

Statistical Analysis of Geodetic Networks for Detecting Regional Events

Robert Granat¹

¹ Jet Propulsion Laboratory , 4800 Oak Grove Dr. , Pasadena, CA
91109 , *granat@aig.jpl.nasa.gov*

ABSTRACT

We present an application of hidden Markov models (HMMs) to analysis of geodetic time series in Southern California. Our model fitting method uses a regularized version of the deterministic annealing expectation-maximization algorithm to ensure that model solutions are both robust and of high quality. Using the fitted models, we segment the daily displacement time series collected by 127 stations of the Southern California Integrated Geodetic Network (SCIGN) over a two year period. Segmentations of the series are based on statistical changes as identified by the trained HMMs. We look for correlations in state changes across multiple stations that indicate region-wide activity. We find that although in one case a strong seismic event was associated with a spike in station correlations, in all other cases in the study time period strong correlations were not associated with any seismic event. This indicates that the method was able to identify more subtle signals associated with aseismic events or long-range interactions between smaller events.

INTRODUCTION

In this work, we apply hidden Markov models (HMMs) to analysis of geodetic time series data. Hidden Markov models are a well known tool that have been

successfully applied to a number of problems. The HMM works by modeling the observations as being generated by a discrete sequence of underlying (hidden) states with Markovian properties. Changes in the statistics of the observation sequence are indicative of changes in the underlying state. Fitting a model to the observation data results in an estimate of the underlying state sequence, allowing classification of observations according to associated state, as well as an estimate of the model statistics.

Hidden Markov models have been used most prominently in the fields of speech synthesis and recognition (continuous output HMMs), and protein matching and analysis (discrete output HMMs). In these domains the difficult non-linear optimization problem of fitting the model to the observation data has primarily been addressed by the addition of explicit and implicit constraints that act to reduce the number of free model parameters. These methods include restrictions on the form of the state-to-state transition probability matrix (Juang & Rabiner 1985, Farago & Lugosi 1989, McGuire, et al. 2000), restrictions on the form of the output distributions (Ephraim, et al. 1989), and parameter tying (Bellegarda & Nahamoo 1990, Young & Woodland 1994, Bocchieri & Mak 2001). These constraints are supported by extensive knowledge about the underlying system being modeled. For instance, in speech analysis, we know not only the rules of language that govern the ordering of sounds and words (Lee & Hon 1989, Lee 1990) but also the details of the actual physical process which generates sound waves (Juang & Rabiner 1991).

However, in the analysis of geodetic time series such constraints are not as readily available. To address this difficulty, we use an alternative optimization approach that uses deterministic annealing and statistics-based regularization. The usual approach to HMM optimization is to use the expectation-maximization (EM) algorithm. The EM algorithm has many good properties and works well in practice for many problems, but only guarantees convergence to a local maxima (the local maxima found is dependent on the initial conditions). When for a given problem the number of local maxima is large, repeated applications of the method can result in very different results. The annealing portion of our method addresses this problem by guiding the solution towards strong local maxima that emerge first from the optimization surface as the problem is “cooled.” The regularization portion of our work acts in a complimentary fashion, by pushing the solution away from solutions with redundant states. The method can be demonstrated to work by empirically measuring the number of local maxima encountered for different

random initializations. A comparison with the standard EM algorithm reveals that the number of local maxima found is dramatically reduced with no loss in solution quality.

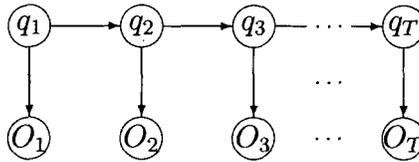
We apply this method to time series of displacement measurements collected by the SCIGN (Southern California Integrated Geodetic Network) array of GPS stations. By training an HMM on an individual time series, we can classify the displacement observations in the series according to the estimate of the underlying state. This allows us to objectively segment the series into different behavioral modes. By repeating this process for all stations in the network, we can get a picture of the overall activity in the region. In particular, we can detect regional events by looking for correlations between state changes at different stations at a given point in time. We present the results of this method used in a two year study of GPS measurements and compare it against the seismic record during that same period.

HIDDEN MARKOV MODELS

We begin with a short review of hidden Markov models (HMMs). A hidden Markov model 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.

Notation

Our notation is similar to that employed by Rabiner (Rabiner 1989) and is as follows: a hidden Markov model λ with N states is composed of a vector of initial state probabilities $\pi = (\pi_1, \dots, \pi_N)$, a matrix of state-to-state transition probabilities $A = (a_{11}, \dots, a_{ij}, \dots, a_{NN})$, and the observable output probability distributions $B = (b_1, \dots, b_N)$. 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



Partially observed Markov chain.

Figure 3.1: A representation of the hidden Markov model, with hidden nodes in underlying system states q , and observable variables O .

probability distributions are denoted by $b_i(y, \theta_{i1}, \dots, \theta_{ij}, \dots, \theta_{iM})$ 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 $b_i(y, \mu_i, \Sigma_i)$. An observation sequence O of length T is denoted $O_1 O_2 \dots O_T$ and a state sequence Q of the model is denoted $q_1 q_2 \dots q_T$.

