Hidden Markov Model Regression

Moshe Fridman *
Institute for Mathematics and its Applications
University of Minnesota
514 Vincent Hall
206 Church Street S.E.
Minneapolis, Minnesota 55455

December 30, 1993

Abstract

Hidden Markov Model Regression (HMMR) is an extension of the Hidden Markov Model (HMM) to regression analysis. We assume that the parameters of the regression model are determined by the outcome of a finite-state Markov chain and that the error terms are conditionally independent normally distributed with mean zero and state dependent variance. The theory of HMM regression is quite new, but some of its development calls on the natural extension of the work by Baum and Petrie. We consider the problem of maximum likelihood estimation of the HMMR parameters and develop analogs for the methods used in HMM's for our regression case. Simulation studies indicate consistency and asymptotic normality of the suggested estimates.

Key words: Hidden Markov model, Forward-Backward procedure, Baum-Welch algorithm, Switching regression model.

AMS 1991 subject classifications. Primary 60J10; Secondary 62M05.

* Research partially supported by ARO Grant DAAL03-91-G-0110

1 Introduction

Hidden Markov models offer a natural tool for dealing with one of the fundamental problems in stochastic modeling: many naturally generated stochastic processes exhibit temporal heterogeneity that is driven by an underlying (but unobservable) change in the signal generating system.

The source of strength of the HMM seems to be due to its ability to acknowledge the relationships between changing regimes when on a short term basis one could adequately model the observed data by an homogeneous process. A second source of strength of HMM's is their exceptional ability to incorporate structural features of the phenomena under study into the structural features of the model. Often the topology of the HMM (the number of states, the transition matrix structure, and observed sequence distributions) is designed to incorporate as many features of the observed process as the underlying science can justify. Although such modeling is not *a priori* effective, history has done much to support the practice. The HMM has been applied with telling success in a variety of scientific contexts with increasing level of model complexity that follows the increase in available computational power.

Early contributions to identifiability and statistical inference problems for HMM's were given in papers by Blackwell and Koopmans (1957), Gilbert (1959), and Baum and Petrie (1966). However, early work was restricted to the case where the observed sequence is a discrete process. The pioneering papers of Baum and Eagon (1967), Petrie (1969), and Baum, Petrie, Soules and Weiss (1970) seem to be the first to introduce a procedure for the maximum likelihood estimation of the HMM parameters for the general case where the observed sequence is a sequence of random variables with log-concave densities. Recent work has extended the theory to multivariate observations and more general densities. In particular, work by Liporace (1982), Juang (1985), and Juang, Levinson and Sondhi (1986), offers generalizations for multivariate stochastic observations of Markov chains with densities that are mixtures of log-concave and elliptically symmetric densities. Extensions of HMM's to ARMA models subject to Markovian changes in regime are studied Hamilton (1989).

A variety of linear models for evolving processes that exhibit discontinuous changes at certain undetermined points in time have been discussed in the statistical literature. In the regression context, the literature refers to such models as *switching regression* models. Switching regressions in which the switching process is modeled via a random walk have been extensively studied by Quandt (1972), Quandt and Ramsey (1978), and Kiefer (1978). In this paper we propose an extension of HMM's to the regression model that results in a switching regression model that is subject to Markovian changes in regime. We no longer have independence in the dependent variable process and as a consequence the analysis of such models is considerably harder. In particular, the analysis of the likelihood function gives rise to analytical and computational difficulties that are not present in the random walk switching case. We provide a practical and simple to implement algorithm for the numerical computation of the maximum likelihood estimate (MLE) for the parameter of the model. The derivation of estimates draws from earlier developments for the HMM. In the following we present some of the basic methodology for HMM's in Section 2, formulate HMM regression and derive the MLE for its parameter in Section 3, and present simulation results in Section 4. In Section 5 we summarize and propose some directions for further research.

2 HMM

We begin with a formal definition of HMM's. We call $\{Y_t\}$ the *observed sequence* of an HMM process if there exists a Markov chain $\{Q_t\}$ on the state space $\mathcal{S} = \{S_1, \ldots, S_N\}$ and cumulative distribution functions $F_1, \ldots, F_N$ such that

$$\begin{aligned}&P(Y_1 \leq c_1, \ldots, Y_T \leq c_T | Q_t = S_i) = P(Y_1 \leq c_1, \ldots, Y_{t-1} \leq c_{t-1}, Q_t = S_i) \cdot \\ &F_i(c_t) \cdot P(Y_{t+1} \leq c_{t+1}, \ldots, Y_T \leq c_T | Q_t = S_i),\end{aligned}$$

for any $1 \leq t \leq T$, $1 \leq i \leq N$, and constants $c_1, \ldots, c_T$. A central hypothesis implied by this definition is that given Q_t the variable Y_t is independent of $\{Y_s, Q_s : s \neq t\}$. The process derives its name from the fact that the Markov chain $\{Q_t\}$ is *unobserved or hidden.* We denote the transition probability matrix by A and the initial distribution by Π.

