as well as other anomalous values.

In conclusion, models chosen with non-robust approaches may be less parsimonious than needed only because of the presence of outliers and not because of a true underlying complex population distribution, even if, as in this example, standard diagnostic tools are able to flag gross outliers.

For the HMM_5 model chosen by the robust procedure, Table 6 gives the estimated mean vectors for the five Gaussian components. For sake of comparison, the MLEs are also given for the case $k = 5$.

It can be noted that there are slight but important differences. The smallest negative classical EM based estimates are more extreme for what concerns the Brazilian and Mexican stock markets, and almost equal for the other markets. On the other hand, the largest positive classical estimates are less extreme for all markets, particularly for the small Peruvian and Chilean markets. The EM estimation leads to a more pessimistic view of the stock markets, likely due to the fact that important decreases in values of the returns happen during very few days, while important increases are more slow after onset. Consequently, predictions based on classical estimates may not be able to catch medium-term rises of the stock market, and would probably over-estimate the importance of short-term falls. As noted by the referee, robust estimation methods should also yield more persistent latent states. This is indeed true. The estimated relative risks of persisting in state j, for $j = 1, \ldots, 5$, range from 1.41 to 1.94 when comparing the robust with the non-robust estimates. Further, after use of the Viterbi algorithm, 58 transitions are estimated robustly, while 216 if parameters are obtained based on the classical EM algorithm.

It is also interesting to note the difference in estimates of the correlation matrix for the first component. The estimates are reported in Table 7. It can be noted that when $k = 5$ the robust estimates of the correlations between the US log-returns and
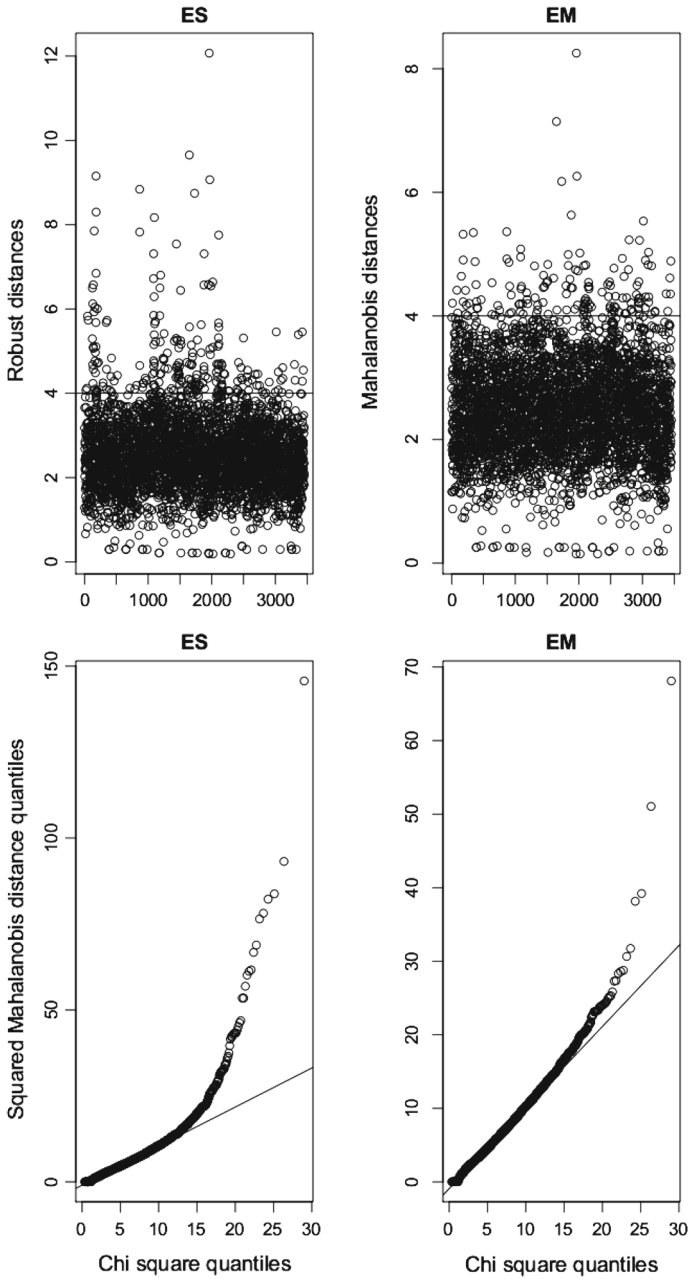
Fig. 1 *S&P* data. *Top* robust (*left*) and classical (*right*) Mahalanobis distances. *Bottom* χ_7^2 Q-Q plot of robust (*left*) and classical (*right*) squared Mahalanobis distances

