A Method of Hidden Markov Model Optimization for Use with Geophysical Data Sets

Robert A. Granat[1]

Jet Propulsion Laboratory Pasadena CA 91106, USA

Abstract. Geophysics research has been faced with a growing need for automated techniques with which to process large quantities of data. A successful tool must meet a number of requirements: it should be consistent, require minimal parameter tuning, and produce scientifically meaningful results in reasonable time. We introduce a hidden Markov model (HMM)-based method for analysis of geophysical data sets that attempts to address these issues. Our method improves on standard HMM methods and is based on the systematic analysis of structural local maxima of the HMM objective function. Preliminary results of the method as applied to geodetic and seismic records are presented.

1 Introduction

In recent years, geophysics research has been faced with a growing need for automated techniques by which to process the ever-increasing quantities of geophysical data being collected. Global positioning system (GPS) networks for measurement of surface displacement are expanding, seismic sensor sensitivity is increasing, synthetic aperture radar missions are planned to measure surface changes worldwide, and increasingly complex simulations are producing vast amounts of data. Automated techniques are necessary to assist in coping of the deluge of information. These techniques are useful in a number of ways: they can analyze quantities of data that would overwhelm human analysts, they can find subtle changes in the data that might evade a human expert, and they assist in objective decision making in cases where even experts disagree (for example, identifying aftershock sequences, or modes in GPS time series). These techniques are not expected to replace human analysis, but rather to be tools for human experts to use as part of the research cycle.

The field of geophysics poses particular challenges for automated analysis. The data is often noisy or of poor quality, due to the nature of the sensor equipment; for similar reasons it is also often sparse or incomplete. Furthermore, the underlying system is unobservable, highly complex, and still poorly understood by theory. Automated analysis is a useful tool only if it can satisfy several criteria. The results produced must be consistent across experiments on the same or similar data. Only minimal parameter tuning can be required, lest the results be considered the arbitrary result of parameter selection. And the method must be computationally tractable, so results can be returned to the user in reasonable time.

In this work, we investigate the use of hidden Markov models (HMMs) [1–5] as the basis for an automated tool for analysis of geophysical data. We begin by giving a brief overview of hidden Markov models and introducing our notation. We then present the standard method for solving for the optimal HMM parameters and discuss the inherent local maxima problem associated with the HMM optimization problem. In answer to this we introduce our modified robust HMM optimization method and present some preliminary results produced by this method.

2 Hidden Markov Models

A hidden Markov model (HMM) is a statistical model for ordered data. The observed data is assumed to have been generated by a unobservable statistical process of a particular form. This process is such that each observation is coincident with the system being in a particular state. Furthermore it is a first order Markov process: the next state is dependent only the current state. The model is completely described by the initial state probabilities, the first order Markov chain state-to-state transition probabilities, and the probability distributions of observable outputs associated with each state.

Our notation is as follows: a hidden Markov model λ with N states is composed of initial state probabilities $\pi = (\pi_1, \ldots \pi_N)$, state-to-state transition probabilities $A = (a_{11}, \ldots, a_{ij}, \ldots, a_{NN})$, and the observable output probability distributions $B = (b_1, \ldots, b_N)$. The observable outputs can be either discrete or continuous. In the discrete case, the output probability distributions are denoted by $b_i(m)$, where m is one of M discrete output symbols. In the continuous case, the output probability distributions are denoted by $b_i(y, \theta_{i1}, \ldots, \theta_{ij}, \ldots, \theta_{iM})$ where y is the real-valued observable output (scalar or vector) and the θ_{ij}s are the parameters describing the output probability distribution. For the normal distribution we have $b_i(y, \mu_i, \Sigma_i)$.
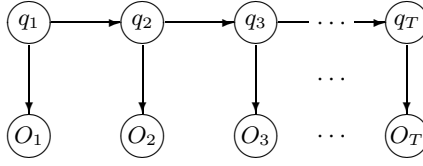
Partially observed Markov chain.

Fig. 1. A representation of the hidden Markov model, with hidden nodes in underlying system states q, and observable variables O.

3 HMM optimization problem

For the series of observations $O = O_1 O_2 \cdots O_T$, we consider the possible model state sequences $Q = q_1 q_2 \cdots q_T$ to which this series of observations could be assigned. For a given fixed state sequence Q, the probability of the observation sequence O is given by

$$P(O|Q, \lambda) = \prod_{t=1}^{T} P(O_t|q_t, \lambda). \tag{1}$$

Assuming statistical independence of observations,

$$P(O|Q, \lambda) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T). \tag{2}$$

The probability of the given state sequence Q is

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}. \tag{3}$$

The joint probability of O and Q is the product of the above, so that

$$P(O, Q|\lambda) = P(O|Q, \lambda) P(Q|\lambda), \tag{4}$$

and the probability of O given the model is obtained by summing this joint probability over all possible state sequences Q:

$$P(O|\lambda) = \sum_{\text{all } Q=q_1 q_2 \cdots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T). \tag{5}$$

Although other optimization criteria are possible, most commonly we wish to optimize the model parameters so as to maximize this likelihood $P(O|\lambda)$. We can pose this as non-convex, non-linear optimization problem with constraints on π, A, and B that reflect the fact that they are probabilities. Often this problem is presented as the equivalent problem of maximizing the *log likelihood*, $\log P(O|\lambda)$.