We will be concerned with continuous value models with parametric families of absolutely continuous distributions of the form $\{F_{\theta_i}\}_{i=1}^N$, i.e., the set of distribution functions for the observed sequence constitute a parametric family with parameter θ in Θ, where Θ is a subset of the n dimensional Euclidean space. For convenience we use the compact notation $\lambda = (A(\lambda), \Theta(\lambda), \Pi(\lambda))$ to indicate a completely specified model. We also will simplify the notation by suppressing the θ from the subscript for the distribution and density functions.

The joint probability of Y and Q under the model λ is given by

$$P_\lambda(Y, Q) = \pi_{q_1} f_{q_1}(y_1) \cdot \prod_{t=2}^{T} a_{q_{t-1}q_t}\, f_{q_t}(y_t). \tag{1}$$

The likelihood function $L(\cdot)$ is given by

$$L(\lambda) = \sum_q P_\lambda(Y, Q) = \sum_q \pi_{q_1} f_{q_1}(y_1) \prod_{t=2}^{T} a_{q_{t-1}q_t}\, f_{q_t}(y_t), \tag{2}$$

where the summation is over all q in $\mathcal{Q} = \{S_1, S_2, \ldots, S_N\}^T$ the product set of all feasible paths through the states space.

2.1 The Evaluation Problem

The evaluation problem is a question in computational efficiency. A naive evaluation of $P_\lambda(Y)$ from equation (2) is computationally infeasible since it involves N^T operations. Instead, we can invoke a simple but powerful procedure to evaluate (2) that avoids having to perform any computation that is exponential in the observation sequence length T. This procedure is called the *forward–backward* procedure, and the discussion that we give here is based on the presentation in Baum (1972). We will make use of the forward–backward procedure in the evaluation of the likelihood of the HMM regression process.

We define the *forward variables*

$$\alpha_t(i) \overset{def}{=} P_\lambda(Y_1 = y_1, Y_2 = y_2, \ldots Y_t = y_t, Q_t = S_i). \tag{3}$$

and the *backward variables*

$$\beta_t(i) \overset{def}{=} P_\lambda(Y_{t+1} = y_{t+1}, Y_{t+2} = y_{t+2}, \ldots Y_T = y_T \mid Q_t = S_i). \tag{4}$$

The key observation that makes the forward and backward variables so useful is that they can be calculated recursively from $\alpha_{t-1}(j)$ and $\beta_{t+1}(j)$, namely

$$\begin{aligned} \alpha_t(i) &= \left[\sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji}\right] f_i(y_t), \quad \text{with} \quad \alpha_1(i) = \pi_i f_i(y_1), \\ \beta_t(i) &= \sum_{j=1}^{N} a_{ij}\, f_j(y_{t+1})\, \beta_{t+1}(j), \quad \text{with} \quad \beta_T(i) = 1\ \forall i. \end{aligned}$$

Using only the forward variables we have from the definition (3) that the *forward only evaluation* formula for $P_\lambda(Y)$ is

$$P_\lambda(Y) = \sum_{i=1}^{N} \alpha_T(i). \tag{5}$$

However, the Forward-Backward procedure provides an effective way of calculating the broader class of probabilities $P_\lambda(Y = y, Q_t = S_i)$, that we shall use in the parameter and state estimation procedures. Such probabilities can be calculated using the *partitioned path probability* formula $P_\lambda(Y = y, Q_t = S_i) = \alpha_t(i)\beta_t(i)$. Note that summing over all possible states $1 \leq i \leq N$ at time t, we obtain the *forward-backward evaluation* formula

$$P_\lambda(Y) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i), \quad \text{for all } 1 \leq t \leq T. \tag{6}$$

Equation (5) is a special case of (6) with $t = T$.

It is easy to verify that the computation of the forward and backward variables require only an order of N^2T operations and an order of $N \times T$ storage units, making the computations feasible. One remaining numerical difficulty is that calculations tend to exceed the standard computer precision range exponentially fast, causing an underflow or overflow condition. In practice, incorporation of a scaling procedure is required to carry out the forward-backward calculations. One such scaling procedure is described in Levinson, Rabiner and Sondhi (1983).

3 HMM Regression

In this section we introduce an extension of HMM's to regression analysis which we call the *hidden Markov model regression* (HMMR). As suggested by its name, the HMMR is a model that relates the *dependent variable* Y to a set of *independent variables* $(X_1, \ldots, X_p)$ through a number of regression planes with distinct regression parameter values. The regression planes describe the relationship between the dependent and independent variables under the various states of the unobserved Markov chain Q.

