A Note on the Mixture Transition Distribution and Hidden Markov Models

By FRANCESCO BARTOLUCCI

Department of Economics, Finance and Statistics, University of Perugia

06123 Perugia, Italy

bart@stat.unipg.it

AND ALESSIO FARCOMENI

Department of Experimental Medicine, University of Rome "La Sapienza"

00186 Roma, Italy

alessio.farcomeni@uniroma1.it

SUMMARY

We discuss an interpretation of the Mixture Transition Distribution (MTD) for discrete-valued time series which is based on a sequence of independent latent variables which are occasion-specific. We show that, by assuming that this latent process follows a first order Markov Chain, MTD can be generalized in a sensible way. A class of models results which also includes the Hidden Markov Model (HMM). For these models we outline an EM algorithm for the maximum likelihood estimation which exploits recursions developed within the HMM literature.

Some key words: Backward-forward Recursions; Discrete-valued time series; EM-algorithm; State-space models.

1. Introduction

Let X_t, $t = 1, \ldots, T$, be a sequence of discrete random variables having support $\{1, \ldots, k\}$ and let x_t denote a realization of X_t. This sequence is said to follow a Mixture Transition Distribution (MTD) of order l, MTD$_l$ for short, when

$$p(x_t|x_1, \ldots, x_{t-1}) = p(x_t|x_{t-l}, \ldots, x_{t-1}) = \sum_h \lambda_h \pi_{x_{t-h}, x_t}, \quad t > l, \tag{1}$$

where λ_h, $h = 1, \ldots, l$, are *weights* and π_{j_1, j_2}, $j_1, j_2 = 1, \ldots, k$, are *transition probabilities*. The former ones are subjected to the constraints $\lambda_h \geq 0$, $h = 1, \ldots, l$, and $\sum_h \lambda_h = 1$. Assumption (1) implies that the joint probability of the entire sequence of random variables is given by

$$p(x_1, \ldots, x_T) = p(x_1, \ldots, x_l) \prod_{t>l} \sum_h \lambda_h \pi_{x_{t-h}, x_t}, \tag{2}$$

where $p(x_1, \ldots, x_l)$ denotes the joint probability of the first l observations which may be arbitrarily defined. This model was introduced by Raftery (1985a); see also Raftery & Tavaré (1994) who discussed more general constraints on the parameters λ_h. For an exhaustive review on MTD, see Berchtold & Raftery (2002).

With respect to a Markov Chain model of order l, an MTD model with the same order has the advantage of being much more parsimonious because it is based on $(l - 1) + k(k - 1)$ parameters, and this number increases linearly with l. We recall that a Markov Chain of order l is instead based on $k^l(k - 1)$ parameters; this number increases exponentially with l. In both cases we do not consider the parameters used to define the initial probability $p(x_1, \ldots, x_l)$. The MTD model can be generalized to the case of continuous random variables $X_1, \ldots, X_T$ by adopting density transition kernels rather than transition probabilities in equation (1). Also note that these transition probabilities can be lag-specific, so that

$$p(x_t|x_1, \ldots, x_{t-1}) = p(x_t|x_{t-l}, \ldots, x_{t-1}) = \sum_h \lambda_h \pi^{(h)}_{x_{t-h}, x_t}, \quad t > l, \tag{3}$$

and then a further generalization of the MTD model results. This generalized model is indicated by gMTD$_l$ (Raftery, 1985b; Berchtold, 1998). Even in this case, the number of parameters increases linearly with l, since it is given by $(l-1)+k(k-1)l$.

Although the MTD model is generally justified by claiming its parsimony and good adaptation when fitting discrete-valued time series, there are different interpretations and justifications that can be additionally put forward. First of all (1) implies that

$$p(X_t = j_2 | X_{t-l} = \cdots = X_{t-1} = j_1) = \pi_{j_1,j_2}, \quad t > l,$$

so that π_{j_1,j_2} is the probability that the chain moves to state j_2 after it persisted in state j_1 for a period of length l. On the other hand, the weights λ_h can be directly interpreted as the relative influence of each of the previous l occasions on the present.