4 Expectation-Maximization

The most common optimization technique employed to solve this problem is the Expectation-Maximization (EM) algorithm [6]. We can pose the EM algorithm generally as follows: we wish to maximize a likelihood $P(\lambda)$ where λ is a set of model parameters. Given $p(x, \lambda)$, a positive real-valued function on $x \times \Lambda$ measurable in x for fixed λ with measure μ, we define

$$P(\lambda) = E[p(x, \lambda)|\lambda] = \int_X p(x, \lambda) d\mu(x) \tag{6}$$

and

$$Q(\lambda, \lambda') = E[\log p(x, \lambda')|\lambda] = \int_X p(x, \lambda) \log p(x, \lambda') d\mu(x). \tag{7}$$

Here x is the so-called *hidden variable*, while $p(x, \lambda)$ is often referred to as the *complete data likelihood*. The function Q is often referred to as the *Q-function*. Note that the function p may be a function of the observable outputs y as well as the parameters of the model λ, so $p = p(x, y, \lambda)$. In this case, the integrals are over $X \to Y(X)$.

It can be shown that for a transformation $\mathcal{F}$ that if $\mathcal{F}(\lambda)$ is a critical point of $Q(\lambda, \lambda')$ as a function of λ', then the fixed points of $\mathcal{F}$ are critical points of P. This gives us the EM algorithm:

1. Start with $k = 0$ and pick a starting $\lambda^{(k)}$.
2. Calculate $Q(\lambda^{(k)}, \lambda)$ (expectation step).
3. Maximize $Q(\lambda^{(k)}, \lambda)$ over λ (maximization step). This gives us the transformation $\mathcal{F}$.
4. Set $\lambda^{(k+1)} = \mathcal{F}(\lambda^{(k)})$. If $Q(\lambda^{(k+1)}, \lambda) - Q(\lambda^{(k)}, \lambda)$ is below some threshold, stop. Otherwise, go to step 2.

Note that this method is inherently sensitive to the initial conditions $\lambda^{(0)}$, and only guarantees eventual convergence to a local maxima of the objective function, not the global maximum. Nevertheless, it is widely used in practice and often achieves good results.

5   Optimization procedure for the HMM

For the hidden Markov model, we employ the EM method in following manner. We have

$$p(q, O, \lambda) = \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T), \tag{8}$$

with $P(\lambda) = E[p(q, O, \lambda)|\lambda]$ defined as in (5). If we let z be a set of state-indicator indicator vectors $z = (z_1, \ldots, z_T)$ such that $z_{it} = 1$ if $q_t = i$, $z_{it} = 0$ otherwise, then we can represent the complete data log likelihood as

$$\sum_{i=1}^{N} z_{i1} \log \pi_i + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} z_{it} z_{j,t+1} \log a_{ij} + \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} \log b_i(O_t). \tag{9}$$

From this we can calculate

$$Q(\lambda, \lambda^{(k)}) = \sum_{i=1}^{N} \tau_{i1}^{(k)} \log \pi_i + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij} + \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \log b_i(O_t) \tag{10}$$

where

$$\tau_{ijt} = P(Z_{it} = 1, Z_{j,t+1} = 1|O, \lambda) \; t = 1, \ldots, T-1, \tag{11}$$
$$\tau_{it} = \qquad P(Z_{it} = 1|O, \lambda) \qquad t = 1, \ldots, T, \tag{12}$$

and Z is a probabilistic component indicator variable analogous to z.

We wish to maximize $Q(\lambda, \lambda^{(k)})$ over λ. We can view Q as the sum of three separable components, $Q = Q_1 + Q_2 + Q_3$:

$$Q_1(\lambda, \lambda^{(k)}) = \sum_{i=1}^{N} \tau_{i1}^{(k)} \log \pi_i, \tag{13}$$

$$Q_2(\lambda, \lambda^{(k)}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij}, \tag{14}$$

$$Q_3(\lambda, \lambda^{(k)}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \log b_i(O_t). \tag{15}$$