Model optimization problem

In this Section we concentrate on maximizing the likelihood of the observation sequence given the model, $P(O|\lambda)$; this is the *maximum likelihood* objective function. However, many other objective functions have been proposed for hidden Markov models, including the state-optimized joint likelihood for the observations and underlying state sequence (Juang & Rabiner 1990), maximum mutual information (MMI), (Bahl, et al. 1986) minimum discrimination information (MDI) (Ephraim et al. 1989), and maximum classification error (MCE) (Chou, et al. 1994). Of these, all but the first require labeled training examples on which to train the models, making them inappropriate for our target application domain. The first, used as the basis for the so-called “segmental K -means” algorithm, suffers from similar initialization-dependent local maxima issues as does the more common maximum likelihood criteria, and so we skip an independent analysis of it in this work.

For the series of observations $O = O_1 O_2 \dots O_T$, we consider the possible model state sequences $Q = q_1 q_2 \dots q_T$ 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

$$P(O|Q, \lambda) = \prod_{t=1}^T P(O_t|q_t, \lambda). \quad (3.1)$$

Assuming statistical independence of observations,

$$P(O|Q, \lambda) = b_{q_1}(O_1)b_{q_2}(O_2) \cdots b_{q_T}(O_T). \quad (3.2)$$

The probability of the given state sequence Q is

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}. \quad (3.3)$$

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

$$P(O, Q|\lambda) = P(O|Q, \lambda)P(Q|\lambda), \quad (3.4)$$

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

$$P(O|\lambda) = \sum_{\text{all } Q=q_1 q_2 \cdots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T). \quad (3.5)$$

We can pose the optimization of $P(O|\lambda)$ as a non-convex optimization problem, often presented in terms of the equivalent problem of maximizing the *log likelihood* $\log P(O|\lambda)$. The most common method for solving this problem is the expectation-maximization (EM) algorithm (Dempster, et al. 1977), although alternative approaches exist, such as those employing genetic algorithms (Kwong, et al. 2001) recursive predictive error techniques (Collings, et al. 1994), or gradient projection (Huo & Chan 1993).

EXPECTATION-MAXIMIZATION

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

$$P(\lambda) = E[p(x, \lambda)|\lambda] = \int_{\mathcal{X}} p(x, \lambda) d\mu(x) \quad (4.1)$$

and

$$Q(\lambda, \lambda') = E[\log p(x, \lambda') | \lambda] = \int_{\mathcal{X}} p(x, \lambda) \log p(x, \lambda') d\mu(x), \quad (4.2)$$

where λ' is also a set of model parameters on Λ . 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 observable outputs y as well as the parameters of the model λ , so we have $p(x, y, \lambda)$. In this case, the integrals are over $\mathcal{X} \rightarrow \mathcal{Y}(\mathcal{X})$.

Assume $Q(\lambda, \bar{\lambda}) \geq Q(\lambda, \lambda)$ for some set of model parameters $\bar{\lambda}$. Then $P(\bar{\lambda}) \geq P(\lambda)$. Proof:

$$\begin{aligned} \log P(\bar{\lambda})/P(\lambda) &= \log \int_{\mathcal{X}} p(x, \bar{\lambda}) d\mu(x) / P(\lambda) \\ &= \log \int_{\mathcal{X}} [p(x, \lambda) d\mu(x) / P(\lambda)] p(x, \bar{\lambda}) / p(x, \lambda) \\ &\geq \int_{\mathcal{X}} [p(x, \lambda) d\mu(x) / P(\lambda)] \log [p(x, \bar{\lambda}) / p(x, \lambda)] \\ &= (P(\lambda))^{-1} [Q(\lambda, \bar{\lambda}) - Q(\lambda, \lambda)] \geq 0. \end{aligned}$$

From this we can show 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.

For the hidden Markov model, we have the complete data likelihood

$$p(Q, O, \lambda) = \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T), \quad (4.3)$$

with $P(\lambda) = E[p(q, O, \lambda) | \lambda]$ defined as in (3.5). The forward-backward method suggested by Baum and colleagues (Baum 1972, Baum & Egon 1967, Baum & Petrie 1966, Baum, et al. 1970, Baum & Sell 1968) can be used to efficiently calculate each step in the EM algorithm.

DETERMINISTIC ANNEALING

Deterministic annealing is a technique based on the principles of statistical mechanics that can be used to modify the EM method to mitigate its inherent sensitivity to initial conditions. Deterministic annealing uses the principle of maximum entropy to specify an alternative posterior probability density for the hidden variables; this allows us to define a new effective cost function depending on the temperature that is analogous to the thermodynamic free energy. Maximization of the likelihood at a given temperature is achieved via minimization of this cost function. Deterministic annealing differs from simulated annealing (Kirkpatrick, et al. 1983), in which a stochastic search is performed on the energy surface, in that the cost function is deterministically optimized at each temperature.