Table 6 *S&P data*

	AR	BR	CN	CL	MX	PE	US
ES ($k = 5$)							
1 (22%)	−0.12	−0.19	−0.12	−0.18	−0.11	−0.18	−0.03
2 (28%)	0.08	0.10	0.10	0.02	0.05	0.04	0.06
3 (16%)	−0.04	−0.03	−0.03	−0.10	−0.07	−0.10	−0.11
4 (17%)	0.10	0.12	0.09	0.27	0.11	0.21	0.07
5 (17%)	−0.03	0.01	−0.01	0.05	0.03	0.05	−0.05
EM ($k = 5$)							
1 (13%)	−0.09	−0.31	−0.15	−0.18	−0.14	−0.19	−0.01
2 (23%)	−0.05	−0.04	−0.04	−0.08	−0.05	−0.10	−0.08
3 (12%)	−0.05	0.04	−0.01	0.03	0.00	0.04	−0.05
4 (22%)	0.03	0.08	0.05	0.03	0.05	0.07	0.09
5 (30%)	0.07	0.09	0.06	0.11	0.07	0.09	0.02
EM ($k = 6$)							
1 (8%)	−0.10	−0.37	−0.14	−0.22	−0.23	−0.23	0.01
2 (18%)	−0.11	−0.13	−0.13	−0.11	−0.11	−0.16	−0.11
3 (12%)	−0.06	0.00	0.00	0.00	0.01	0.02	−0.07
4 (20%)	0.04	0.02	0.05	−0.09	−0.01	−0.05	0.04
5 (28%)	0.07	0.08	0.06	0.10	0.07	0.08	0.02
6 (14%)	0.05	0.17	0.06	0.18	0.13	0.21	0.09

ES (upper panel) and EM (lower panels) estimated mean parameter vectors for the Gaussian components. In parentheses, the posterior estimates for the percentage of days spent in each state

the same for all other markets are slightly larger (with maybe the exception of MX, which is almost equal). This happens likely because of the fact that the relationship between the other markets and the US market is masked by short-term falls of the smaller markets due to local reasons. After down-weighting of extremal situations, the strong relationship between US and other markets is underlined better. The EM is able to catch a similar behavior only by adding a further component to the mixture.

6 Conclusions

A simple and effective algorithm for robust fitting of dynamic mixtures in the form of HMMs has been proposed. The algorithm has been focused on Gaussian mixtures but it can be generalized to other assumptions. The key point is that the M-step in the EM algorithm can be replaced by robust estimation, as long as the robust approach leads to an increase in the likelihood. In this paper, we have shown this happens when using S-estimators, which have also exhibited good properties both theoretically and empirically in the brief simulation study. Moreover, in our example we have argued that outliers can lead to select an overly complex model when they are not properly taken into account.

We strongly believe that robust methods should be more frequently used in the practice of data analysis. On this point we refer to the discussion in Farcomeni and

Table 7 *S&P* data

	AR	BR	CN	CL	MX	PE	US
ES ($k = 5$)							
AR	1.00	0.68	0.44	0.47	0.62	0.44	0.52
BR	0.68	1.00	0.42	0.50	0.61	0.47	0.53
CN	0.44	0.42	1.00	0.43	0.44	0.36	0.67
CL	0.47	0.50	0.43	1.00	0.50	0.38	0.41
MX	0.62	0.61	0.44	0.50	1.00	0.42	0.53
PE	0.44	0.47	0.36	0.38	0.42	1.00	0.27
US	0.52	0.53	0.67	0.41	0.53	0.27	1.00
EM ($k = 5$)							
AR	1.00	0.51	0.24	0.47	0.46	0.37	0.28
BR	0.51	1.00	0.35	0.68	0.64	0.49	0.38
CN	0.24	0.35	1.00	0.36	0.40	0.33	0.69
CL	0.47	0.68	0.36	1.00	0.58	0.54	0.42
MX	0.46	0.64	0.40	0.58	1.00	0.50	0.48
PE	0.37	0.49	0.33	0.54	0.50	1.00	0.25
US	0.28	0.38	0.69	0.42	0.48	0.25	1.00
EM ($k = 6$)							
AR	1.00	0.77	0.44	0.69	0.69	0.51	0.51
BR	0.77	1.00	0.41	0.67	0.65	0.51	0.43
CN	0.44	0.41	1.00	0.49	0.47	0.42	0.71
CL	0.69	0.67	0.49	1.00	0.57	0.53	0.47
MX	0.69	0.65	0.47	0.57	1.00	0.51	0.51
PE	0.51	0.51	0.42	0.53	0.51	1.00	0.31
US	0.51	0.43	0.71	0.47	0.51	0.31	1.00