Maximization of each component may be pursued separately. We have

$$\pi_i = \frac{\pi_i^{(k)} b_i^{(k)}(O_1)}{\sum_{j=1}^{N} \pi_j^{(k)} b_j^{(k)}(O_1)}, \tag{16}$$

as the maximizing solution for Q_1 and

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}^{(k)}}{\sum_{t=1}^{T-1} \tau_{it}^{(k)}}, \tag{17}$$

as the maximizing solution for Q_2. If the outputs of the model are discrete, the maximizing solution for Q_3 is

$$b_i(m) = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{\sum_{t=1}^{T} \tau_{it}^{(k)}} \tag{18}$$

where m is a possible output symbol. If the outputs of the model are continuous, then there is no general explicit formula for the maximum value of the output distribution parameters. However, for certain special forms of the output

distribution, the maximizing values can be calculated analytically. For example, in the case of multivariate Gaussian output distributions ($b_i(y) = n(\det(\Sigma_i))^{-1/2} \exp(-(y - \mu_i)^T \Sigma_i^{-1}(y - \mu_i)/2)$, where n is a normalizing factor), we have

$$\mu_i = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} O_t}{\sum_{t=1}^{T} \tau_{it}^{(k)}}, \tag{19}$$

and

$$\Sigma_i = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)}(O_t - \mu_i^{(k+1)})(O_t - \mu_i^{(k+1)})^T}{\sum_{t=1}^{T} \tau_{it}^{(k)}}. \tag{20}$$

What remains is to calculate the probabilities τ_{it} and τ_{ijt}. To do so, we make use of the lattice structure of the HMM to perform an iterative calculation, known as the *forward-backward* procedure. Consider the forward variable $\alpha_t(i)$ defined as

$$\alpha_t(i) = P(O_1 \cdots O_t, Z_{it} = 1|\lambda). \tag{21}$$

This is the probability of observing the partial sequence $O_1 \cdots O_t$ and that the system is in state i at time t, given the model λ. We can solve for $\alpha_t(i)$ inductively as follows:

1. Initialization:
$$\alpha_1(i) = \pi_i b_i(O_1), \ i = 1, \ldots, N. \tag{22}$$

2. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1}), \ t = 1, \ldots, T-1,$$

$$j = 1, \ldots, N. \tag{23}$$

This is an $O(N^2 T)$ computation. Note that it also gives us an efficient way to calculate the value of the objective function, since

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{24}$$

As the second part of the forward-backward procedure, we consider the backward variable $\beta_t(i)$ defined as

$$\beta_t(i) = P(O_{t+1} \cdots O_T | Z_{it} = 1, \lambda). \tag{25}$$

This is the probability of observing the partial sequence $O_{t+1} \cdots O_T$, given that the system is in state i at time t and the model λ. Once again we can solve for $\beta_t(i)$ inductively:

1. Initialization:
$$\beta_T(i) = 1, \ i = 1, \ldots, N. \tag{26}$$

2. Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \ t = T-1, \ldots, 1,$$

$$i = 1, \ldots, N. \tag{27}$$

This is also an $O(N^2 T)$ computation.

Now we can calculate the probabilities τ using the forward and backwards variables. For instance,

$$\tau_{it} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{28}$$

is the probability of being in state i at time t, given the observation sequence and the model. Note that we can use τ_{ti} to solve for the individually most likely state q_t at time t, as

$$q_t = \operatorname*{argmax}_{1 \leq i \leq N}(\tau_{it}), \ t = 1, \ldots, T. \tag{29}$$

We can also now calculate τ_{ijt}, the probability of being in state i in time t and state j at time $t+1$, given the model and the observation sequence. Using our definitions of the forward-backward variables, we can write

$$\tau_{ijt} = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}. \tag{30}$$

6 Multimodality of the HMM objective function

As previously noted, the EM algorithm only guarantees convergence to a local maximum. Since the algorithm is deterministic, the initial model parameter selection controls which local maxima is eventually reached. In many cases, the EM algorithm functions well; this is one reason for its popularity. However, the likelihood of an HMM has potentially an exponential number of local maxima; this makes the optimization problem much more difficult.

Consider a set of HMM parameters for which $\pi_i, a_{ij} \in \{0, 1\}$ for $i, j = 1, \ldots, N$. Let $Q^* = q_1^* \cdots q_T^*$ be the state sequence for some particular π^*, A^* chosen from this set. Then

$$\alpha(i) = \begin{cases} b_{q_1^*}(O_1) \cdots b_{q_t^*}(O_t) & \text{if } i = q_t^* \\ 0 & \text{otherwise} \end{cases}, \tag{31}$$

and

$$\beta(i) = \begin{cases} b_{q_T^*}(O_T) \cdots b_{q_{t+1}^*}(O_{t+1}) & \text{if } i = q_t^* \\ 0 & \text{otherwise} \end{cases}, \tag{32}$$

assuming that $b_{q_t^*}(O_t) > 0$ for all t. This implies that

$$\tau_{it} = \begin{cases} 1 & \text{if } i = q_t^* \\ 0 & \text{otherwise} \end{cases}, \tag{33}$$

and that

$$\tau_{ijt} = \begin{cases} 1 & \text{if } i = q_t^* \text{ and } j = q_{t+1}^* \\ 0 & \text{otherwise} \end{cases}. \tag{34}$$

From this we can derive the updates:

$$\pi_i^{(k+1)} = \begin{cases} 1 & \text{if } \pi_i^{(k)} = 1 \\ 0 & \text{otherwise} \end{cases},$$

$$a_{ij}^{(k+1)} = \begin{cases} 1 & \text{if } a_{ij}^{(k)} = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{35}$$

As such, this solution is a fixed point of the EM transformation $\mathcal{F}$, and therefore a critical point of the likelihood $P(O|\lambda)$. Since there are N^{N+1} different solutions of this form, there are also at least that many critical points of the likelihood function. The question is, how many of these critical points are local maxima?

