MULTIPLE-PREDICTION-HORIZON RECURSIVE IDENTIFICATION OF HIDDEN MARKOV MODELS *

Iain B. Collings † *John B. Moore* ‡

† Cooperative Research Centre for Sensor Signal and Information Processing, SPRI Building, Technology Park, The Levels, SA 5095, Australia

‡ Department of Systems Engineering, RSISE, Australian National University, Canberra ACT 0200, Australia.

ABSTRACT

This paper considers on-line identification of hidden Markov models via multiple-prediction-horizon recursive prediction error (RPE) methods. Working with multiple prediction horizons ensures that there is consistent parameter estimation, under appropriate excitation conditions. Simulation studies are included to illustrate the advantages of the proposed approach when compared to standard methods (which do not ensure consistent parameter estimation).

1. INTRODUCTION

Recently, Hidden Markov models (HMMs) with states in a finite-discrete set have been widely applied in many areas of signal processing. Applications include communication systems [1], speech processing [2], frequency tracking [3], and biological signal processing [4]. In each of these areas, on-line identification can have many advantages.

In [5], a sequential linear convergent expectation-maximisation (EM) algorithm is presented for on-line identification of HMMs. A quadratically convergent scheme is achieved in [6], via recursive prediction error (RPE) techniques. Unfortunately, when applied to HMMs, the RPE approach sometimes results in convergence to local, rather than global, minimuma, or at least to biased parameter estimates.

This paper presents a modification to the scheme in [6], which provides consistent parameter estimation in cases where previous on-line schemes have identifiability problems (especially in low noise environments). The parameters to be identified are the transition probabilities and state values of the Markov chain (The measurement noise variance can also be estimated, however this is not presented here).

A key to the approach is that instead of simply using a prediction one time step ahead, we use predictions over multiple time horizons. This achieves improved parameter observability. In cases where a biased estimate is found for one step prediction, the scheme presented here, achieves consistent estimation (under persistence of excitation conditions) by combining the estimates obtained from multiple-horizon predictions.

The model parametrisation considered here, uses the square root of the transition probabilities constrained to the surface of a sphere in $\mathbb{R}^N$ [6]. The derivatives of the prediction error, for the RPE scheme, are thus constrained to the tangent space of the smooth manifold. The advantage of working on the sphere is that estimates of transition probabilities are assured to be non-negative, and derivatives are smooth.

Simulation examples are presented to illustrate the comparative advantages of the proposed algorithms. These examples show that the proposed schemes can satisfactorily identify HMM parameters in cases where previous prediction error approaches result in biased estimates.

*Supported by the Cooperative Research Centres for Sensor Signal and Information Processing (CSSIP), and for Robust and Adaptive Systems ((CR)^{2}ASYS)

2. SIGNAL MODEL

2.1. State Space Signal Model

Let X_k be a discrete-time homogeneous, first order Markov process belonging to a finite-discrete set. The term finite-discrete is used to indicate that the set has a finite number of elements which have discretized values. The state space of X, without loss of generality, can be identified with the set of unit vectors $\mathbf{S} = \{e_1, e_2, ..., e_N\}$, where $e_i = (0, ..., 0, 1, 0, ..., 0)' \in \mathbb{R}^N$, with 1 in the i^{th} position. These vectors are termed *indicator vectors*, as they indicate which of the discrete-states is active at each time k.

The probability of transitioning from state i to state j is denoted by $a_{ij} = P(X_{k+1} = e_j \mid X_k = e_i)$. These probabilities are the elements of the state transition probability matrix, $\mathbf{A}$. Of course $a_{ij} \geq 0$, for all i, j, and $\sum_{j=1}^{N} a_{ij} = 1$, for each i.

The dynamics of X_k are given by the state equation

$$X_{k+1} = \mathbf{A}' X_k + M_{k+1}$$

where M_{k+1} is a martingale increment [7]. This equation, while not being used explicitly in the remainder of this paper, provides a major clue to the use of recursive prediction error techniques for HMM identification. It is now possible to see, more clearly, the parallels between HMMs and standard linear and nonlinear systems.