A more interesting interpretation of the MTD model may be obtained by introducing occasion-specific latent variables Z_t, $t = l+1, \ldots, T$, which are independent and identically distributed and are also independent of $X_1, \ldots, X_T$. Each variable Z_t has a discrete distribution with support $\{1, \ldots, l\}$ and mass probabilities $\lambda_1, \ldots, \lambda_l$. In particular, we can easily show that the MTD model, formulated in its generalized version based on (3), is equivalent to a model based on the assumption

$$p(x_t | x_1, \ldots, x_{t-1}, z_l, \ldots, z_t) = p(x_t | x_{t-l}, \ldots, x_{t-1}, z_t) = \sum_h I(z_t = h) \pi^{(h)}_{x_{t-h}, x_t}, \quad t > l, \tag{4}$$

where $I(\cdot)$ is the indicator function. According to (4), the response variable X_t depends only on the lagged variable X_{t-h}, where the lag h is chosen by a random mechanism which is not directly observable. Then, for $t > l$ and given $Z_t = h$ and $X_{t-h} = j_1$, π_{j_1,j_2} is the conditional probability of $X_t = j_2$, i.e. $p(X_t = j_2 | X_{t-h} = j_1, Z_t = h) = \pi_{j_1,j_2}$. This latent variable interpretation of the MTD model motivates the use of the EM algorithm (Dempster *et al.*, 1977) for parameter estimation; see also Le *et al.* (1996).

The assumption that $Z_{l+1}, \ldots, Z_T$ is a sequence of independent random variables implies that, at each time occasion, the lag on which to rely is independent of the lags previously adopted. In several contexts, this is far to be realistic. Then, we propose a generalization of the MTD and gMTD models based on the assumption that the

sequence $Z_{l+1}, \ldots, Z_T$ follows an hidden Markov Chain. Further generalizations are possible, but are easily seen to lead to models in which the number of parameters can be high and whose fit involves computationally intensive algorithms. The proposed generalization is illustrated in §2, where we show that the resulting model also generalizes the Hidden Markov Model (HMM); see MacDonald & Zucchini (1997). Likelihood inference for the proposed model is discussed in §3.

2. Hidden Markov Extension of the Mixture Transition Distribution

The proposed generalization is based on assumption (4) with Z_t, $t = l+1, \ldots, T$, that follows a homogenous first-order Markov Chain with initial probabilities $\rho_h = p(Z_{l+1} = h)$, $h = 1, \ldots, l$, and transition probabilities $\phi_{h_1,h_2} = p(Z_t = h_2 | Z_{t-1} = h_1)$, $h_1, h_2 = 1, \ldots, l$, for $t > l+1$.

In order to compute the conditional probability $p(x_{l+1}, \ldots, x_T)$, and then $p(x_1, \ldots, x_T)$ as in (2), we can exploit a forward recursion which recalls a well-known recursion in the HMM literature. First of all consider that

$$p(x_{l+1}, z_{l+1} | x_1, \ldots, x_l) = \rho_h \pi^{(h)}_{x_{l+1-h}, x_{l+1}} \tag{5}$$

and that, for any $t > l+1$, we have

$$p(x_{l+1}, \ldots, x_t, z_t | x_1, \ldots, x_l) = \sum_h p(x_{l+1}, \ldots, x_{t-1}, Z_{t-1} = h | x_1, \ldots, x_l) \phi_{h,z_t} \pi^{(z_t)}_{x_{t-z_t}, x_t}. \tag{6}$$

By computing (5) and then (6) for $t = l+2, \ldots, T$, we obtain $p(x_{l+1}, \ldots, x_T, z_T | x_1, \ldots, x_l)$ and consequently the conditional probability of the last $t-l$ observations given the first l observations as

$$p(x_{l+1}, \ldots, x_T | x_1, \ldots, x_l) = \sum_h p(x_{l+1}, \ldots, x_T, z_T = h | x_1, \ldots, x_l).$$

Moreover, we have

$$p(x_{l+1} | x_1, \ldots, x_l) = \sum_h \rho_h \pi^{(h)}_{x_{l+1-h}, x_{l+1}}$$

which is the same as (3), whereas, for $t > l + 1$, the above assumptions imply that

$$p(x_t|x_1, \ldots, x_{t-1}) = \sum_h \lambda_h^{(t)}(x_1, \ldots, x_{t-1}) \pi^{(h)}_{x_{t-h}, x_t}, \tag{7}$$

with $\lambda_h^{(t)}(x_1, \ldots, x_{t-1})$ denoting the conditional probability of $Z_t = h$ given all the previous observations, which may be computed as

$$\lambda_h^{(t)}(x_1, \ldots, x_{t-1}) = \frac{\sum_m p(x_{l+1}, \ldots, x_{t-1}, Z_{t-1} = m|x_1, \ldots, x_l)\phi_{m,h}}{\sum_m p(x_{l+1}, \ldots, x_{t-1}, Z_{t-1} = m|x_1, \ldots, x_l)}.$$