Assume that for a given solution $\lambda^* = (\pi^*, A^*, B^*)$ and resulting state sequence Q^*, the output probability distributions B^* are such that b_i^* is an optimal estimator for the set of observable outputs $O_{q_t^*}$ such that $q_t^* = i$. If this set is empty, that is, there are no observable outputs associated with the distribution b_i, the let $b_i = b_j$ where b_j is an optimal estimator for some nonempty subset of observations. In the case where the solution is such that $q_1^* = \cdots = q_T^*$, we also therefore have $b_1 = \cdots = b_N$. This implies that for small perturbations of π, A, B designated $\epsilon_\pi, \epsilon_A, \epsilon_\theta$,

$$b_{q_1^*}(O_1) \cdots b_{q_T^*}(O_T) \geq b_{q_1}(O_1, \theta_{q_1} + \epsilon_{\theta q_1}) \cdots b_{q_T}(O_T, \theta_{q_T} + \epsilon_{\theta q_T}), \tag{36}$$

for any sequence Q. Since $\sum_{\text{all}Q} P(Q|\lambda) = 1$,

$$P(O|\lambda^*) = b_{q_1^*}(O_1) \cdots b_{q_T^*}(O_T) \geq \sum_{\text{all}Q} (\pi_{q_1} + \epsilon_{\pi q_1}) b_{q_1}(O_1, \theta_{q_1} + \epsilon_{\theta q_1})(a_{q_1 q_2} + \epsilon_{A q_1 q_2}) b_{q_2}(O_2, \theta_{q_2} + \epsilon_{\theta q_2})$$

$$\cdots (a_{q_{T-1} q_T} + \epsilon_{A q_{T-1} q_T}) b_{q_T}(O_T, \theta_{q_T} + \epsilon_{\theta q_T}). \tag{37}$$

So λ^* is a local maximum. However, λ^* is not a unique maximum but rather part of a locally maximum continuous region of fixed points of $\mathcal{F}$ for which $b_1 = \cdots = b_N$ are optimal estimators of the joint observations and π and A are unrestricted. To see this, consider

$$\alpha'_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} \text{ with } \alpha_1(j) = \pi_j, \tag{38}$$

$$\beta'_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} \text{ with } \beta_T(i) = 1. \tag{39}$$

For $b_1 = \cdots = b_N$ we have then

$$\tau_{it} = \frac{\alpha'_t(i)\beta'_t(i)}{\sum_{i=1}^N \alpha'_t(i)\beta'_t(i)}, \tag{40}$$

$$\tau_{ijt} = \frac{\alpha'_t(i)a_{ij}\beta'_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha'_t(i)a_{ij}\beta'_{t+1}(j)}. \tag{41}$$

We note that $\sum_{i=1}^N \alpha'_t(i) = 1$ and $\beta'_t(i) = 1$ and so

$$\tau_{it} = \alpha'_t(i) \tag{42}$$

$$\tau_{ijt} = a_{ij}\alpha'_t(i), \tag{43}$$

and therefore $\pi_i^{(k+1)} = \pi_i^{(k)}$ and $a_{ij}^{(k+1)} = a_{ij}^{(k)}$.

Suppose we wish to exclude all such degenerate solutions from our analysis. Then we can consider a particular data set, one composed of S distinct segments s, each starting at $t_1(s)$ and ending at $t_T(s)$. For each segment the outputs $O_s = O_{t_1(s)} \cdots O_{t_T(s)}$ are all a single unique value, m_s. For this data set, the local maxima are solutions in which the possible output values for each state are unique, so that if $b_i(m) \neq 0$, then $b_j(m) = 0$ for all $i \neq j$, and are contiguous in the time sequence. More specifically, let the N_{s_i} segments $s_i(1), \ldots, s_i(N_{s_i})$ be associated with the state i; that is, let $b_i(m_{s_i(k)}) \neq 0, k = 1, \ldots, N_{s_i}$. Furthermore, let $L_{s_i(k)}$ be the length of the segment $s_i(k)$. Then a locally maximum model λ^* is such that

$$\pi_i^* = \begin{cases} 1 \text{ if } O_1 = m_{s_i(k)} \text{ for some } k \\ 0 \text{ otherwise} \end{cases},$$

$$a_{ij}^* = \begin{cases} \frac{\sum_{k=1}^{N_{s_i}} L_{s_i(k)}-1}{\sum_{k=1}^{N_{s_i}} L_{s_i(k)}} & \text{if } i = j \\ \frac{1}{\sum_{k=1}^{N_{s_i}} L_{s_i(k)}} & \text{if } t_T(s_i(k)) + 1 = t_1(s_j(l)) \text{ for some } k, l \\ 0 & \text{otherwise} \end{cases},$$

$$b_i(m)^* = \begin{cases} \frac{L_{s_i(k)}}{\sum_{k=1}^{N_{s_i}} L_{s_i(k)}} & \text{if } m = m_{s_i(k)} \text{ for some } m_{s_i(k)} \\ 0 & \text{otherwise} \end{cases}. \tag{44}$$