The *observation process*, y_k, is a linear function of the state X_k plus additive noise. It is in this sense that the Markov model is *hidden*. Without loss of generality we can write y_k in the following form:

$$y_k = g' X_k + w_k \tag{1}$$

where $w_k \sim N[0, \sigma_w^2]$, and $g \in \mathbb{R}^N$ is the vector of *state-values* of the Markov chain. Let $\mathcal{Y}_l$ be the σ-field generated by y_k, $k \leq l$, and define $Y_k \triangleq (y_0, \ldots, y_k)$.

It should be noted that any nonlinear function of an indicator vector can be represented in a linear form. Therefore, (1) is a quite general function, not limited to linear systems.

2.2. Model Parameterisation

Consider that the HMM is parametrised by an unknown vector θ (that is, the elements which define the HMM, namely $\mathbf{A}$ and g, are functions of θ). Of course, the values this parameter can take are constrained by the fact that $\mathbf{A}$ is a stochastic matrix.

As in [6], the following parametrisation is used:

$$\theta = (g_1, ..., g_N, s_{11}, ..., s_{1N}, s_{21}, ..., s_{NN})' ,$$

where $a_{ij} = s_{ij}^2$. Defining θ in this way ensures the constraint manifold is differentiable at all points. Also, estimates of a_{ij} will always be positive, and only the equality constraint of the sphere surface, $\mathbf{S}^{N-1}$, remains, where

$$\mathbf{S}^{N-1} = \left\{ s_{ij} : \sum_{j=1}^{N} s_{ij}^2 = 1 \right\} . \qquad (2)$$

2.3. Parametrised Information State Signal Model

Let $\hat{X}_{k|\theta}$ denote the conditional filtered state estimate of X_k at time k, that is,

$$\hat{X}_{k|\theta} \triangleq E[X_k|\mathcal{Y}_k, \theta] = \langle \alpha_{k|\theta}, \underline{1} \rangle^{-1} \alpha_{k|\theta} ,$$

where $\underline{1}$ is the column vector containing all ones, and the "forward" variable $\alpha_{k|\theta}$ is such that the i^{th} element $\alpha_{k|\theta}(i) \triangleq P(Y_k, X_k = e_i \mid \theta)$.

The *information state*, $\alpha_{k|\theta}$, is conveniently computed using the following recursion [2]:

$$\alpha_{k+1|\theta} = \mathbf{B}(y_{k+1}, \theta)\, \mathbf{A}'(\theta)\, \alpha_{k|\theta} \qquad (3)$$

where $\mathbf{B}(y_k, \theta) = \text{diag}(\, b(y_k, g_1), ..., b(y_k, g_N)\,)$, and $b(y_k, g_i) = P[y_k \mid X_k = e_i, \theta]$.

3. THE MULTIPLE-PREDICTION-HORIZON APPROACH

The aim of the identification task is to estimate θ, based on the observations $\mathcal{Y}_k$. The approach in [6] employs an RPE algorithm which evaluates a prediction error, one-step ahead, and uses it to update $\hat{\theta}_k$, the recursive estimate of the parameter vector based on $\mathcal{Y}_k$. The one-step prediction error is defined by

$$\hat{n}_{k|k-1} = y_k - \hat{y}_{k|k-1} ,$$

where $\hat{y}_{k|k-i}$ denotes the predicted output at time k based on measurements up to time $k-i$. For one-step prediction it is given by

$$\hat{y}_{k|k-1} = \hat{g}'_{k-1,1} \hat{\mathbf{A}}'_{k-1,1} \hat{X}_{k-1|\hat{\Theta}_{k-1,1}} , \qquad (4)$$

where $\hat{g}_{k,1} = g(\hat{\theta}_{k,1})$ and $\hat{\mathbf{A}}_{k,1} = \mathbf{A}(\hat{\theta}_{k,1})$ (this notation is used to indicate that the estimates for g and $\mathbf{A}$ are functions of the parameter estimate $\hat{\theta}$). Also, $\hat{\Theta}_{k,1} \triangleq \{\hat{\theta}_{1,1}, ..., \hat{\theta}_{k,1}\}$, and

$$\hat{X}_{k-i|\hat{\Theta}_{k-i,i}} = \langle \alpha_{k-i|\hat{\Theta}_{k-i,i}}, \underline{1} \rangle^{-1} \alpha_{k-i|\hat{\Theta}_{k-i,i}} ,$$

where the second subscript denotes that the parameters are those evaluated by the i-step-ahead prediction scheme (where $i = 1$ in equation (4)).