Clearly, expression (7) is a generalization of (3) in which the mixing weights are time-varying and depend on the previous observations. The way in which each weight varies according to t and the previous observations depends on the latent transition probabilities. It is also clear that the above model generalizes not only the MTD and gMTD models, but also the HMM; then we will indicate it by HM-gMTD$_l$, where l is the lag order.

It is worth noting that the HM-gMTD$_l$ model specializes into the gMTD$_l$ model when $\phi_{h_1,h_2} = \rho_{h_2}$, $h_1, h_2 = 1, \ldots, l$, and then the latent variables Z_t are independent of each other and have the same distribution with mass probabilities $\rho_1, \ldots, \rho_l$. On the other hand, the HM-gMTD$_l$ model specializes into the HMM when $\pi^{(h)}_{j_1,j_2} = \pi^{(h)}_{j_2}$, $j_1, j_2 = 1, \ldots, k$, $h = 1, \ldots, l$, so that the distribution of each observation does not depend on the previous observations, but only on the corresponding latent variable. Note that when such an assumption is made on the manifest probabilities, the latent process can be considered to start at $t = 1$. Other different models can arise according to the constraints which are put on the parameters of the HM-gMTD$_l$ model.

The above points are summarized in Table 1, where we also indicate how to compute the number of parameters of the HM-gMTD$_l$ model and the most important submodels; see also Table 2 for numerical examples about the application of these rules.

It can be appreciated that the HM-gMTD class is flexible enough to contain many models commonly used for discrete-value time series. The HM-MTD specialization provides a generalization of MTD which is still quite parsimonious while providing interesting insights into persistency phenomena of the series. Also note that the number

Model	Nested models	Constraints	#parameters
HM-gMTD$_l$	HM-MTD$_l$, gMTD$_l$, HM, MTD$_l$	-	$l^2-1+kl(k-1)$
HM-MTD$_l$	MTD$_l$	$\pi^{(h)}_{j_1,j_2}=\pi_{j_1,j_2}$	$l^2-1+k(k-1)$
gMTD$_l$	MTD$_l$	$\phi_{h_1,h_2}=\rho_{h_2}$	$l-1+kl(k-1)$
HMM		$\pi^{(h)}_{j_1,j_2}=\pi^{(h)}_{j_2}$	$l^2-1+l(k-1)$
MTD$_l$		$\phi_{h_1,h_2}=\rho_{h_2}, \pi^{(h)}_{j_1,j_2}=\pi_{j_1,j_2}$	$l-1+k(k-1)$

Table 1: *List of models nested into the HM-gMTD model with the corresponding number of parameters.*

	l ($k=2$)				
Model	1	2	3	4	5
HM-gMTD$_l$	2	7	14	23	34
HM-MTD$_l$	2	5	10	17	26
gMTD$_l$	2	5	8	11	14
HMM	1	5	11	19	29
MTD$_l$	2	3	4	5	6
Markov Chain	2	4	8	16	32
	l ($k=3$)				
Model	1	2	3	4	5
HM-gMTD$_l$	6	15	26	39	54
HM-MTD$_l$	6	9	14	21	30
gMTD$_l$	6	13	20	27	34
HMM	2	7	14	23	34
MTD$_l$	6	7	8	9	10
Markov Chain	6	18	54	162	486
	l ($k=4$)				
Model	1	2	3	4	5
HM-gMTD$_l$	12	27	44	63	84
HM-MTD$_l$	12	15	20	27	36
gMTD$_l$	12	25	38	51	54
HMM	3	9	17	27	39
MTD$_l$	12	13	14	15	16
Markov Chain	12	48	192	768	3072

Table 2: *Comparison between the models listed in Table 1 and a Markov Chian model of order l in terms of number of parameters.*

of parameters of the HM-gMTD$_l$ model is $l^2-1+kl(k-1)$ which increases quadratically, rather than linearly, in l. In any case, this number is usually much smaller than that of an ordinary Markov Chain model with the same lag, especially when the manifest transition probabilities $\pi^{(h)}_{j_1,j_2}$ are assumed to be constant in h, and then the HM-MTD$_l$

model results. A further reduction in the number of parameters can be achieved by assuming a specific structure for the latent transition matrix with elements ϕ_{h_1,h_2}. For instance, we can assume this matrix to be symmetric, tridiagonal, or even with off-diagonal elements equal to each other. For an illustration of constraints on this type in a similar context see Bartolucci (2006).