Use of deterministic annealing has been proposed for vector quantization (Rose, et al. 1992) and for clustering problems (Buhmann & Kuhnel 1993, Wong 1993). Yuille and colleagues (Yuille, et al. 1994) showed that the EM algorithm can be used in conjunction with deterministic annealing. Recently the deterministic annealing technique has been applied to a variety of problems (Rose 1998). The particular framework we present here was first applied by Ueda and Nakano to mixture density estimation problems (Ueda & Nakano 1994) and then extended to the general case (Ueda & Nakano 1998), and involves a reformulation of the EM algorithm so that it incorporates deterministic annealing.

How does the annealing process help in avoidance of local maxima? In effect, the method involves optimizing over a series of smoothed approximations to the original objective function. By slowly increasing the computational temperature parameter γ , the effect of each observation is gradually localized. At $\gamma = 1$, the parameterized Q -function is equivalent to the original Q -function for the problem.

We start the algorithm at a γ_{min} such that the modified objective function has a single maximum in λ . We thereafter assume that at each new γ , the global maximum of the new objective function is close to that of the previous, so that the method tracks the global maximum as γ increases. In cases where this assumption does not hold true, the method will fail to track the global optimum.

Our application of the deterministic annealing method to HMM optimization was is similar to that presented by Rose and Rao (Rose & Rao 2001) but differs from it in some important respects. First, it is not a supervised training method, and optimizes the likelihood rather than the minimum classification error. Second, it employs EM rather than gradient descent at each temperature. Our method is described in full in (Granat & Donnellan 2001) but can be summarized as follows: on the k th iteration at each temperature we optimize over the function

$$U(\gamma, \lambda | \lambda^{(k)}) = \sum_{i=1}^N \tau_{i1}^{(k)}(\gamma) \log \pi_i + \sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}^{(k)}(\gamma) \log a_{ij} + \sum_{i=1}^N \sum_{t=1}^T \tau_{it}^{(k)}(\gamma) \log b_i(O_t). \quad (5.1)$$

REGULARIZATION

As was observed by (Whiley & Titterington 2002), many local maxima of the deterministic annealing EM method are located where the states are underutilized, in other words where $b_i = b_j$. Our approach is design regularization terms that act to push the optimization procedure away from these parts of the parameter space.

In general when applying regularization terms it is convenient to work directly with the so-called Q -function for the HMM which is maximized during each EM iteration:

$$Q(\lambda, \lambda^{(k)}) = \sum_{i=1}^N \tau_{i1}^{(k)} \log \pi_i + \sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij} + \sum_{i=1}^N \sum_{t=1}^T \tau_{it}^{(k)} \log b_i(O_t). \quad (6.1)$$

Since this is separable in π , A , and B , we can divide this into the sum of three functions, $Q_1(\pi)$, $Q_2(A)$, and $Q_3(B)$ which can be individually maximized. Since

we are interested in the output distributions we concentrate our attention on the last of these.

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

$$Q'_3 = \sum_{i=1}^N \sum_{t=1}^T \tau_{it}^{(k)} \left(\log n - \frac{1}{2} \log \det(\Sigma_i) - \frac{1}{2} (m_i - \mu_i)^T \Sigma_i^{-1} (m_i - \mu_i) - \frac{1}{2} (O_t - m_i)^T \Sigma_i^{-1} (O_t - m_i) + \frac{\omega_{Q_3}}{2} \sum_{j=1}^N (\mu_i - \mu_j)^T (\mu_i - \mu_j) \right). \quad (6.2)$$

To find the means and covariances we solve the simultaneous equations

$$\left(N\omega_{Q_3} I_{N^2 \times N^2} + \omega_{Q_3} \begin{bmatrix} I_{N \times N} & \cdots & I_{N \times N} \\ \vdots & \ddots & \vdots \\ I_{N \times N} & \cdots & I_{N \times N} \end{bmatrix} \right) U = \begin{bmatrix} \Sigma_1^{-1} & & \\ & \ddots & \\ & & \Sigma_N^{-1} \end{bmatrix} M, \quad (6.3)$$

and

$$\Sigma_i = \frac{\sum_{t=1}^T \tau_{it}^{(k)} (O_t - \mu_i)(O_t - \mu_i)^T}{\sum_{t=1}^T \tau_{it}^{(k)}}. \quad (6.4)$$

We note that for sufficiently large values of ω_{Q_3} equation 6.2 is no longer concave in the means and covariances, thereby invalidating the M-step of the EM algorithm. To address this, we require that $\omega_{Q_3} \leq \|\Sigma_i^{-1}\|/2N$ at each iteration; this guarantees concavity of the modified Q -function.

SINGLE STATION RESULTS

In this Section, we present some results of using the combined deterministic annealing and regularization techniques to train hidden Markov models on a GPS time series collected by a station in Claremont, California.