Good results can be obtained from this approach. Unfortunately, however, in some cases it is possible that even though the *product* $g'\mathbf{A}'$ is consistently estimated, g and $\mathbf{A}$ themselves, are not. In order to overcome this identifiability problem, we propose a multiple-prediction-horizon RPE scheme.

Our approach is to have a number of HMM/RPE on-line identification algorithms operating in parallel, each with a different prediction horizon. For example, take a state model with $N = 2$ for simplicity. The two-step ahead measurement prediction is

$$\hat{y}_{k|k-2} = \hat{g}'_{k-2,2} \hat{\mathbf{A}}'_{k-2,2} \hat{\mathbf{A}}'_{k-2,2} \hat{X}_{k-2|\hat{\Theta}_{k-2,2}} . \qquad (5)$$

Exploiting both (4) and (5) in an RPE scheme allows consistent estimation of both $g'\mathbf{A}'$ and $g'\mathbf{A}'\mathbf{A}'$, with $\hat{X}_{k-i|\hat{\Theta}_{k-i,i}}$ persistently exciting for $i = 1, 2$. Now, even if $\hat{g}_{k,i}$ and $\hat{\mathbf{A}}_{k,i}$ do not approach g and $\mathbf{A}$ in either the one-step-ahead ($i = 1$) or two-step-ahead ($i = 2$) prediction case, there can be consistent estimation by solving the following simultaneous equations:

$$\begin{aligned} \hat{g}'_k \hat{\mathbf{A}}'_k &= \hat{g}'_{k,1} \hat{\mathbf{A}}'_{k,1} \\ \hat{g}'_k \hat{\mathbf{A}}'_k \hat{\mathbf{A}}'_k &= \hat{g}'_{k,2} \hat{\mathbf{A}}'_{k,2} \hat{\mathbf{A}}'_{k,2} \\ \hat{\mathbf{A}}_k \underline{1} &= \underline{1} \end{aligned}$$

where $\hat{g}_k$ and $\hat{\mathbf{A}}_k$ are the resultant estimates derived from estimates $\hat{g}_{k,i}$ and $\hat{\mathbf{A}}_{k,i}$ from the i-step ahead RPE scheme.

The fact that the products of g and $\mathbf{A}$ are estimated correctly, might suggest a different choice of parametrisation, namely $\theta = (g'\mathbf{A}', g'\mathbf{A}'\mathbf{A}')'$. Unfortunately, however, the derivatives necessary for any gradient descent algorithm can not be calculated for such a parametrisation, as will be seen in Section 5 (specifically, $\partial\alpha/\partial\theta$ can not be evaluated).

A block diagram of the multiple-prediction-horizon approach is given in Figure 1.

4. THE MULTIPLE-PREDICTION-HORIZON RPE ALGORITHM

In this section we present the on-line prediction error algorithm for estimation of θ. Here the estimate of α_k is recursively computed at each iteration, by substituting the best estimate of θ at each time k into (3):

$$\hat{\alpha}_{k+1|\hat{\Theta}_k} = \mathbf{B}(y_{k+1}, \hat{\theta}_k)\, \mathbf{A}'(\hat{\theta}_k)\, \hat{\alpha}_{k|\hat{\Theta}_{k-1}} \qquad (6)$$

The RPE parameter update equations for the ith prediction horizon are, ([8] p. 94)

$$\hat{\theta}_k = \Gamma_{proj}\{\hat{\theta}_{k-1} + \gamma_k R_k^{-1} \psi_{k|k-i}\, \hat{n}_{k|k-i}\}$$

$$R_k^{-1} = \frac{1}{1-\gamma_k} \left(R_{k-1}^{-1} - \frac{\gamma_k R_{k-1}^{-1} \psi_{k|k-i} \psi'_{k|k-i} R_{k-1}^{-1}}{(1-\gamma_k) + \gamma_k \psi'_{k|k-i} R_{k-1}^{-1} \psi_{k|k-i}} \right)$$

where γ_k is a gain sequence (often referred to as step size), and $\hat{n}_{k|k-i}$ is the i step ahead prediction error, $\hat{n}_{k|k-i} =$

$y_k - \hat{y}_{k|k-i}$. Also $\psi_{k|k-i}$ is the gradient

$$\psi_{k|k-i} \triangleq \left(-d\hat{n}_{k|k-i}/d\theta\right)'\Big|_{\theta=\hat{\theta}_{k-i}}, \qquad (7)$$

and R_k is the Hessian, or covariance matrix, approximation. The notation $\Gamma_{proj}\{.\}$ represents a projection into the constraint domain, given by the manifold $\mathbf{S}^{N-1}$ in (2).