3.1 HMMR Definition

Let $\{Q\}_{t=1}^{T}$ denote an homogeneous ergodic Markov chain with transition matrix A and state space $\{S_1, \ldots, S_N\}$. We define the HMMR as

$$Y_t = X_t^T \beta_{Q_t} + \sigma_{Q_t} \epsilon_t, \quad 1 \leq t \leq T, \tag{7}$$

where $\beta_{Q_t} = \beta_i$ and $\sigma_{Q_t} = \sigma_i$ if $Q_t = S_i$ and the error terms ϵ_t are independent and identically distributed as $N(0, 1)$. The vector of covariates $X_t = (1, x_{t1}, \ldots, x_{tp})^T$ is an $(p+1) \times 1$ vector of known constants at time t and $\theta_i = (\beta_i^T, \sigma_i^2)$, where $\beta_i^T = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{pi})$, are the regression parameters associates with state S_i. By the distribution assumption for the error terms, the sequence of dependent variables $\{Y_t\}_{t=1}^T$ are conditionally independent and normally distributed with mean $X_t^T \beta_i$ and variance σ_i^2, given $Q_t = S_i$. As usual, we assume that the $T \times (p+1)$ covariates matrix X is of full column rank.

3.2 The Baum-Welch Approach

The *Baum-Welch* (BW) algorithm, as presented in Baum *et al.* (1970), is a numerical procedure for the local maximization of the likelihood of an HMM. We adopt the Baum-Welch approach to provide a procedure for maximum likelihood estimation of the parameter of an HMMR. The approach is based on the Kullback-Leibler inequality that says that for any two distributions P_{θ_1} and P_{θ_2} we have

$$\sum_x P_{\theta_1}(x) \ln \frac{P_{\theta_1}(x)}{P_{\theta_2}(x)} \geq 0.$$

This approach provides an elegant iterative uphill stepping algorithm that requires only first derivatives of the function $\ln P_\lambda(Y, Q)$ (differentiation of $P_\lambda(Y)$ is much more involved), and allows for the easy incorporation of the forward-backward procedure, making the computations feasible. The BW algorithm is equivalent to what latter became known in its general form as the EM algorithm.

In general terms, the BW method utilizes an auxiliary function

$$Q(\lambda, \lambda^{'}) = \sum_q P_\lambda(Y, Q) \ln P_{\lambda^{'}}(Y, Q),$$

to define a transformation τ from the parameter space Λ into itself given by

$$\tau : \lambda \rightarrow \hat{\lambda} \overset{def}{=} \arg\max_{\lambda^{'} \in \Lambda} Q(\lambda, \lambda^{'}),$$

that under general conditions on the distribution family $\{F_\theta\}$ produces a sequence of increasing likelihood values when successively applied to an initial parameter $\lambda_0 \in \Lambda$.

3.3 Maximum Likelihood Estimation

We next apply the BW approach for the HMMR case and show that the sequence $L(\tau^n(\lambda))$ is monotone nondecreasing. Furthermore, we provide a workable form for this transformation.

The likelihood function for an HMMR sequence $(Y_1, \ldots, Y_T)$ is

$$L(\lambda) \;=\; \sum_q \prod_{t=1}^{T} a_{q_{t-1},q_t}(2\pi\sigma^2_{q_t})^{-1/2}\exp(-\frac{(Y_t - X_t^T\beta_{q_t})^2}{2\sigma^2_{q_t}}), \qquad (8)$$

where $a_{q_0,q_1} = \pi_{q_1}$. The auxiliary function $Q(\lambda, \lambda^{'})$ can be written as a sum of three functions, each operating on a different set of primed parameters,

$$Q(\lambda,\lambda^{'}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\nu_{ij}\ln a^{'}_{ij} + \sum_{i=1}^{N}\zeta_i \ln \pi^{'}_i + \qquad (9)$$

$$\sum_{j=1}^{N}\sum_{q}\sum_{t_j} P_\lambda(y,q)\ln\left[(2\pi\sigma_j^{2'})^{-1/2}\exp\{-(Y_t - X_t^T\beta^{'}_j)^2/(2\sigma_j^{2'})\}\right].$$

Here $\nu_{ij} = \sum_q(\sum_{t=2}^{T} 1_{(q_{t-1}=i,q_t=j)})P_\lambda(Y,Q)$, $\zeta_i = \sum_q(1_{(q_1=i)})P_\lambda(Y,Q)$, and $t_j = \{t : q_t = S_j, 1 \le t \le T\}$. Hence, it will be useful to analyze these three functions separately. We begin by stating as a lemma a well known result on constrained optimization which one can prove by the method of Lagrange multipliers.