This data set, which we designate `clar`, consists of relative displacement measurements in three dimensions (north-south, east-west, and vertical) collected

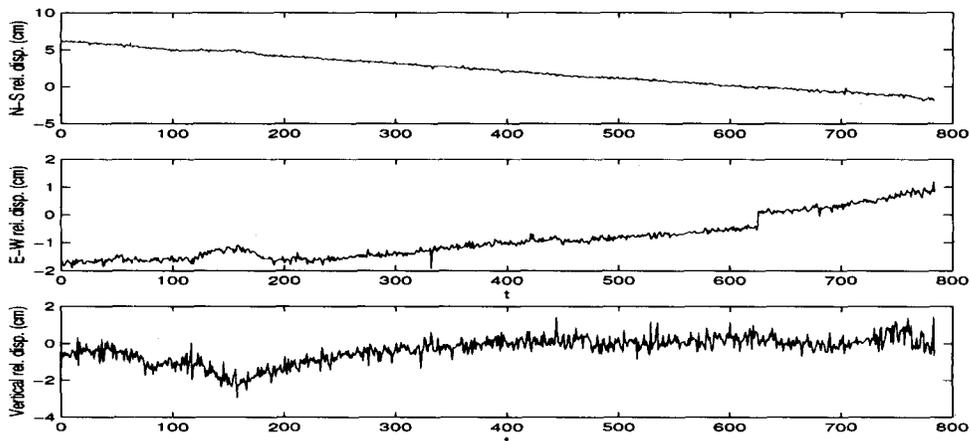


Figure 7.1: The data set `clar`, collected by a SCIGN GPS station in Claremont, California.

daily over about two years spanning 1998-1999. We choose this particular data set because it contains certain clear signals of deformation processes which have been identified by scientists, thereby providing some measure of ground truth against which we can evaluate models fit to this data. The figure 7.1 shows this data set; note the slow, recovering displacement around days 100-200 and the sudden east-west jump on day 626. The former is the result of ground water pumping and subsequent refilling of a local aquifer, the latter is an effect of the 1999 Hector Mine earthquake (magnitude 7.1).

Our regularization scheme was based on the squared Euclidean distance as described in the preceding Section. Instead of choosing a particular set value for ω_{Q_3} we set the value to the upper bound throughout the optimization procedure. That is, at each iteration $\omega_{Q_3} = \min_i \|\Sigma_i^{-1}\|/2N$. We note that because of this recalculation of the regularization weight, our procedure is in fact not a true EM optimization method. However, our implementation does require that the log likelihood function decrease at every iteration and so our procedure satisfies the requirements of a generalized expectation-maximization (GEM) method, guaranteeing convergence to a local maxima.

To evaluate the effectiveness of our method in avoid local maxima, we need to develop a metric for determining model solution distance. Our approach is to use the Hamming distance between the individually most likely state assignments for

the observation sequence (i.e., the classification results):

$$q_t = \operatorname{argmax}_{1 \leq i \leq N}(\tau_{it}), \quad t = 1, \dots, T \quad (7.1)$$

where τ_{it} is an estimate of the probability of being in state i at time t , given the observation sequence and the model. We use a linear assignment method based on bipartite graph matching (Ford & Fulkerson 1956) to resolve equivalent state permutations. Using this metric, we consider solutions with distance greater than zero to be different maxima. This means that models that produce identical classification sequences are considered to be the same local maxima, even if the model parameters are not identical. To determine the number of maxima found by an algorithm when applied to a particular data set, we can run repeated trials with uniform random initializations of the model parameters and count the number of different solutions based on this criterion. While this method does not guarantee identification of all local maxima, we can have confidence in the results if after some number of tests the number of identified local maxima fails to increase.

In figure 7.2 we present results of the method as applied to the data set `clar`. We note that the combined method has fewer local maxima than both the standard EM approach and the deterministic annealing alone for all three annealing schedules. In fact, we observe that there is only a single solution for the combined method at the slowest annealing schedule for $N = 1, \dots, 6$ and only two solutions for $N = 7$. However, after this point there is an abrupt rise in the number of experimentally determined local maxima. We propose that this rise is due to the fact that we have exceeded the true number of classes in the data set: since the combined method acts to reduce the number of redundant maxima, if we exceed the true number of maxima in the data set, then we expect radically worse results as the method forces the existence of additional, distinct classes. Figure 7.3 displays a classification result of the combined method for $N = 7$. We see that the method has identified all the major modes of the system including not only the before and after Hector Mine earthquake states and the water pumping signal but also a number of more subtle signals.