Raftery (1985a) showed that the MTD model has the same equilibrium distribution as the first-order Markov Chain with the same transition probabilities, no matter the MTD order. In parallel with that result, we prove below that for any finite l, the stationary distribution of the HM-MTD$_l$ model coincides with that of the corresponding first order Markov Chain with transition probabilities π_{j_1,j_2}, $j_1, j_2 = 1, \ldots, k$. It is then straightforward to see that any HM-gMTD$_l$ model has stationary distribution given by a suitable mixture of the stationary distributions associated to each matrix of transition probabilities with elements $\pi^{(h)}_{j_1,j_2}$, $h = 1, \ldots, l$.

Theorem 1. *Let $X_1, X_2, \ldots$ be distributed according to the HM-MTD$_l$ model, with l finite, and let $\pi_1, \ldots, \pi_k$ denote the probability masses of the stationary distribution associated to the transition probabilities π_{j_1,j_2}, $j_1, j_2 = 1, \ldots, k$. Then, as t goes to infinity, $p(X_t = j) \rightarrow \pi_j$, $j = 1, \ldots, k$.*

Proof. First of all consider that

$$p(X_t = j) = \sum_h p(X_t = j | Z_t = h) p(Z_t = h), \quad t > l.$$

For any h and j, $p(X_t = j | Z_t = h) \rightarrow \pi_j$ as t goes to infinity. Then the result obviously holds because $\sum_h p(Z_t = h) = 1$. □

3. Likelihood Inference

In the following, we outline an EM algorithm (Dempster *et al.*, 1977) which may be used for the maximum likelihood estimation of the parameters of the HM-gMTD$_l$ model

and then of each nested model listed in Table 1. The algorithm is formulated for the case in which we observe a single time series $x_1, \ldots, x_T$, but it can be easily adapted to the case of panel data in which we observe short sequences of observations for a sample of n statistical units.

When we observe a single time series, the *log-likelihood* to be maximized is

$$\ell(\theta) = \log p(x_{l+1}, \ldots, x_T | x_1, \ldots, x_l) + \log p(x_1, \ldots, x_l),$$

where θ is the vector of all model parameters and the first component at rhs may be computed by the recursion illustrated in §2. The second component at rhs, i.e. $\log p(x_1, \ldots, x_l)$, is not of direct interest and it is treated as a constant term.

The EM algorithm is based on the maximization of a suitable expectation of the log-likelihood of the *complete data* which are represented by $z_{l+1}, \ldots, z_T$ further to the observations $x_1, \ldots, x_T$. This log-likelihood has expression

$$\begin{aligned} \ell^*(\theta) &= \log p(x_{l+1}, \ldots, x_T, z_{l+1}, \ldots, z_T | x_1, \ldots, x_l) = \\ &= \sum_{t>l} \sum_h d_{t,h} \log(\pi^{(h)}_{x_{t-h}, x_t}) + \sum_h d_{l+1,h} \log(\rho_h) + \sum_{h_1} \sum_{h_2} \log(\phi_{h_1,h_2}) \sum_{t>l+1} d_{t-1,h_1} d_{t,h_2}, \end{aligned}$$

where $d_{t,h} = I(z_t = h)$ is a dummy variable equal to 1 if the latent process is in state h at occasion t and to 0 otherwise. Consequently, $\sum_{t>l+1} d_{t-1,h_1} d_{t,h_2}$ is equal to the number of transitions from state h_1 to state h_2.