Lemma 3.1 *If $c_i > 0, x_i > 0$ for $1 \le i \le N$ then subject to the constraint $\sum_i x_i = 1$, the function $F(\boldsymbol{x}) = \sum_i c_i \ln x_i$ attains its unique global maximum when $x_i = c_i / \sum_k c_k$.*

□ The main result is summarized in the following theorems. Proofs are given in the Appendix.

Theorem 3.1 *Let $\tau : \lambda \to \hat{\lambda} = \arg\max_{\lambda^{'}\in\Lambda} Q(\lambda,\lambda^{'})$. Then, we have*

1. *$L(\tau(\lambda)) \ge L(\lambda)$ for all $\lambda \in \Lambda$ with equality holding if and only if λ is a fixed point of τ.*

2. *λ is a critical point of L if and only if it is a fixed point of τ.*

3. *All limit points of $\tau^n(\lambda_0)$ are fixed points of τ for any $\lambda_0 \in \Lambda$.*

□

From Theorem 3.1, we can derive a maximum likelihood estimation method for the HMMR parameters by applying τ repeatedly until convergence. If $\tau^n(\lambda_0)$ converges, up to a prespecified level of precision, to a fixed point of τ, the resulting parameter $\tilde{\lambda}$ approximates a critical point of the likelihood function. In theory, this critical point need not be a local maximum of $L(\cdot)$, but it could be a saddle point. However, it is unlikely in practice, since the basins of attraction of such saddle points are low-dimensional stable manifolds. It is usually a good idea to attempt to find a global maximum by choosing the best among several estimates, obtained using different initializations λ_0.

To make the result of Theorem 3.1 useful, we need to reexpress the transformation in a computable form, such as in the following result.

Theorem 3.2 *Let $\alpha_t(i), \beta_t(i)$ be the forward and backward variables respectively, and let $\phi(X_t^T\beta, \sigma^2, y_t)$ be the value of a normal density with mean $X_t^T\beta$ and variance σ^2 at the point y_t. Then, given $\lambda \in \Lambda$, the transformation $\hat{\lambda} = \tau(\lambda)$ can be expressed as,*

$$
\begin{aligned}
\hat{\beta}_i &= \begin{bmatrix} ss_{x_0,x_0} & ss_{x_0,x_1} & \cdots & ss_{x_0,x_p} \\ \vdots & \vdots & & \vdots \\ ss_{x_p,x_0} & ss_{x_p,x_1} & \cdots & ss_{x_p,x_p} \end{bmatrix}^{-1} \times \begin{bmatrix} ss_{x_0,y} \\ \vdots \\ ss_{x_p,y} \end{bmatrix} \\
\hat{\sigma}_i^2 &= \frac{\sum_{t=1}^T \alpha_t(i)\beta_t(i)(y_t - X_t^T\hat{\beta}_i)^2}{\sum_{t=1}^T \alpha_t(i)\beta_t(i)} \\
\hat{a}_{ij} &= \sum_{t=1}^{T-1} \alpha_t(i) a_{ij} \phi(X_{t+1}^T\beta_j, \sigma_j^2, y_{t+1})\beta_{t+1}(j) / \sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i) \\
\hat{\pi}_i &= \alpha_1(i)\beta_1(i) / \sum_{j=1}^{N} \alpha_T(j)
\end{aligned}
$$

where $ss_{x_\ell,x_{\ell'}} \overset{def}{=} \sum_{t=1}^T \alpha_t(i)\beta_t(i)x_{t\ell}x_{t\ell'}, \quad ss_{x_\ell,y} \overset{def}{=} \sum_{t=1}^T \alpha_t(i)\beta_t(i)x_{t\ell}y_t.$

□

Careful examination of the reestimation formulas reveals a simple method of solution that requires only to perform an iterative reweighted least squares calculation with weights $\sqrt{\alpha_t(i)\beta_t(i)}$.

HMMR Algorithm

```
Initialization:
    Choose $\hat{\lambda} \in \Lambda$ and a level of precision $\epsilon > 0$;
Iteration:
    For $\lambda \leftarrow \hat{\lambda}$, calculate $\alpha_t(i), \beta_t(i)$ for $1 \leq i \leq N, 1 \leq t \leq T$,
    using the forward-backward algorithm;
    Calculate new parameter values $\hat{\lambda}$, according to the equations
    in Theorem 3.2 as follows:
    Calculate $\hat{a}_{ij}, \hat{\pi}_i$;
    Calculate $\hat{\beta}_i, \hat{\sigma}_i^2$, by means of performing a weighted
    least squares regression with weights $\sqrt{\alpha_t(i)\beta_t(i)}$;
    Repeat until the distance $\mid \hat{\lambda} - \lambda \mid < \epsilon$;
```

Remarks:

1. The requirement that $\tau(\lambda) \in \Lambda$ is automatically satisfied in our case, as it is evident from the formulas in Theorem 3.2.