MULTIPLE STATION RESULTS

In the preceding Section we presented results of our method applied to a single SCIGN GPS station in Claremont, California. Here we are interested in detecting

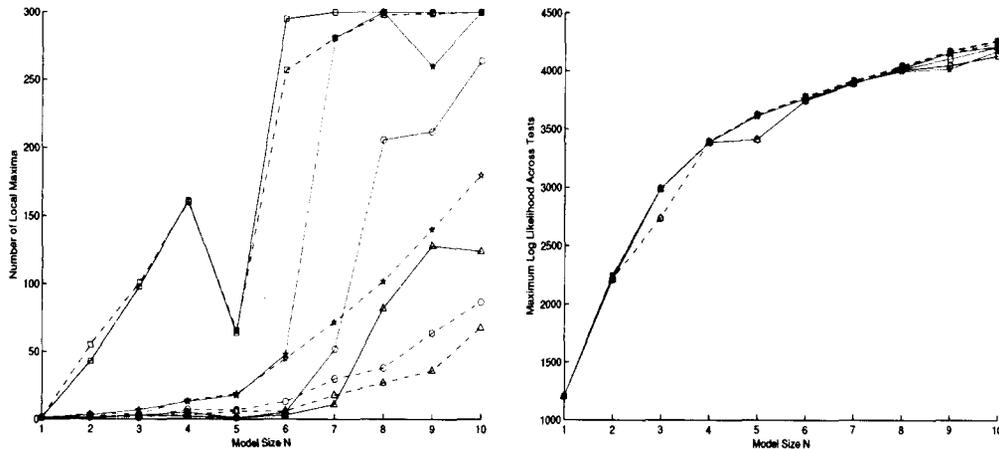


Figure 7.2: Left: Number of experimentally determined local maxima for HMMs with varying numbers of hidden states applied to the data set `clar`. Right: Maximum log likelihood among all experiments for HMMs with varying numbers of hidden states applied to the data set `clar`. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles results with schedule $\Delta\beta = 0.01$; red triangles results with schedule $\Delta\beta = 0.001$. Dashed lines are the results with deterministic annealing only, solid lines are the results of the combined annealing and regularization technique.

geophysical events with geographically disperse signatures and therefore wish to use the entire network. As background to our study we note that while earthquake events are of course of considerable interest, recently the geophysics community has become interested in aseismic events linked to crustal block motion or stress transfer between earthquake faults. These types of events have been observed in a few instances (Melbourne & Webb 2003, Rogers & Dragert 2003, Melbourne & Webb 2002, Melbourne, et al. 2002, Miller, et al. 2002, Hirose, et al. 1999, Heki, et al. 1997), but detections remain rare due to the subtlety of the signals. We hope to observe evidence of not only seismic but also aseismic events in the SCIGN data.

To do this, we extract GPS signals from all 127 available stations in a 820 day window. When GPS displacement values for a given station are not available on a particular day due to signal dropout or incomplete installation, we assume a zero displacement measurement for that day at that station. We note that since actual measurements are almost never of zero displacement, this in effect adds an additional “dropout” class to the data. Our next step is to train N -state hidden

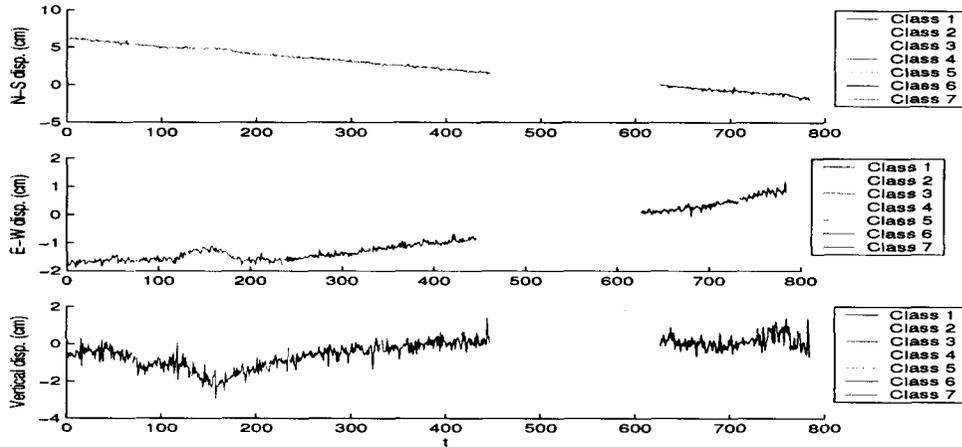


Figure 7.3: Classification results for a seven-state HMM applied to the data set `clar`.

Markov models on each of these GPS signals. Since the GPS signals have similar statistics to one another, we can use the results of our experiments on the data set `clar` to estimate the model size. We see that there was a single local maxima for $N < 7$ and two local maxima for $N = 7$, rising rapidly after that. So we can guess that a good number of states to use would be in the range of 5-7, with an additional state added to account for the dropout class. Once all models of a particular size have been trained on each of the GPS time series, we can use the models to perform state assignments of each observation. We suspect that interesting geophysical events will manifest themselves as changes in the signals across multiple GPS stations, so we look for correlations in state changes across the network.