ES (upper panel) and EM (lower panels) estimated correlation matrix for the extremal Gaussian component

Ventura (2012), and to the excellent book Heritier et al. (2009). The proposed algorithm may be used in practically any situation in which continuous outcomes are repeatedly measured over time. The focus may be on prediction, estimation of the parameters of the manifest distribution, latent state estimation. In all these cases, contamination may lead to bias; and we therefore believe that robust estimation procedures shall be routinely used when modeling by means of dynamic mixtures. There are many applications of HMM with continuous outcomes in econometrics and finance. Other areas include monitoring of gene expression over time (e.g., Schliep et al. (2003); Farcomeni and Arima (2012)), modeling of ion channels (Michalek et al. 2001), and much more.

There are different possibilities for further work, three of which are worth mentioning. First of all, our approach should be extended to large dimensions. This can be done for instance constraining the covariance matrices of the components (e.g., that they are all equal, that all off-diagonal elements are equal, and so on), but at the price in general of more complex estimation strategies. Secondly, robust model selec-

tion strategies could be better explored and employed and the properties of available strategies after robust estimation could be investigated. Finally, note that there are no robustness issues at the E step, given that it substantially is a maximization with respect to a bounded parameter space. In this sense, m_{ij} and $m_{ij_1j_2}$ can not break down. A more general definition of breakdown can be found in Genton and Lucas (2003), and modification of the E step in this light is also grounds for further work.

Appendix 1: Multivariate M-estimators

Put $V = \Sigma^{-1}$. Recall that, for a matrix A and a vector b

$$\frac{\partial |A|}{\partial A} = |A|A^{-1}$$

$$\frac{\partial b^T A b}{\partial b} = (A + A^T)b$$

$$\frac{\partial b^T A b}{\partial A} = bb^T .$$

By taking the derivatives of (2) with respect to μ and V therefore one obtains

$$\frac{\partial \ell(\theta)}{\partial \mu} = \sum_{i=1}^{n} \frac{\partial \rho(d_i)}{\partial d_i} \frac{\partial d_i}{\partial \mu} = \sum_{i=1}^{n} \frac{\partial \rho(d_i)}{\partial d_i} \frac{V(y_i - \mu)}{d_i}$$

$$\frac{\partial \ell(\theta)}{\partial V} = \sum_{i=1}^{n} \frac{\partial \rho(d_i)}{\partial d_i} \frac{\partial d_i}{\partial V} = \frac{n}{2} V^{-1} - \sum_{i=1}^{n} \frac{\partial \rho(d_i)}{\partial d_i} \frac{(y_i - \mu)(y_i - \mu)^T}{2d_i} .$$

Put $w_i = w(d_i) = \frac{\partial \rho(d_i)}{\partial d_i} \frac{1}{d_i}$, and (3) follows.

Appendix 2: S-estimators

Set $V = \Sigma^{-1}$ and let λ denote a Lagrangian multiplier. Besides the constraint in (4), $(\hat{\mu}_S, \hat{V}_S, \hat{\lambda})$ satisfies

$$\frac{\partial L_n}{\partial \mu} = -V\lambda \sum_{i=1}^{n} (y_i - \mu)w_i = 0 \tag{19}$$

$$\frac{\partial L_n}{\partial V} = -\frac{1}{2} \sum_{i=1}^{n} \left[V^{-1} + \lambda(y_i - \mu)(y_i - \mu_j^T w_i \right] = 0$$

where L_n is the Lagrangian

$$L_n(\mu, \Sigma, \lambda) = \log|\Sigma| - \lambda \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(d_i) - \delta \right\}.$$

On simplification of the second (matrix) equation, we have that

$$\sum_{i=1}^{n} \left[I_p + \lambda V(y_i - \mu)(y_i - \mu)^T w_i \right] = 0$$

and by taking the trace

$$\sum_{i=1}^{n} \left[p + \lambda d_i^2 w_i \right] = 0,$$

from which one obtains the expression for the Lagrangian multiplier

$$\lambda = -\frac{p}{\sum_{i=1}^{n} v_i}. \tag{20}$$

Substitute (20) into (19) and obtain (6) for $\Sigma = V^{-1}$.