2. One can prove that for most practical cases the MLE for Π is a unit vector with point mass at a single state. In fact, it is clear that consistent estimation of Π from a single sequence is impossible.

3. In the degenerate case with a single state, the HMMR algorithm will converge to a fixed point after a single iteration and so provide the same estimator as the one obtained by the OLS estimates.

3.4 State Estimation

In many applications one is often interested in providing an estimate of the unobserved state sequence that led to a given observation sequence. The state estimation method we shall adopt here is the *maximal aposteriori probability* (MAP) method, by which we estimate Q_t by the state S_i that maximizes the marginal aposteriori probability $P_\lambda(Q_t|Y^s)$, $1 \leq s \leq T$, where λ is substituted by its MLE and $Y^s \overset{def}{=} (Y_1, \ldots, Y_s)$.

We begin by considering the smoothed state estimates, i.e. $s = T$. The posterior state probabilities are then

$$P_\lambda(Q_t = S_j|Y^T) = P_\lambda(Y^T, Q_t = S_j)/P_\lambda(Y^T)$$

$$= \alpha_t(j)\beta_t(j)/P_\lambda(Y^T) = \alpha_t(j)\beta_t(j)/\sum_k \alpha_T(k).$$

For filtering purposes we need to maximize over

$$P_\lambda(Q_t = S_j|Y^t) = \frac{P_\lambda(Y_t|Q_t = S_j, Y^{t-1}, X^{t-1})P_\lambda(Q_t = S_j|Y^{t-1})}{P_\lambda(Y^t|Y^{t-1})}$$
$$= P_\lambda(Y_t|Q_t = S_j)P_\lambda(Q_t = S_j, Y^{t-1})/P_\lambda(Y^t) = \alpha_t(j)/\sum_k \alpha_t(k),$$

while for one-step forecasting we need to maximize over

$$P_\lambda(Q_{t+1} = S_j|Y^t) = \sum_\ell P_\lambda(Q_{t+1} = S_j, Q_t = S_\ell|Y^t)$$
$$= \sum_\ell P_\lambda(Y^t|Q_{t+1} = S_j, Q_t = S_\ell)P_\lambda(Q_{t+1} = S_j, Q_t = S_\ell)/P_\lambda(Y^t)$$
$$= \sum_\ell P_\lambda(Y^t, Q_t = S_\ell)a_{\ell j}/P_\lambda(Y^t) = \sum_\ell \alpha_t(\ell)a_{\ell j}/\sum_k \alpha_t(k).$$

One can similarly derive m-step state predictions, for example a two-step state MAP prediction is obtained by maximizing over

$$P_\lambda(Q_{t+2} = S_j|Y^t) = \sum_r \sum_\ell \alpha_t(r)a_{r\ell}a_{\ell j}/\sum_k \alpha_t(k).$$

In all of these examples we obtain simple formulas in the forward and backward variables that are readily available from the estimation process.

4 Simulation Studies

We conducted a series of experiments to explore some important properties of the algorithm and the resulting *HMMR estimator.* We report on two experiments that were designed to address the following issues:

(A) The most important question we ask about the HMMR estimator is whether it is a consistent estimator for λ. Once we are assured of consistency, the natural question to ask is what are reasonable sequence sizes required to obtain accurate and stable estimates?

(B) The most classical question is to ask if the estimates are asymptotically normally distributed, and what sequence size is needed for such an approximation.

Table 1: Parameter values for true model λ_0 in Simulation A.

parameter							
a_{11}	a_{22}	β_{01}	β_{02}	β_{11}	β_{12}	σ_1^2	σ_2^2
0.5	0.9	10	15	1	1	1	25

For simplicity, the simulation studies we present are based on simple HMMR's with only two possible states. The single covariate for the regression analysis was generated uniformly on the interval $[0-10]$, thus implying low probability for high leverage. Data for the simulations were generated using the *Splus* statistical software. A program in *C* was developed for the estimation of the parameters using the HMMR algorithm.