At the E-step, the algorithm computes the conditional expected value of each $d_{t,h}$ and $d_{t-1,h_1} d_{t,h_2}$ given the observed data. Note that

$$\begin{aligned} E(d_{t,h} | x_1, \ldots, x_T) &= p(Z_t = h | x_1, \ldots, x_T), \\ E(d_{t-1,h_1} d_{t,h_2} | x_1, \ldots, x_T) &= p(Z_{t-1} = h_1, Z_t = h_2 | x_1, \ldots, x_T); \end{aligned}$$

these *posterior probabilities* may be obtained by recursions taken from the HMM literature which we describe below. See MacDonald & Zucchini (1997) for a general description and Bartolucci (2006) for an efficient implementation based on the matrix notation. Also see Bartolucci & Besag (2002) for alternative recursions.

For $t > l$, let

$$\begin{aligned}\alpha_t(h) &= p(x_{l+1}, \ldots, x_t, Z_t = h | x_1, \ldots, x_l),\\ \beta_t(h) &= p(x_{t+1}, \ldots, x_T | x_1, \ldots, x_t, Z_t = h),\end{aligned}$$

so that $p(x_{l+1}, \ldots, x_T | x_1, \ldots, x_l) = \sum_h \alpha_T(h)$. The first quantity corresponds to (5) when $t = l + 1$ and, because of (6), may be recursively computed as

$$\alpha_t(h) = \sum_m \alpha_{t-1}(m)\phi_{m,h}\pi^{(h)}_{x_{t-h},x_t}$$

for $t > l + 1$. Similarly, $\beta_t(h)$ may be computed by the backward recursion

$$\beta_t(h) = \sum_m \beta_{t+1}(m)\phi_{h,m}\pi^{(m)}_{x_{t+1-m},x_{t+1}},$$

initialized with $\beta_T(h) = 1$ for $h = 1, \ldots, l$. It is straightforward to see that

$$p(Z_t = h | x_1, \ldots, x_T) = \frac{\alpha_t(h)\beta_t(h)}{p(x_{l+1}, \ldots, x_T | x_1, \ldots, x_l)}, \quad t > l,$$

and

$$p(Z_{t-1} = h_1, Z_t = h_2 | x_1, \ldots, x_T) = \frac{\alpha_{t-1}(h_1)\phi_{h_1,h_2}\pi^{(h_2)}_{x_{t-h_2},x_t}\beta_t(h_2)}{p(x_{l+1}, \ldots, x_T | x_1, \ldots, x_l)}.$$

At the M-step, the algorithm updates the parameter estimates by maximizing the expected value of $\ell^*(\theta)$, obtained by substituting to each $d_{t,h}$ and $d_{t-1,h_1}d_{t,h_2}$ the expected values computed as above. Under the largest model, HM-gMTD$_l$, explicit solutions are available, i.e.

$$\pi^{(h)}_{j_1,j_2} = \frac{\sum_{t>l} p(Z_t = h | x_1, \ldots, x_T) I(x_{t-h} = j_1, x_t = j_2)}{\sum_{t>l} p(Z_t = h | x_1, \ldots, x_T) I(x_{t-h} = j_1)}, \quad h = 1, \ldots, l,\ j_1, j_2 = 1, \ldots, k, \tag{8}$$

for what concerns the manifest transition probabilities,

$$\rho_h = p(Z_{l+1} = h | x_1, \ldots, x_T), \quad h = 1, \ldots, l,$$

for the initial probabilities of the latent process, and

$$\phi_{h_1,h_2} = \frac{\sum_{t>l+1} p(Z_{t-1} = h_1, Z_t = h_2 | x_1 \ldots, x_T)}{\sum_{t>l+1} p(Z_{t-1} = h_1 | x_1 \ldots, x_T)}, \quad h_1, h_2 = 1, \ldots, l,$$

for its transition probabilities.

Note that in case the HM-MTD$_l$ model is assumed, the manifest transition probabilities are updated as

$$\pi_{j_1,j_2} = \frac{\sum_h \sum_{t>l} p(Z_t = h|x_1, \ldots, x_T) I(x_{t-h} = j_1, x_t = j_2)}{\sum_h \sum_{t>l} p(Z_t = h|x_1, \ldots, x_T) I(x_{t-h} = j_1)}, \quad j_1, j_2 = 1, \ldots, k,$$

instead of by (8). Moreover, when the MTD$_l$ model is assumed, the initial probabilities of the latent process are updated as

$$\rho_h = \frac{\sum_{t>l} p(Z_t = h|x_1, \ldots, x_T)}{T - l}, \quad h = 1, \ldots, l,$$

and we let $\phi_{h_1,h_2} = \rho_{h_2}$, $h_1, h_2 = 1, \ldots, l$, since in this case the latent transition matrix is assumed to have each row equal to $\rho_1, \ldots, \rho_l$. In case the HMM is assumed, the algorithm reduces to a standard EM algorithm to fit this model. Finally, under more elaborated constraints on the latent transition matrix, e.g. this matrix is assumed tridiagonal, updating the estimates of its elements requires more sophisticated rules which may be taken from Bartolucci (2006).