We first present a simple illustrative example. Consider the sequence $O = 112233$ of length $T = 6$, on which we train a model of size $N = 2$. Consider

$$\lambda_1 = \left\{ \pi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, A = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, b_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, b_2 = \begin{pmatrix} 1/5 \\ 2/5 \\ 2/5 \end{pmatrix} \right\},$$

$$\lambda_2 = \left\{ \pi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, A = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}, b_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, b_2 = \begin{pmatrix} 0 \\ 1/2 \\ 1/2 \end{pmatrix} \right\},$$

$$\lambda_3 = \left\{ \pi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, A = \begin{pmatrix} 2/3 & 1/3 \\ 0 & 1 \end{pmatrix}, b_1 = \begin{pmatrix} 2/3 \\ 1/3 \\ 0 \end{pmatrix}, b_2 = \begin{pmatrix} 0 \\ 1/3 \\ 2/3 \end{pmatrix} \right\}.$$

Then $P(O|\lambda_1) = 0.00512, P(O|\lambda_2) = 0.015625, P(O|\lambda_3) = 0.01$, so λ_2 is a local maximum. A second local maximum exists for which $q_1 \cdots q_4 = 1, q_5 q_6 = 2$; a third maximum is one for which the entire sequence is in the same state.

Now we present the general case and demonstrate that λ^* of the form described in (44) is in fact a local maximum. For ease of notation, we assume without loss of generality that $t_1(s_{i+1}(1)) > t_T(s_i(1))$, that is, the segment labels increase monotonically with t. We furthermore define

$$\mathcal{L}_i = \sum_{k=1}^{N_{s_i}} L_{s_i(k)}. \tag{45}$$

Then we have

$$P(O|\lambda^*) = \left(\frac{\mathcal{L}_1 - 1}{\mathcal{L}_1}\right)^{\mathcal{L}_1 - 1} \left(\frac{1}{\mathcal{L}_1}\right) \cdots \left(\frac{\mathcal{L}_N - 1}{\mathcal{L}_N}\right)^{\mathcal{L}_N - 1} \cdot \prod_{k=1}^{N_{s_1}} \left(\frac{L_{s_1(k)}}{\mathcal{L}_1}\right)^{L_{s_1(k)}} \cdots \prod_{k=1}^{N_{s_N}} \left(\frac{L_{s_N(k)}}{\mathcal{L}_N}\right)^{L_{s_N(k)}} \tag{46}$$

Now consider a model λ which is slightly perturbed from λ^* so that

$$b_1(m_{s_2(1)}) = \frac{1}{\mathcal{L}_1 + 1},$$

$$b_1(m_{s_1(k)}) = \frac{L_{s_1(k)}}{\mathcal{L}_1 + 1}, \ \ k = 1, \ldots, N_{s_1}$$

$$b_2(m_{s_2(1)}) = \frac{L_{s_2(1)} - 1}{\mathcal{L}_2 - 1},$$

$$b_2(m_{s_2(k)}) = \frac{L_{s_2(k)}}{\mathcal{L}_2 - 1}, \ \ k = 2, \ldots, N_{s_2}, \tag{47}$$

and

$$a_{11} = \frac{\mathcal{L}_1}{\mathcal{L}_1 + 1}, \ a_{12} = \frac{1}{\mathcal{L}_1 + 1}, \tag{48}$$

$$a_{22} = \frac{\mathcal{L}_2 - 2}{\mathcal{L}_2 - 1}, \ a_{23} = \frac{1}{\mathcal{L}_2 - 1}. \tag{49}$$

In other words, this model λ corresponds to a state sequence Q such that $q_t = 1$ for $t = 1, \ldots, \mathcal{L}_1 + 1$. We have

$$P(O|\lambda) = \sum_{n=0}^{L_{s_2(1)} - 1} \left\{ \left(\frac{\mathcal{L}_1}{\mathcal{L}_1 + 1} \right)^{(\mathcal{L}_1 - 1 + n)} \left(\frac{1}{\mathcal{L}_1 + 1} \right) \left(\frac{\mathcal{L}_2 - 2}{\mathcal{L}_2 - 1} \right)^{(\mathcal{L}_2 - 1 - n)} \left(\frac{1}{\mathcal{L}_2 - 1} \right) \cdots \left(\frac{\mathcal{L}_N - 1}{\mathcal{L}_N} \right)^{(\mathcal{L}_N - 1)}.$$

$$\prod_{k=1}^{N_{s_1}} \left(\frac{L_{s_1(k)}}{\mathcal{L}_1 + 1} \right)^{L_{s_1(k)}} \left(\frac{1}{\mathcal{L}_1 + 1} \right)^{n} \left(\frac{L_{s_2(1)} - 1}{\mathcal{L}_2 - 1} \right)^{L_{s_2(1)} - n}.$$

$$\prod_{k=2}^{N_{s_2}} \left(\frac{L_{s_2(k)}}{\mathcal{L}_2 - 1} \right)^{L_{s_2(k)}} \prod_{k=1}^{N_{s_3}} \left(\frac{L_{s_3(k)}}{\mathcal{L}_3} \right)^{L_{s_3(k)}} \cdots \prod_{k=1}^{N_{s_N}} \left(\frac{L_{s_N(k)}}{\mathcal{L}_N} \right)^{L_{s_N(k)}} \right\}, \tag{50}$$