5. ON-LINE GRADIENT VECTOR AND PROJECTION CALCULATIONS

In this section we consider only the two step prediction case, however extension to M-steps follows directly.

The derivative vector, $\psi_{k|k-i}$, defined in (7), is given, for $m, n \in [1, ..., N]$, by

$$\psi_{k|k-i} = \frac{\partial \hat{y}_{k|k-i}}{\partial \theta}\Big|_{\theta=\hat{\theta}_{k-i}} = \left(\frac{\partial \hat{y}_{k|k-i}}{\partial g_m}, \frac{\partial \hat{y}_{k|k-i}}{\partial s_{mn}}\right)'\Big|_{\theta=\hat{\theta}_{k-i}}$$

where the one-step and two-step output predictions, $\hat{y}_{k|k-1}$ and $\hat{y}_{k|k-2}$ are given in (4) and (5) respectively. In the remainder of this paper we omit the obvious dependence of $\hat{\alpha}_k$ on $\hat{\Theta}_{k-1}$ (from (6)), and we denote $N_k \triangleq \langle \hat{\alpha}_k, \underline{1} \rangle^{-1}$.

We now have the following expressions for the gradients, evaluated on the surface of the constraint manifold $\mathbf{S}^{N-1}$:

$$\begin{aligned}\frac{\partial \hat{y}_{k+1|k}}{\partial g_m} &= N_k a'_{(.)m}\hat{\alpha}_k + N_k g'\mathbf{A}'\eta_k(m) \\ &\quad - N_k^2 \underline{1}'\eta_k(m) g'\mathbf{A}'\hat{\alpha}_k \qquad (8)\end{aligned}$$

$$\begin{aligned}\frac{\partial \hat{y}_{k+1|k}}{\partial s_{mn}} &= 2N_k\hat{\alpha}_k(m)s_{mn}g'\left(e_n - \mathrm{diag}(s_{m(.)})s'_{m(.)}\right) \\ &\quad + N_k g'\mathbf{A}'\xi_k(m,n) - N_k^2\underline{1}'\xi_k(m,n)g'\mathbf{A}'\hat{\alpha}_k \qquad (9)\end{aligned}$$

$$\begin{aligned}\frac{\partial \hat{y}_{k+1|k-1}}{\partial g_m} &= N_{k-1}a'_{(.)m}\mathbf{A}'\hat{\alpha}_{k-1} \\ &\quad + N_{k-1}g'\mathbf{A}'\mathbf{A}'\eta_{k-1}(m) \\ &\quad - N_{k-1}^2\underline{1}'\eta_{k-1}(m)g'\mathbf{A}'\mathbf{A}'\hat{\alpha}_{k-1} \qquad (10)\end{aligned}$$

$$\begin{aligned}\frac{\partial \hat{y}_{k+1|k-1}}{\partial s_{mn}} &= 2N_{k-1}\left[\hat{\alpha}_{k-1}(m)s_{mn}g'\mathbf{A}'\right. \\ &\quad \left. + \left(\mathbf{A}'\hat{\alpha}_{k-1}\right)(m)s_{mn}g'\right]\left(e_n - \mathrm{diag}(s_{m(.)})s'_{m(.)}\right) \\ &\quad + N_{k-1}g'\mathbf{A}'\mathbf{A}'\xi_{k-1}(m,n) \\ &\quad - N_{k-1}^2\underline{1}'\xi_{k-1}(m,n)g'\mathbf{A}'\mathbf{A}'\hat{\alpha}_{k-1} \qquad (11)\end{aligned}$$

where $a_{(.)m} = (a_{1m}, \ldots, a_{Nm})'$ and $s_{m(.)} = (s_{m1}, \ldots, s_{mN})$. The N-dimensional vectors $\eta_k(j,m) \triangleq \partial\hat{\alpha}_k(j)/\partial g_m$ and $\xi_k(j,m,n) \triangleq \partial\hat{\alpha}_k(j)/\partial s_{mn}$ can be expressed *recursively* by the following equations, obtained using (6):