4.1 Simulation A: Accuracy and Stability of the HMMR Estimates

In the first experiment we focus on the behavior of the HMMR estimate as the observation sequence size T is increased from 50 to 700. A natural metric to measure the distance of estimators from the true model parameters is the Kullback-Leibler divergence,

$$K(\lambda_0, \lambda) \overset{def}{=} \lim_{T} \frac{1}{T}\{\log L(o_1, \ldots, o_T; \lambda_0) - \log L(o_1, \ldots, o_T; \lambda)\},$$

where λ_0 is the true parameter. For a finite sequence of length T, we define the *sample Kullback-Leibler divergence* between two parameter points as

$$K_T(\lambda_0, \lambda) \overset{def}{=} \frac{1}{T}\{\log L(o_1, \ldots, o_T; \lambda_0) - \log L(o_1, \ldots, o_T; \lambda)\}.$$

We shall use the stochastic distance function $\tilde{K}_T(\lambda_0, \tilde{\lambda})$ to measure the distance between the HMMR estimate $\tilde{\lambda}$ and λ_0. This distance measure was effectively used in earlier studies (see Juang and Rabiner (1985)). The values for the parameter λ_0 were chosen as shown in Table 1.

Values of $\lambda^{(0)}$, the initial parameter for the HMMR algorithm, were obtained by randomly perturbing the true parameter values by up to 20% of

Figure 1: Simulation A: Box plots of $\tilde{K}(\lambda_0, \tilde{\lambda}_T)$ for various sequence sizes T.

their true value, provided the result lies in the interior of Λ. Termination of the HMMR algorithm occurred when the relative change in each component of the estimated parameter values, are all smaller than a threshold value chosen as .0001 (initial probabilities excluded).

For each value of T, the estimation procedure was carried out twenty times, and the distances $\tilde{K}_T(\lambda_0, \tilde{\lambda}_T)$ between each of the twenty estimators and the true parameter λ_0 were evaluated on a new sequence, independent of the first twenty sequences used to obtain the estimators. This way we prevent the potential underestimation of the distance as a result of estimating the parameter and evaluating its performance on the same sequence.

Box plots of the sets of distances for the various values of T are presented under a unified scale in Figure 1. The plots clearly show a general decrease in average and spread of the distances with increasing T. Given that small values of $\tilde{K}_T$ imply similarity between λ_0 and $\tilde{\lambda}_T$, the results of this experiment suggest increasing accuracy and stability of the sequence of HMMR estimators as T increases.

4.2 Simulation B: Asymptotic Normality of the HMMR Estimates

This experiment investigates the asymptotic distribution of $\tilde{\lambda}_T$, the HMMR estimate. We generated 100 sequences of size T following the same proce-

Table 2: Summary of results for Simulation B.

				parameter				
	a_{11}	a_{22}	α_1	α_2	β_1	β_2	σ_1^2	σ_2^2
Initial values	0.3	0.9	3	3	1	1	1	15
True model	0.9	0.75	4	1	1	2	1	25
Mean	.9004	.7531	3.992	.9419	1.002	1.996	.9919	24.59
Median	.9008	.7577	3.993	1.026	1.003	1.996	.9850	24.19
STD	.0239	.0674	.1489	1.085	.0235	.1860	.1150	4.668
P value	.728	.394	.765	.984	.055	.580	.434	.614

dure as described in Simulation A. Parameter initialization for all the 100 replications was fixed at prespecified values. The values of the initial parameter $\lambda^{(0)}$ and the true model parameter λ_0 are given in the first and second rows of Table 2. Note that the true parameter values here are chosen so that the two regression lines intersect (at $(x, y) = (3, 7)$) and therefore under this model the points are in particular hard to classify to the correct state.

We used the Shapiro-Wilk statistic to test the univariate normality of each component of $\tilde{\lambda}_T$. The sequence size T was increased until the normality null hypotheses could not be rejected at the 0.05 level of significance.

Rows three to five of Table 2 show the mean, median and standard deviation of the parameter estimates for a simulation with $T = 300$, and the last row gives the significance value for the normality test. The results suggest that for large T the estimates distribution tends to be normal. A sequence size of 300 and up would be required to use a normal approximation for the distribution of the HMMR parameter estimates parameter of a simple HMMR with two states.

5 Conclusions

Two powerful and popular tools of modeling data are linear models and HMM's. In this paper we propose a model that combines these tools and call it hidden Markov model regression. We formulate the model and develop maximum likelihood estimates for its parameter by applying the Baum-Welch approach. Simulation results provide empirical evidence of the usefulness of the estimates. A rigorous study of the inferential properties of likelihood methods is a difficult task, since the process under consideration is in general not Markovian and not homogeneous in time.

An interesting direction for further research is the generalization of our model to one where one allows for a dependence of both the dependent and independent variables on the states of the Markov chain. In many situations it would be natural to assume such a dependence, that would result in a random covariates model in contrast to our fixed covariate model. In the opposite direction, one can also think of the case where the state transition probabilities are not homogeneous in time, but depend on the previous state and the previously observed covariates levels. The study of such models would provide a further step in the extension of hidden Markov models to regression analysis and allow for further flexibility in applications.