From which it is evident that $P(O|\lambda^*) > P(O|\lambda)$. A similar analysis follows for the model λ perturbed from λ^* corresponding to the state sequence Q such that $q_t = 1$ for $t = 1, \ldots, \mathcal{L}_1 - 1$. We can extend this analysis to all such models λ such that A and B are perturbed in a like manner from the segment boundaries from A^* and B^*, so that $P(O|\lambda^*) > P(O|\lambda)$. From this we can conclude that λ^* is in fact a local maximum.

We note that for S unique segments there are $\binom{S-1}{N-1}$ local maxima λ^* of this form utilizing all N states, since we choose $N - 1$ of the $S - 1$ possible transitions between segments as our state transition points. We further note that this same analysis holds true for all models for which less than the full number of states are utilized. So in total there are $\sum_{n=1}^{N} \binom{S-1}{n-1}$ local maxima for this data set and model size N. If $S \geq N$, then $\sum_{n=1}^{N} \binom{S-1}{n-1} \geq 2^{N-1}$, so the lower bound on the number of local maxima is exponential in the model size.

An additional problem arises for certain forms of the output distribution B. For these forms there are values of the parameters θ_{im} such that the likelihood achieves an unfavorable global maximum. By unfavorable, we mean that these globally maximum model parameters are less informative about the values of the hidden variables than models with merely local maxima. For example, in the case of Gaussian output probability distributions, the likelihood goes approaches infinity as the eigenvalues of the variances approach zero. We can identify $\sum_{n=1}^{N} \binom{N}{n} \sum_{d=1}^{D} \binom{D}{d}$ such unfavorable global maxima, where D is the dimension of the observations, since the likelihood will approach infinity if even one eigenvalue of the variance of a single state approaches zero. This implies that the number of such global maxima is exponential in both the number of states and in the dimension of the observable data.

7 Q-function penalty terms

The analysis of the previous section indicates that many fixed points of the EM transformation and sub-optimal local maxima are located in the model parameter space at predictable points: where $\pi_i, a_{ij} \in \{0, 1\}$ and $b_i = b_j$. It would therefore appear to be advantageous to augment to the standard optimization procedure so as to avoid these parts of the parameter space. One way to do this is to add penalty terms to the Q-function.

For instance, we can modify Q_1 and Q_2 by adding log barrier terms:

$$Q_1'(\lambda, \lambda^{(k)}) = \sum_{i=1}^{N} \tau_{i1}^{(k)} \log \pi_i + \omega_{Q_1} \sum_{i=1}^{N} \log \pi_i, \tag{51}$$

$$Q_2'(\lambda, \lambda^{(k)}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij} + \omega_{Q_2} \sum_{i=1}^{N} \sum_{i=1}^{N} a_{ij}, \tag{52}$$

where $\omega_{Q_1}, \omega_{Q_2} > 0$ are small weighting terms. Our update rules are then

$$\pi_i = \frac{\pi_i^{(k)} b_i^{(k)}(O_1) + \omega_{Q_1}}{\sum_{j=1}^{N} \pi_j^{(k)} b_j^{(k)}(O_1) + N\omega_{Q_1}}, \tag{53}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}^{(k)} + \omega_{Q_2}}{\sum_{t=1}^{T-1} \tau_{it}^{(k)} + N\omega_{Q_2}}, \tag{54}$$

which cannot lie in $\{0,1\}$.

No general penalty term exists to assist in avoiding the condition where $b_i = b_j$. However, for particular forms of the output distribution penalty terms can be devised. For example, for discrete output distributions, we can add a penalty term based on the inner product:

$$Q_3' = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \tau_{it}^{(k)} \delta(O_t - m) \log b_i(m) - \omega_{Q_3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{m=1}^{M} b_i(m) b_j(m) \tag{55}$$

where $\omega_{Q_3} > 0$ is a small weighting factor. As a second example, we consider the case of Gaussian output distributions. We add a penalty term based on the squared Euclidean distance:

$$Q_3' = \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \left(\log n - \frac{1}{2} \log \det(\Sigma_i) - \frac{1}{2}(m_i - \mu_i)^T \Sigma_i^{-1}(m_i - \mu_i) - \frac{1}{2}(O_t - m_i)^T \Sigma_i^{-1}(O_t - m_i) \right.$$
$$\left. + \frac{\omega_{Q_3}}{2} \sum_{j=1}^{N} (\mu_i - \mu_j)^T (\mu_i - \mu_j) \right). \tag{56}$$

In both these cases conditions on the weighting terms ω_{Q_3} can be found such that the function Q_3 remains concave and thus has a single local maxima. Computing the solution to either maximization problem requires an iterative procedure with a computational cost per iteration which is cubic in the dimension of the observations. As the optimization of the original cost function requires inversion of the covariance matrices at each EM iteration, the modified method merely introduces a constant factor for a bounded number of iterations in the inner loop. In practice, solutions to Q_3' can be found in very small (< 10) numbers of iterations, and good approximations in merely one or two.