$$\begin{aligned}\eta_{k+1}(m) &= \mathbf{B}(y_{k+1},\theta)\mathbf{A}'\eta_k(m) \\ &\quad + \mathrm{diag}(e_m)\left(\frac{y_{k+1}-g_m}{\sigma_w^2}\right)\mathbf{B}(y_{k+1},\theta)\mathbf{A}'\hat{\alpha}_k \qquad (12)\end{aligned}$$

$$\begin{aligned}\xi_{k+1}(m,n) &= \mathbf{B}(y_{k+1},\theta)\mathbf{A}'\xi_k(m,n) \\ &\quad - 2s_{mn}\hat{\alpha}_k(m)\mathbf{B}(y_{k+1},\theta)\mathrm{diag}(s_{m(.)})s'_{m(.)} \\ &\quad + 2\hat{\alpha}_k(m)\mathrm{diag}(e_n)\mathbf{B}(y_{k+1},\theta)s'_{m(.)} \qquad (13)\end{aligned}$$

Of course, in this two state ($N = 2$) case, equations (8) and (9) would be used in the one-step-ahead RPE algorithm, while equations (10) and (11) would be for the two-step-ahead RPE algorithm. Equations (12) and (13) relate to both RPE algorithms, however of course, they would have different estimates of g and $\mathbf{A}$ in each case.

6. SIMULATION STUDIES

These examples demonstrate that the on-line multi-step algorithm presented in this paper, provides the global solution, for cases where the single step algorithm does not have a unique solution.

Example 1: in this example, the parameters of the Markov chain are $g = [0, 1]'$ and $a_{ii} = 0.8$. The SNR is therefore given by $10\log(1/\sigma_w^2)$. In this example a relatively low noise level is used in order to clearly demonstrate the undesirable behaviour of the single-step algorithm. Initial parameter estimates were $\hat{g} = [0.4, 0.6]'$ and $\hat{a}_{ii} = 0.1$.

The estimates for a typical observation sequence are shown in Figures 2 and 3. The estimates of the state values are presented in Figure 2 and the estimates of the transition probabilities are presented in Figure 3. In each case, the dashed lines are the single step estimates and the solid lines are the combined multi-step estimates. The figures show that the multi-step HMM/RPE algorithm converges to the correct values, even though the single-step algorithm does not.

Example 2: This example presents the outcome from more extensive Monticarlo simulations. The results are shown in Tables 1 to 3. Table 1 presents results for the HMM given in Example 1, while Tables 2 and 3 show the algorithm performance for different HMMs. Again, these results demonstrate the superior performance of the multi-step HMM/RPE algorithm presented in this paper.

REFERENCES

[1] I. B. Collings and J. B. Moore, "An HMM approach to adaptive demodulation of QAM signals in fading channels," *Int. Jour. of Adaptive Control and Signal Processing*, vol. 8, pp. 457–474, October 1994.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[3] X. Xianya and R. J. Evans, "Multiple target tracking and multiple frequency line tracking using hidden Markov models," *IEEE Trans. on Signal Processing*, vol. 39, pp. 2659–2676, December 1991.

[4] S. H. Chung, V. Krishnamurthy, and J. B. Moore, "Adaptive processing techniques based on hidden Markov models for characterising very small channel currents buried in noise and deterministic interferences," *Philosophical Transactions of the Royal Society, Lond., Series B*, vol. 334, pp. 357–384, 1991.

[5] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. on Signal Processing*, vol. 41, no. 8, pp. 2557–2573, 1993.

[6] I. B. Collings, V. Krishnamurthy, and J. B. Moore, "On-line identification of hidden Markov models via recursive prediction error techniques," *IEEE Trans. on Signal Processing*, vol. 42, pp. 3535–3539, December 1994.

[7] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov models : estimation and control.* New York: Springer-Verlag, 1995.

[8] L. Ljung and T. Söderström, *Theory and practice of recursive identification.* Cambridge, Massachusetts: MIT Press, 1983.