The EM algorithm described above is guaranteed to lead to a local maximum of the likelihood. To increase the chance of catching the global maximum, common strategies involve multistart and/or initialization from opportune starting values (for instance obtained from maximum likelihood estimation of models nested in the assumed one).

Once the maximum likelihood estimate has been obtained, we can predict the most likely sequence of latent states through a Viterbi algorithm (Viterbi, 1967) along the same lines as Bartolucci & Farcomeni (2008). We also refer to Bartolucci and Farcomeni (2008) for a method to compute the standard errors for the parameter estimates which is based on the numerical derivative of the score vector; the latter is directly obtained from the EM algorithm. These standard errors may be used to construct confidence intervals and testing statistical hypotheses on the parameters. A more general way to test such hypotheses is by the likelihood ratio statistic. Note, however, that the null asymptotic distribution of this statistic is not ensured to be a standard chi-squared distribution when the hypothesis of interest is that certain elements of the latent transition

matrix are equal to 0. This happens, for instance, when we assume that this matrix is tridiagonal. In this case, the asymptotic distribution is of chi-bar-squared type (Bartolucci, 2006), i.e. a mixture of chi-squared distributions with suitable weights; for a general description of this distribution see Shapiro (1988).

Finally, a fundamental point concerns model choice with respect to both the order l of the lag and possible constraints on the parameters; see Table 1. In the MTD literature, the Bayesian Information Criterion of Schwarz (1978) seems to be preferred among the available selection criteria. This criterion is based on the minimization of the index $BIC = -2\ell(\hat{\theta}) + g \log(T - l)$, where $\hat{\theta}$ is the vector of parameter estimates obtained at convergence of the EM algorithm and g is the number of non-redundant parameters. Modifications of the penalization terms are required with panel data in order to take into account the sample size also. In the HMM literature, BIC is known to perform well in choosing the order of the model even if its theoretical properties are not so clear; see Celeux & Durand (2006) and the references therein. These reasons lead us to suggest BIC as an adequate selection criterion for the proposed model, as an alternative to other criteria such as the Akaike Information Criterion (Akaike, 1973).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International symposium on information theory*. Petrov, B. N. and Csaki F. (eds), 267–81, Budapest: Akademiai Kiado.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, Series B* **68**, 155–78.

Bartolucci, F.& Besag, J. (2002). A recursive algorithm for Markov random fields. *Biometrika* **89**, 724–30.

Bartolucci, F. & Farcomeni, A. (2008). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the*

American Statistical Association, in press, available at
`http://www.stat.unipg.it/bartolucci/marg_long28.pdf`.

Berchtold, A. (1998) *Chaînes de Markov et Modèles de Transition: Applications aux Sciences Sociales.* Hermes, Paris.

Berchtold, A. & Raftery, A.E. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17:328-356.

Celeux, G. & Durand, J-B. (2006). Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics* **23**, 541–64.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Le, N.D., Martin, D. & Raftery, A.E. (1996). Modeling flat streches, bursts, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, 91:1504-1515.

MacDonald, I.L. & Zucchini, W. (1997). *Hidden Markov and other models for discrete valued time series.* London: Chapman and Hall.

Raftery, A.E. (1985a). A model for high-order Markov chains. *Journal of the Royal Statistical Society, Series B* **47**, 528–39.

Raftery, A.E. (1985b). A new model for discrete-valued time series: Autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni* **3**, 149-162.

Raftery, A.E. & Tavaré, S. (1994). Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model. *Applied Statistics* **43**, 179-199.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis, *International Statistical Review* **56**, 49–62.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics* **6**, 461-464.

Silvapulle, M.J. & Sen, P.K. (2004). *Constrained Statistical Inference.* Wiley, New York.

Viterbi, A.J. (1967), Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *Transactions on Information Theory* **13**, 260–269.