7 Appendix 3: S-estimation of location and scatter in HMM

Let us consider the partial derivatives of $L_{n_j^{(s)}}$. Put $V_j = \Sigma_j^{-1}$ and make use of the results in "Appendix 1". Besides the constraints in (14), $(\hat{\mu}_j, \hat{V}_j, \hat{\lambda}_j)$ satisfies

$$\frac{\partial L_{n_j^{(s)}}}{\partial \mu_j} = -V_j \lambda_j \sum_{i=1}^{n} (y_i - \mu_j) w_{ij}^{(s)} m_{ij}^{(s)} = 0$$

$$\frac{\partial L_{n_j^{(s)}}}{\partial V_j} = -\frac{1}{2} \sum_{i=1}^{n} m_{ij}^{(s)} \left[V_j^{-1} + \lambda_j (y_i - \mu_j)(y_i - \mu_j)^T w_{ij}^{(s)} \right] = 0. \tag{21}$$

On simplification of the second (matrix) equation, we have that

$$\sum_{i=1}^{n} m_{ij}^{(s)} \left[I_p + \lambda_j V_j (y_i - \mu_j)(y_i - \mu_j)^T w_{ij}^{(s)} \right] = 0,$$

and by taking the trace

$$\sum_{i=1}^{n} m_{ij}^{(s)} \left[p + \lambda_j d_{ij}^2 w_{ij}^{(s)} \right] = 0,$$

one obtains the expression for the j-th Lagrangian multiplier

$$\lambda_j = -\frac{p \sum_{i=1}^{n} m_{ij}^{(s)}}{\sum_{i=1}^{n} m_{ij}^{(s)} v_{ij}^{(s)}} . \tag{22}$$

If we substitute (22) into (21) then (16) follows for $\Sigma_j = V_j^{-1}$.

8 Appendix 4: Monotonicity

In order to prove (18), first consider that the matrices $|\Sigma_j^{(s+1)}|$ are the solutions to the constrained minimization problems $\mathcal{P}_n^j$, hence at the sth S-step $|\Sigma_j^{(s+1)}| \leq |\Sigma_j^{(s)}|$, where equality holds at convergence. Therefore, it is possible to show that the quantity

$$\sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{(s)} \left[\rho \left(d_{ij}^{(s)} \right) - \rho \left(d_{ij}^{(s+1)} \right) \right], \tag{23}$$

is not smaller than zero.

Let us recall that $\rho(\cdot) = -\log h(\cdot)$, where $h(\cdot)$ is a positive function. This yields that (23) is not smaller than

$$\sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{(s)} w_{ij}^{(s)} \left[d_{ij}^{2,(s)} - d_{ij}^{2,(s+1)} \right], \tag{24}$$

with equality in the Gaussian case for which $\rho(d) = d^2$ [see Maronna et al. (2006, Ch. 9), for further details]. By using the fact that $\mu_j^{(s+1)}$ minimizes $\sum_{i=1}^{n} (y_{ij} - \mu)^T V (y_{ij} - \mu)$ for any positive definite matrix V and therefore $\sum_{i=1}^{n} d_{ij}^{2,(s)} \geq \sum_{i=1}^{n} \tilde{d}_{ij}^{2,(s)}$, with $\tilde{d}_{ij}^{(s)} = d(y_i; \mu_j^{(s+1)}, \Sigma_j^{(s)})$, one can establish that (24) is not smaller than

$$\sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^{(s)} w_{ij}^{(s)} \left[\tilde{d}_{ij}^{2,(s)} - d_{ij}^{2,(s+1)} \right] = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^T \left(\Sigma_j^{(s)} \right)^{-1} u_{ij}$$

$$- \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^T \left(\Sigma_j^{(s+1)} \right)^{-1} u_{ij} , \tag{25}$$

where $u_{ij} = \sqrt{w_{ij}^{(s)} m_{ij}^{(s)}} \left(y_i - \mu_j^{(s+1)} \right)$. Since $\Sigma_j^{(s+1)}$ is the sample covariance matrix of the u_{ij}'s, it minimizes the sum of squared Mahalanobis distances for each component j, then (25) is not smaller than zero and equal to zero when $\Sigma_j^{(s+1)} = \Sigma_j^{(s)}$ and (18) follows.