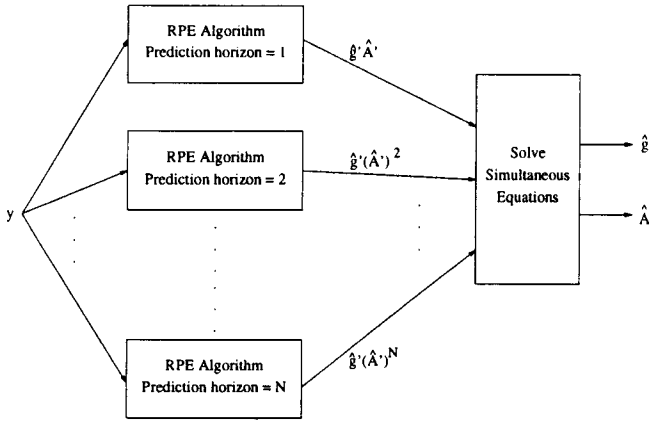

Figure 1. Block diagram of Multiple-Prediction-Horizon RPE scheme

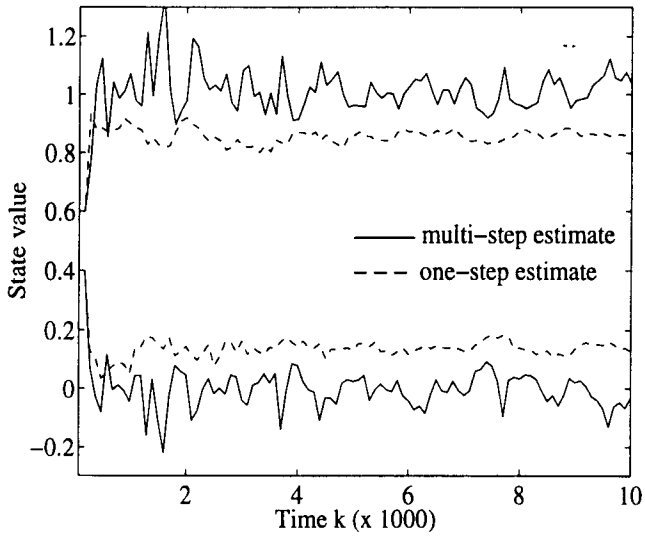

Figure 2. State value estimates

Parameter	True Value	Mean of Estimate	
		One-step	Two-step
g_1	0	0.3187	-0.0736
g_2	1	0.6786	1.0506
a_{11}	0.8	0.8895	0.7755
a_{12}	0.2	0.1105	0.2245
a_{21}	0.2	0.0830	0.2227
a_{22}	0.8	0.9170	0.7773

Table 1. Mean Values of Parameter Estimates

Parameter	True Value	Mean of Estimate	
		One-step	Two-step
g_1	0	0.5378	-0.0840
g_2	2	1.8955	2.0244
a_{11}	0.7	0.9394	0.6716
a_{12}	0.3	0.0606	0.3284
a_{21}	0.1	0.0793	0.1075
a_{22}	0.9	0.9207	0.8925

Table 2. Mean Values of Parameter Estimates

Parameter	True Value	Mean of Estimate	
		One-step	Two-step
g_1	0	0.2269	-0.0507
g_2	1	0.7721	1.0603
a_{11}	0.7	0.8194	0.6893
a_{12}	0.3	0.1806	0.3107
a_{21}	0.3	0.1807	0.3143
a_{22}	0.7	0.8193	0.6857

Table 3. Mean Values of Parameter Estimates

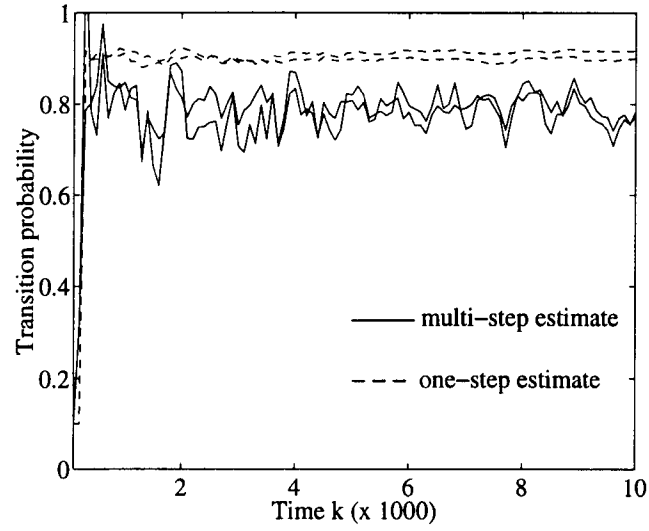

Figure 3. Transition probability estimates