Figure 8.1 shows the number of coincident state changes across all observation days with training done with six-state models. We see that there are a number of strong peaks indicating correlated state changes. Of note is the strong peak on day 652, which corresponds to the Hector Mine earthquake visible as an E-W displacement jump in the `clar` data. We also observe that there is an increasing trend in the average number of coincident state transitions; this is because of the increasing number of stations coming on line during the observation period. In figure 8.2 we compare the results of using the baseline EM algorithm (blue) for training the HMMs used in this study against the results of using the regularized deterministic annealing EM training (red). We see that the noise level in the

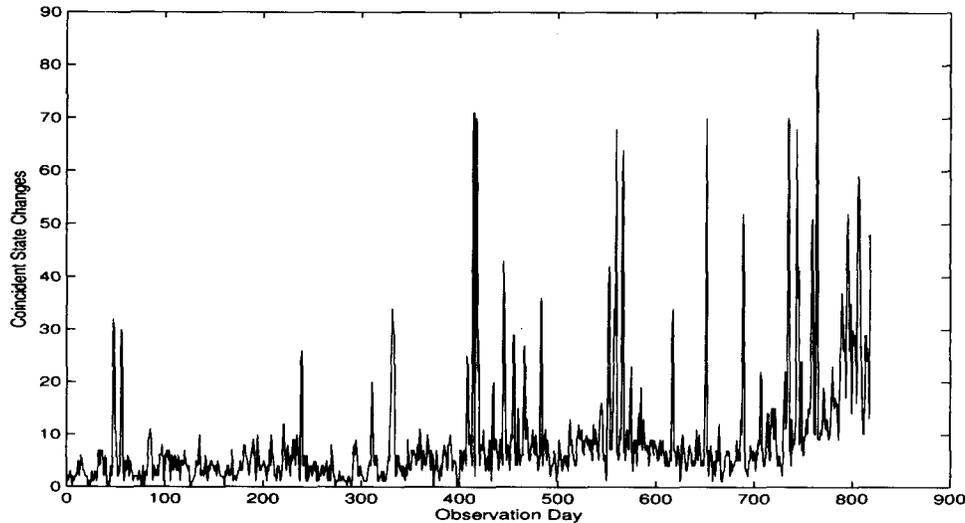


Figure 8.1: Coincident state changes for six-state HMMs trained on signals from each of 127 SCIGN GPS stations.

coincident state transition signal is significantly reduced by employing the latter method. We compare the coincident state changes against the earthquake record during the same time period in figure 8.3. We see that correlations across the GPS network (blue) are only strongly correlated with an earthquake event (red) in the case of the aforementioned Hector Mine earthquake. There are no other strong earthquakes in the time window studied. The implication of this is that the regional activity indicated by the state transition correlations is either an aseismic effect or the result of subtle long-range interactions between small (magnitude ≤ 4.0) events.

CONCLUSIONS

We have presented a method that uses deterministic annealing and regularization to modify the standard expectation-maximization (EM) method for fitting hidden Markov models (HMMs). We show that for typical geodetic time series the method greatly improves the robustness of the solution, as measured by the propensity to converge to different local maxima given random initial conditions, while still preserving solution quality as indicated by the solution log likelihood measure and by comparison with ground truth as identified by domain experts.

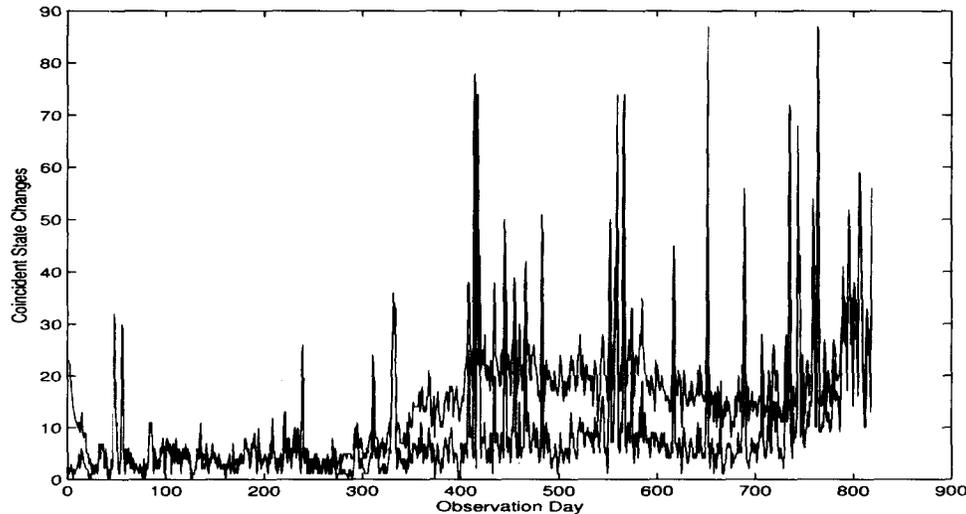


Figure 8.2: Coincident state changes for six-state HMMs trained using standard EM (blue) and regularized deterministic annealing EM (red) on signals from each of 127 SCIGN GPS stations.