Appendix: Proofs

Proof of Theorem 3.1 We begin by showing that successive applications of the transformation τ on any $\lambda \in \Lambda$ provides an increasing sequence of likelihood function values. By the Kullback-Leibler inequality we have that for any $\lambda, \hat{\lambda} \in \Lambda$,

$$\sum_q P_\lambda(Q|Y) \ln P_\lambda(Q|Y) \geq \sum_q P_\lambda(Q|Y) \ln P_{\hat{\lambda}}(Q|Y). \tag{10}$$

Also, it is easy to verify that

$$\ln P_{\hat{\lambda}}(Y) = \frac{Q(\lambda, \hat{\lambda})}{P_\lambda(Y)} - \sum_q P_\lambda(Q|Y) \ln P_{\hat{\lambda}}(Q|Y). \tag{11}$$

Hence, if we let $\hat{\lambda} = \tau(\lambda)$ and show that $Q(\lambda, \lambda')$ is strictly concave in λ' then by (10) and (11) we have that $\ln P_{\hat{\lambda}}(Y) \geq \ln P_\lambda(Y)$. Furthermore, we get that $\hat{\lambda}$ is the unique solution for $\bigtriangledown_{\lambda'} Q(\lambda, \lambda') = 0$. By the partition (9) of $Q(\lambda, \lambda')$, we need only to consider the terms involving the primed parameter $\theta' = (\theta_1', \ldots, \theta_N')$. The rest of the terms involving the primed parameters A' and Π' are strictly concave by Lemma 3.1. Let y be a fixed observed sequence, and let

$$h(\theta_j', y) \overset{def}{=} \sum_q \sum_{\{t:q_t=S_j\}} P_\lambda(y, Q) \ln \left[(2\pi\sigma_j^{2'})^{-1/2} \exp\{-(y_t - X_t^T \beta_j')^2/(2\sigma_j^{2'})\} \right]$$

$$= \quad -\sum_{t=1}^{T} c_{tj}(y)\ln(2\pi\sigma_j^{2'}) - \sum_{t=1}^{T} c_{tj}(y)(y_t - X_t^T\beta_j^{'})^2/\sigma_j^{2'},$$

where $c_{tj}(y) \overset{def}{=} \sum_{\{q:q_t=S_j\}} P_\lambda(y,Q)/2$ and $\theta_j^{'} = (\beta_{0j}^{'},\ldots,\beta_{pj}^{'},\sigma_j^{2'})$, $1 \leq j \leq N$. It is straightforward to check that under the assumptions that X is of full column rank and A is ergodic, the function $h(\theta,y)$ is strictly concave in each element of θ for any y. This is done by showing that $h(\theta,y) \to -\infty$ as θ approaches the boundary of Θ, and then showing that for every critical point θ_0 of $h(\cdot)$ the Hessian matrix at θ_0, $H(\theta_0)$, is negative definite. Then, by Morse theory, $h(\cdot)$ has only one local maximum, that is a unique global maximum. This completes the proof of the first statement of the Theorem.

The proof of the second statement of the theorem on the equivalence between fixed points of τ or a critical points of $L(\cdot)$ follows from the fact that, since $P_\lambda(Y,Q)$ is continuously differentiable in λ, we have

$$\left.\frac{\partial L(\lambda)}{\partial \lambda_i}\right|_{\lambda} = \left.\frac{\partial Q(\lambda,\lambda^{'})}{\partial \lambda_i^{'}}\right|_{\lambda^{'}=\lambda}, \tag{12}$$

where λ_i denotes a single component of λ.

The last part of the theorem addresses the convergence properties of the iterates of the transformation τ. If λ^* is a limit point of $\tau^n(\lambda)$ then

$$L(\lambda^*) \leq L(\tau(\lambda^*)) = \lim_i L(\tau^{n_i+1}(\lambda_0)) \leq \lim_i L(\tau^{n_{i+1}}(\lambda_0)) = L(\lambda^*),$$

implying that $\tau(\lambda^*) = \lambda^*$, namely all limit point of τ^n are fixed points of τ.