We note in that these penalty terms do not help to escape from local maxima when the model parameters are already at a point where $b_i = b_j$. Although random initialization of the model parameters makes this unlikely, alternate initialization methods can make this more problematic. In such cases, one way to escape from the local maximum is to perturb the distributions by some small amount when the case $b_i = b_j$ is detected.

In the case of Gaussian output distributions we can impose an additional penalty term in order to deal with unfavorable global maxima located where the covariance matrices become singular. Our penalty term is based on the trace of the inverse of the covariance matrix, since

$$\mathrm{Tr}\, \Sigma_i^{-1} = \sum_{d=1}^{D} \frac{1}{\lambda_{id}} \tag{57}$$

where D is the dimension of the observations and $\lambda_{i1}, \ldots, \lambda_{iD}$ are the eigenvalues of the ith covariance matrix. The modified Q-function is

$$Q_3' = \sum_{t=1}^{T} \sum_{i=1}^{N} \tau_{it}^{(k)} \left(\log n + \frac{1}{2} \log \det(\Sigma_i^{-1}) - \frac{1}{2}(m_i - \mu_i)^T \Sigma_i^{-1}(m_i - \mu_i) - \frac{1}{2} \mathrm{Tr}\, \Sigma_i^{-1} S_i - \frac{\omega_\Sigma}{2} \mathrm{Tr}\, \Sigma_i^{-1} \right), \tag{58}$$

where ω_Σ is a weighting factor. This leads us to an optimum solution in which we add a diagonal matrix $\omega_\Sigma I$ to each covariance matrix.

Incorporating all of the above, our modified EM algorithm is then:

1. Start with $k = 0$ and pick a starting $\lambda^{(k)}$.
2. Calculate $Q'(\lambda^{(k)}, \lambda)$ (expectation step).
3. Maximize $Q'(\lambda^{(k)}, \lambda)$ over λ (maximization step). This gives us the transformation $\mathcal{F}$.
4. Set $\lambda^{(k+1)} = \mathcal{F}(\lambda^{(k)})$. If $Q'(\lambda^{(k+1)}, \lambda) - Q'(\lambda^{(k)}, \lambda)$ is below some threshold, stop. Otherwise, go to step 2.
5. Check to see if $b_i = b_j$ for any $i \neq j$. If so, then perturb the current model so that $\theta_i = \theta_i + \epsilon_\theta$, and go to step 2. Otherwise, stop.

8 Experimental Results

We applied our robust HMM method to GPS and seismicity data collected in the southern California region. In our implementation we assume Gaussian output probability distributions for both FMM and HMM for simplicity and ease of computation. Presented here are some preliminary experimental results.

The GPS data consists of surface displacement signals collected from a number of sites scattered around the southern California region. The data was three dimensional, consisting of east-west displacement, north-south displacement, and vertical displacement measurements, collected daily. Figure 2 shows a representative example of the results of the method applied to GPS data collected in the city of Claremont, California. The method determined that a five state model was optimal for this data set. Using a five state model, the HMM was able to separate the data into distinct classes that correspond to physical events. These classes are indicated in the figure by different shades and vertical lines. There is one instance of class 2 in the midst of class 3, corresponding to sharp north-south and vertical movements at that time sample, but otherwise the classes are sequential. The states before and after the Hector Mine quake of October 1999 are clearly separated, and distinct in turn from a period in 1998 in which well ground water drainage caused displacement in the vertical direction. Sharp movements in the north-south direction (as yet unattributed) were also isolated as a separate class.
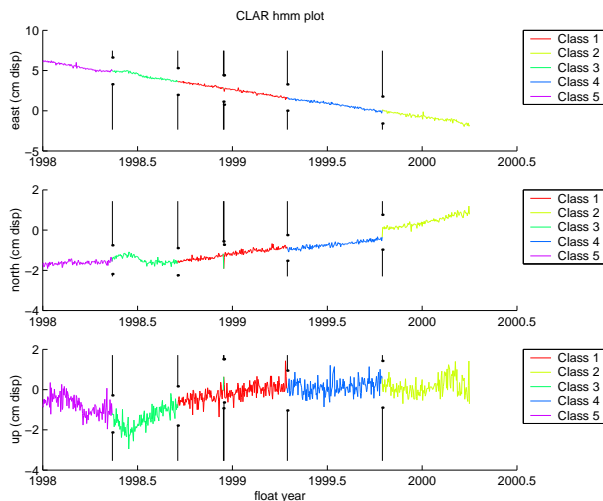

Fig. 2. HMM analysis results of global positioning system (GPS) relative displacement data collected from a receiver located in Claremont, California. Classes associated with different regimes are indicated by line coloration and vertical indicator lines.