References

Altman R (2007) Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. J Am Stat Assoc 102:201–210

Bartolucci F, Farcomeni A (2010) A note on the mixture transition distribution and hidden Markov models. J Time Ser Anal 31:132–138

Bartolucci F, Farcomeni A, Pennoni F (2013) Latent Markov models for longitudinal data. Chapman & Hall/CRC Press, Boca Raton, FL

Bashir S, Carter E (2005) High breakdown mixture discriminant analysis. J Multivar Anal 93:102–111

Bashir S, Carter E (2007) Performance of high breakdown mixture discriminant analysis under different biweight functions. Commun Stat Simul Comput 36:177–183

Baum L, Petrie T (1966) Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat 37:1554–1563

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

Bulla J (2011) Hidden Markov models with T components. Increased persistence and other aspects. Quant Financ 11(3):459–475

Cambbell N (1984) Mixture models and atypical values. Math Geol 16:465–477

Cerioli A, Farcomeni A (2011) Error rates for multivariate outlier detection. Comput Stat Data Anal 55:544–553

Cerioli A, Farcomeni A, Riani M (2013) Robust distances for outlier free goodness-of-fit testing. Comput Stat Data Anal 65:29–45

Cuesta-Albertos J, Matran C, Mayo-Iscar A (2008) Robust estimation in the normal mixture model based on robust clustering. J R Stat Soc Ser B 70:779–802

Davies P (1987) Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. Ann Stat 15:1269–1292

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38

Dias J, Vermunt J, Ramos S (2008) Heterogeneous hidden Markov models. In: Proceedings of Compstat 2008

Farcomeni A (2012) Quantile regression for longitudinal data based on latent Markov subject-specific parameters. Stat Comput 22:141–152

Farcomeni A (2013) Snipping for robust k-means clustering under component-wise contamination. Stat Comput 1–13. doi:10.1007/s11222-013-9410-8

Farcomeni A (2014) Robust constrained clustering in presence of entry-wise outliers. Technometrics 56:102–111

Farcomeni A, Arima S (2012) A Bayesian autoregressive three-state hidden Markov model for identifying switching monotonic regimes in Microarray time course data. Stat Appl Genet Mol Biol 11 (article 3)

Farcomeni A, Ventura L (2012) An overview of robust methods in medical research. Stat Methods Med Res 21:111–133

Fraley C, Raftery A (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. J Am Stat Assoc 41:578–588

Gallegos M, Ritter G (2009) Trimmed ML estimation of contaminated mixtures. Sankhya 71:164–220

Genton M, Lucas A (2003) Comprehensive definitions of breakdown points for independent and dependent observations. J R Stat Soc Ser B 65:81–94

Hennig C (2004) Breakdown point for maximum likelihood estimators of location-scale mixtures. Ann Stat 32:1313–1340

Heritier S, Cantoni E, Copt S, Victoria-Feser MP (2009) Robust methods in biostatistics. Wiley, Chichester

Huber P, Ronchetti E (2009) Robust statistics, 2nd edn. Wiley, New York

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Humburg P, Bulger D, Stone G (2008) Parameter estimation for robust HMM analysis of ChIP-chip data. Bioinformatics 9:343

Juang B, Rabiner L (1991) Hidden Markov models for speech recognition. Technometrics 33:251–272

Lopuhaa H (1989) On the relation between S-estimators and M-estimators of multivariate location and covariance. Ann Stat 17:1662–1683

MacDonald IL, Zucchini W (1997) Hidden Markov and other models for discrete-valued time series. Chapman and Hall, London

Markatou M (2000) Mixture models, robustness, and the weighted likelihood metodology. Biometrics 56:483–486

Maronna R, Martin D, Yohai V (2006) Robust statistics: theory and methods. Wiley, Chichester

Maruotti A (2014) Robust fitting of hidden Markov regression models under a longitudinal setting. J Stat Comput Simul 11(8):1728–1747

McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York

Michalek S, Wagner M, Timmer J (2001) Finite sample properties of the maximum likelihood estimator and of likelihood ratio tests in hidden Markov models. Biom J 43:863–879

Neykov N, Filzmoser P, Dimova R, Neytchev P (2007) Robust fitting of mixtures using the trimmed likelihood estimator. Comput Stat Data Anal 52:299–308

Riani M, Cerioli A, Torti F (2014) On consistency factors and efficiency of robust s-estimators. TEST 23(2):356–387

Ronchetti E (1997) Robust inference by influence functions. J Stat Plan Inference 57(1):59–72

Schliep A, Schönhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19:i255–i263

Shah S, Xuan X, DeLeeuw R, Khojasteh M, Lam W, Ng R, Murphy K (2006) Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22:431–439

Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 13:260–269

Yu SZ (2010) Hidden semi-Markov models. Artif Intell 174:215–243