Proof of Theorem 3.2 First, we differentiate (9) with respect to $\beta_{\ell i}^{'}, \sigma_i^{2'}, a_{ij}^{'}$, and $\pi^{'}$, and set the derivatives equal to zero in order to find the critical point of $Q(\lambda,\lambda^{'})$ in $\lambda^{'}$. We obtain the following so called *reestimation formulas*,

$$\hat{\beta}_i = (\hat{\beta}_{0i},\ldots,\hat{\beta}_{pi})^T = M_i^{-1}V_i \tag{13}$$

$$\hat{\sigma}_i^2 = \frac{\sum_q P_\lambda(Y,Q)\sum_{\{t:q_t=S_i\}}(Y_t - X_t\hat{\beta}_i)^2}{\sum_q P_\lambda(Y,Q)\sum_{\{t:q_t=S_i\}}1} \tag{14}$$

$$\hat{a}_{ij} = \frac{\sum_q \sum_{t=2}^T P_\lambda(Y,Q)1_{(q_{t-1}=S_i,q_t=S_j)}}{\sum_q \sum_{t=2}^T P_\lambda(Y,Q)1_{(q_{t-1}=S_i)}} \tag{15}$$

$$\hat{\pi}_i = \frac{\sum_q P_\lambda(Y,Q)1_{(q_1=S_i)}}{P_\lambda(Y)} \tag{16}$$

where M_i is a $(p+1) \times (p+1)$ matrix and V_i is a $(p+1) \times 1$ vector with elements

$$\begin{aligned} (M_i)_{\ell,\ell'} &\overset{def}{=} \sum_q P_\lambda(Y,Q) \sum_{\{t:q_t=S_i\}} x_{t\ell} x_{t\ell'}, \quad 0 \le \ell, \ell^{'} \le p, \\ (V_i)_\ell &\overset{def}{=} \sum_q P_\lambda(Y,Q) \sum_{\{t:q_t=S_i, 1\le t \le T\}} x_{t\ell} Y_t, \quad 0 \le \ell \le p. \end{aligned}$$

The existence of the inverse matrices M_i^{-1} can be easily verified. The form of the reestimation formulas stated in the result is obtained by using the definition of $\alpha_t(i), \beta_t(i)$ and interchanging the summations over q and t in all the double summations in the equations (13)-(15). For example,

$$\begin{aligned} \sum_q P_\lambda(Y,Q) \sum_{\{t:q_t=S_i\}} x_{t\ell}^2 = \sum_{t=1}^{T} \{ \sum_{\{q:q_t=S_i\}} P_\lambda(Y,Q) \} x_{t\ell}^2 = \\ \sum_{t=1}^{T} P_\lambda(Y, Q_t = S_i) x_{t\ell}^2 = \sum_{t=1}^{T} \alpha_t(i) \beta_t(i) x_{t\ell}^2. \end{aligned} \tag{17}$$

As for $\hat{\pi}_i$, we need only to use the definition of $\alpha_t(i), \beta_t(i)$.

Acknowledgements This work is drawn from the authors Ph.D. dissertation written at the university of Pennsylvania under the direction of Professor J. Michael Steele to whom I am indebted for his invaluable guidance and advice.

References

Baum, L.E. (1972), "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities,III: Proceedings of the symposium*, (Shisha, Qved ed.), New York: Academic Press, 1-8.

Baum, L.E. and Eagon, J.A. (1967), "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of Ecology," *Bull. Amer. Math. Soc.*, **73**, 360-363.

Baum, L.E. and Petrie, T. (1966), "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, **37**, 1554-1563.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970), "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, **41**, 164-171.

Blackwell, D. and Koopmans, L. (1957), "On the identifiability problem for functions of finite Markov chains," *Ann. Math. Statist.*, **28**, 1011-1015.

Gilbert, E.J. (1959), "On the identifiability problem for functions of finite Markov chains," *Ann. Math. Statist.*, **30**, 688-697.

Hamilton, J.D. (1989), "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, **57**, 357-384.

Juang, B.H. (1985), "Maximum Likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT & T Tech. J.*, **64**, 1235-1249.

Juang, B.H., Levinson, S.E. and Sondhi, M.M. (1986), "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform. Theory*, **IT-32**, 307-309.

Juang, B.H. and Rabiner, L.R. (1985), "A probabilistic distance measure for hidden Markov models," *AT & T Tech. J.*, **64**, 391-408.

Kiefer, N.M. (1978), "Discrete parameter variation: Efficient estimation of a switching regression model," *Econometrica*, **46**, 427-434.

Leroux, B.G. (1992), "Maximum-likelihood estimation for hidden Markov models," *Stochastic Process. Appl.*, **40**, 127-143.

Levinson, S.E., Rabiner, L.R. and Sondhi, M.M (1983), "An introduction to the applications of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, **62**, 1035-1074.

Liporace, L.A. (1982), "Maximum Likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, **IT-28**, 729-734.

Petrie, T. (1969), "Probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, **40**, 97-115.

Quandt, R.E. (1972), "A new approach to estimating switching regressions," *J. Amer. Statist. Assoc.*, **67**, 306-310.

Quandt, R.E. and Ramsey, J.B. (1978), "Estimating mixtures of Normal distributions and switching regressions," *J. Amer. Statist. Assoc.*, **73**, 730-738.