Application of this method to training HMMs on displacement time series data collected by the Southern California Integrated Geodetic Network (SCIGN) enabled the reliable statistical segmentation of 127 of those time series over approximately a two year span. Comparing the timing of state changes across all of the stations, it was found that large correlations between multiple stations were found at particular points in time. In only one case, that of the Hector Mine earthquake, were these correlations found to be connected with a seismic event. The implication, therefore, is that these correlations are indicative of aseismic activity or of more subtle interactions between smaller scale events.

Further study will involve extending the analysis to longer time series and inclusion of data collected by all of the more than 250 SCIGN GPS stations. In addition, correlation spikes will be analyzed in more detail to determine if the correlated stations can be associated with any particular crustal block motion or particular fault interactions. Extensions of the method include use of techniques such as generalized conjugate-gradient acceleration (Jamshidian & Jennrich 1993) to speed up solution convergence, particularly in flat portions of the objective function, as well as combination with Kalman filter type approaches to estimate continuous state trajectory dynamics.

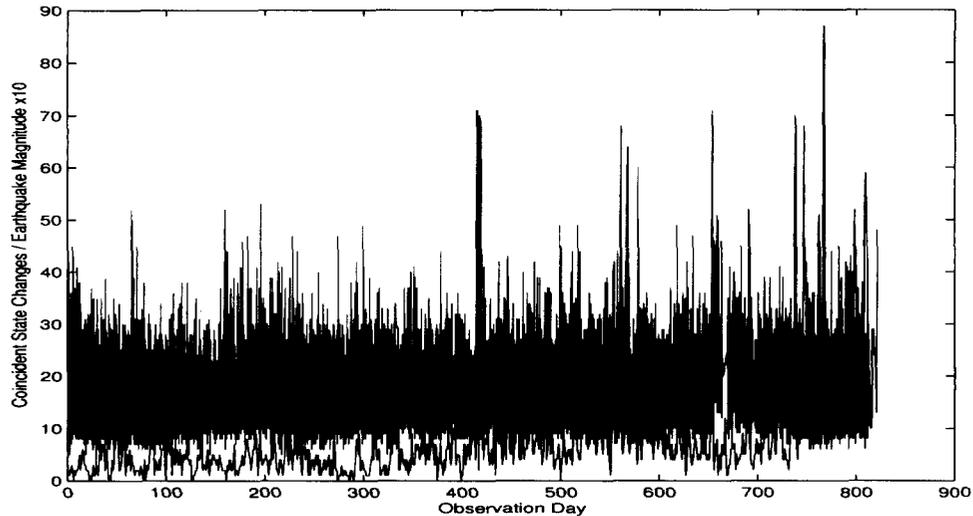


Figure 8.3: Comparison of coincident state changes for six-state HMMs trained using the regularized deterministic annealing EM (blue) and the Southern California earthquake record (red). Earthquake magnitudes, exaggerated by a factor of 10 for visibility, are presented on the vertical axis.

ACKNOWLEDGEMENTS

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

References

- R. Bahl, et al. (1986). 'Maximum mutual information estimation of hidden Markov model parameters for speech recognition'. In *Proc. 1986 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 49–52, Tokyo.
- L. E. Baum (1972). 'An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes'. *Inequalities* 3:1–8.
- L. E. Baum & J. A. Egon (1967). 'An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology'. *Bull. Amer. Math. Soc.* 73:360–363.

- L. E. Baum & T. Petrie (1966). 'Statistical inference for probabilistic functions of finite state Markov Chains'. *Ann. Math. Stat.* **37**:1554–1563.
- L. E. Baum, et al. (1970). 'A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains'. *Ann. Math. Stat.* **41**(1):164–171.
- L. E. Baum & G. R. Sell (1968). 'Growth functions for transformations on manifolds'. *Pac J Math* **27**(2):211–227.
- J. Bellegarda & L. Nahamoo (1990). 'Tied mixture continuous parameter modeling for speech recognition'. *IEEE Trans. on Acoustics, Speech, and Signal Proc.* **38**(12):2033–2045.
- E. Bocchieri & B. Mak (2001). 'Subspace distribution clustering hidden Markov model'. *IEEE Trans. on Speech and Audio Proc.* **9**(3):264–275.
- J. Buhmann & H. Kuhnel (1993). 'Complexity optimized data clustering by competitive neural networks'. *Neural Computation* **5**:75–88.
- W. Chou, et al. (1994). 'A minimum error rate pattern recognition approach to speech recognition'. *Int. J. Pattern Recogn. Artificial Intelligence* **8**(1):5–31.
- I. Collings, et al. (1994). 'Online identification of hidden Markov models via recursive prediction error techniques'. *IEEE Transactions On Signal Processing* **42**(12):3535–3539.
- A. D. Dempster, et al. (1977). 'Maximum Likelihood from Incomplete Data via the EM Algorithm'. *Journal of the Royal Statistical Society* **B-39**:1–37.
- Y. Ephraim, et al. (1989). 'A minimum discrimination information approach for hidden Markov modeling'. *IEEE Transactions On Information Theory* **35**(5):1001–1013.
- A. Farago & G. Lugosi (1989). 'An algorithm to find the global optimum of left-to-right hidden Markov model parameters'. *Problems Of Control And Information Theory-Problemy Upravleniya I Teorii Informatsii* **18**(6):435–444.
- L. R. Ford & D. R. Fulkerson (1956). 'Maximal flow through a network'. *Canadian Journal of Mathematics* **8**:399–404.
- R. A. Granat & A. Donnellan (2001). 'Deterministic Annealing Hidden Markov Models for Geophysical Data Exploration'. In *Proc. 3rd Workshop Sci. Datamining*.
- K. Heki, et al. (1997). 'Silent fault slip following an interplate thrust earthquake at the Japan Trench'. *Nature* **386**(6625):595–598.