The seismicity data was taken from the Southern California Earthquake Center (SCEC) catalog. For this experiment, the original data set was processed to produce six components for each observed seismic event between January 1st, 1960 and December 31st, 1999: latitude, longitude, depth, magnitude, time to next event, and time to previous event. Events of less than magnitude four were removed. The method determined that a model with 17 states would be optimal for this data sets. The data was grouped into scientifically meaningful classes, including clusters of aftershocks for the Hector Mine, Landers, and Northridge earthquakes, Transverse Range events, and swarm events in the Salten Sea area. Furthermore, relationships between the classes as indicated by the transition probabilities reveal evidence of scientifically meaningful phenomenon such as stress waves. Figure 3 show examples of the classifications produced by the method. Circles indicate the location of earthquakes; circle size corresponds to magnitude. Lines represent the major faults.

9 Conclusions and Future Work

We have presented a tool for geophysical data analysis that is based around the use of hidden Markov models (HMMs). The tool employs a method for estimating the optimal HMM parameters that is based on the analytical analysis of certain local maxima of the HMM objective function that originate in the model structure itself rather than the data. This analysis is then used to modify the standard optimization procedure through the application of penalty functions which enable the solution to avoid many local maxima. This improves both the quality and consistency
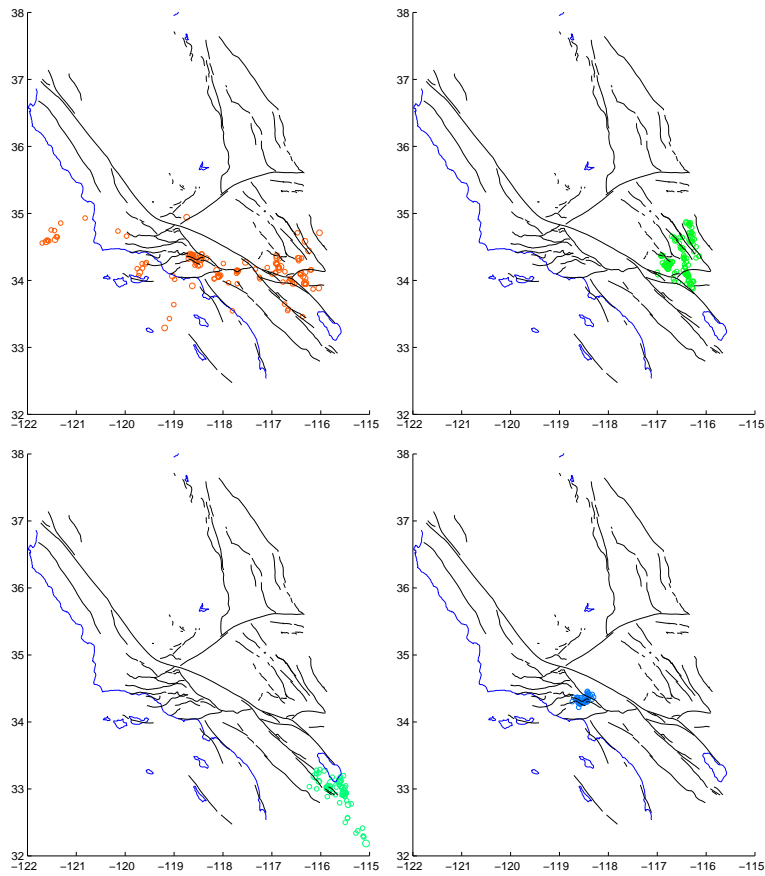
Fig. 3. HMM analysis result for SCEC catalog seismicity data. Upper left: the class of Transverse Range events; upper right: the class of Hector Mine and Landers earthquake aftershocks; bottom left: the class of Salten Sea swarm events; bottom right: the class of Northridge earthquake aftershocks.

of results. Preliminary experiments employing this method in the analysis of geodetic and seismic record data have yielded scientifically meaningful results.

As part of our continued work on this method we are performing large-scale systematic analysis of the effect of the modifications on the final solution. In addition, we are applying the method to a more diverse assortment of geophysical data sets.

References

1. L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.
2. L. E. Baum and J. A. Egon. An inequality with applications to statistical estimation for probabilistic functions of a markov process and to a model for ecology. *Bull Amer Meteorol Soc*, 73:360–363, 1967.
3. L. E. Baum and T. Petric. Statistical inference for probabilistic functions of finite state markov chains. *Ann Math Stat*, 37:1554–1563, 1966.
4. L. E. Baum, T. Petrie, G. Soules, and H. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *Ann Math Soc*, 41(1):164–171, 1970.
5. L. E. Baum and G. R. Sell. Growth functions for transformations on manifolds. *Pac J Math*, 27(2):211–227, 1968.
6. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorith. *J Roy Stat Soc*, 39(1):1–38, 1977.