- H. Hirose, et al. (1999). 'A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan'. *Geophysical Research Letters* **26**(21):3237–3240.
- Q. Huo & C. Chan (1993). 'The gradient projection method for the training of hidden Markov models'. *Speech Communication* **13**:307–313.
- M. Jamshidian & R. I. Jennrich (1993). 'Conjugate gradient acceleration of the EM algorithm'. *Journal of the American Statistical Association* **88**:221–228.
- B. Juang & L. Rabiner (1985). 'Mixture autoregressive hidden Markov models for speech signals'. *IEEE Transactions On Acoustics Speech And Signal Processing* **33**(6):1404–1413.
- B. Juang & L. Rabiner (1990). 'The segmental k-means algorithm for estimating parameters of hidden Markov models'. *IEEE Transactions On Acoustics Speech And Signal Processing* **38**(9):1639–1641.
- B. Juang & L. Rabiner (1991). 'Hidden Markov models for speech recognition'. *Technometrics* **33**(3):251–272.
- S. Kirkpatrick, et al. (1983). 'Optimization by simulated annealing'. *Science* **220**(4598):671–680.
- S. Kwong, et al. (2001). 'Optimisation of HMM topology and its model parameters by genetic algorithms'. *Pattern Recognition* **34**(2):509–522.
- K. Lee (1990). 'Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition'. *IEEE Transactions On Acoustics Speech And Signal Processing* **38**(4):599–609.
- K. Lee & H. Hon (1989). 'Speaker-independent phone recognition using hidden Markov models'. *IEEE Transactions On Acoustics Speech And Signal Processing* **37**(11):1641–1648.
- G. McGuire, et al. (2000). 'A Bayesian model for detecting past recombination events in DNA multiple alignments'. *Journal Of Computational Biology* **7**(1-2):159–170.
- T. Melbourne & F. Webb (2002). 'Precursory transient slip during the 2001 $M_w=8.4$ Peru earthquake sequence from continuous GPS'. *Geophysical Research Letters* **29**(21):art. no.–2032.
- T. Melbourne & F. Webb (2003). 'Slow but not quite silent'. *Science* **300**(5627):1886–1887.

- T. Melbourne, et al. (2002). 'Rapid postseismic transients in subduction zones from continuous GPS'. *Journal Of Geophysical Research – Solid Earth* **107**(B10):art. no.–2241.
- M. Miller, et al. (2002). 'Periodic slow earthquakes from the Cascadia subduction zone'. *Science* **295**(5564):2423–2423.
- L. R. Rabiner (1989). 'A tutorial on hidden Markov models and selected applications in speech recognition'. *Proc. IEEE* **77**(2):257–286.
- G. Rogers & H. Dragert (2003). 'Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip'. *Science* **300**(5627):1942–1943.
- K. Rose (1998). 'Deterministic annealing for clustering, compression, classification, regression, and related optimization problems'. *Proc. IEEE* **86**(11):2210–2239.
- K. Rose, et al. (1992). 'Vector quantization by deterministic daing'. *IEEE Trans. Inf. Theory* **38**(4):1249–1257.
- K. Rose & A. V. Rao (2001). 'Deterministically annealed design of hidden Markov model speech recognizers'. *IEEE Trans. on Speech and Audio Processing* **9**(2):111–126.
- N. Ueda & R. Nakano (1994). 'Mixture density estimate via EM algorithm with deterministic daing'. *Proceedings of the IEEE Neural Networks for Signal Processing* pp. 69–77.
- N. Ueda & R. Nakano (1998). 'Deterministic annealing EM algorithm'. *Neural Networks* **11**(2):271–282.
- M. Whiley & D. M. Titterington (2002). 'Applying the deterministic annealing expectation maximisation algorithm to naive bayesian networks'. *Univ. Glasgow Tech. Report*
- Y. Wong (1993). 'Clustering data by melting'. *Neural Computation* **5**:89–104.
- S. Young & P. Woodland (1994). 'State clustering in hidden Markov model-based continuous speech recognition'. *Computer Speech And Language* **8**(4):369–383.
- A. L. Yuille, et al. (1994). 'Statistical physics, mixtures of distributions and the EM algorithm'. *Neural Computation* **6**:334–340.