University of California

Los Angeles

Regularized Deterministic Annealing EM for Hidden Markov Models

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical Engineering

by

Robert A. Granat

2004

The dissertation of Robert A. Granat is approved.

Chiara Sabatti

Ali H. Sayed

Lieven Vandenberghe

Vwani Roychowdhury, Committee Chair

University of California, Los Angeles

2004

Table of Contents

1 Introduction **1**

1.1 Experimental Data 3

1.2 Preview of Results 5

1.2.1 Science Results 6

1.3 Outline 10

2 Hidden Markov Models **14**

2.1 Notation 15

2.2 Model optimization problem 15

2.3 Expectation-Maximization 17

2.4 Optimization procedure for the HMM 19

2.4.1 HMM Q-function Maximization 20

2.4.2 Forward-Backward Procedure 24

2.5 Finite mixture models and HMMs 26

2.5.1 Finite Mixture Models 26

2.5.2 Mixture Hidden Markov Models 29

3 Maximum Likelihood **31**

4 Deterministic Annealing **36**

4.1 Deterministic annealing for HMMs 41

5 Regularization **46**

5.1 Initial and state transition terms 48
5.2 Output distribution terms 50
5.2.1 Discrete output terms 50
5.2.2 Gaussian output terms: Mahalanobis distance 52
5.2.3 Gaussian output terms: Euclidean distance 56

6 Results and Discussion 60
6.1 Experiment Design 60
6.2 Synthetic Data Results 61
6.3 Field Data Results 66
6.4 Altering Local Maxima Criteria 68
6.5 Discussion 69

7 Local Maxima 74
7.1 Initial and Transition Probabilities 74
7.2 Output Distribution Functions 76
7.2.1 Discrete Output 76
7.2.2 Continuous Output 80

8 Science Applications 86
8.1 SCIGN GPS 86
8.2 Seismicity 89
8.3 Seismic Waveforms 93

9 Conclusions 100

References **103**

List of Figures

1.1 Two experimental data sets used in this work. Left: The data set `clar`, collected daily by a GPS station in Claremont, California. Note the unusual ground water pumping signal around days 100-200, and the abrupt shift caused by the 1999 Hector Mine earthquake on day 626. Right: The synthetic data set `step`, with a known ground truth. . . . 4

1.2 Classification results for a six-state HMM trained on the data set `clar` using the standard EM method. Note that classes (states) 1 and 6 are learning noise rather than any actual mode of the signal. 5

1.3 Classification results for a six-state HMM trained on the data set `clar` using the standard EM method. Note that the method fails to separate the signal into classes before and after the east-west shift around day 450. 6

1.4 Classification results for a six-state HMM trained on the data set `clar` using the standard EM method. Note that class 6 is learning noise rather than any actual mode of the signal. 6

1.5 Classification results for a seven-state HMM trained on the data set `clar` using the regularized deterministic annealing EM method. . . . 7

1.6 Earthquakes assigned to state 1 of a hidden Markov model trained on the Southern California seismic record (1960-1999). 8

1.7 Earthquakes assigned to state 22 of a hidden Markov model trained on the Southern California seismic record (1960-1999). 9

1.8 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. 10

1.9 Classification result of a six state HMM trained on the data set `clar`. States 4 and 5 are identical, resulting in state 5 becoming an empty state. ("E" in the legend denotes an empty state). 11

2.1 A representation of the hidden Markov model, with hidden nodes in underlying system states q, and observable variables O. 15

3.1 Left: Classification results for a ten state HMM for data set `step`. Log likelihood = 234.371. Right: Classification results for a ten state HMM for data set `step`. Log likelihood = 233.369. 34

3.2 Left: Number of local maxima for the data set `step` for HMMs with up to ten states. Right: Number of local maxima for the data set `clar` for HMMs with up to ten states. 35

4.1 Left: Number of experimentally determined local maxima for HMMs with varying numbers of hidden states applied to the data set `step`. Right: Maximum log likelihood among all experiments for HMMs with varying numbers of hidden states applied to the data set `step`. Blue squares show results for the baseline HMM with standard EM optimization; magenta stars results with schedule $\Delta\gamma = 0.1$, green circles results with schedule $\Delta\gamma = 0.01$; red triangles results with schedule $\Delta\gamma = 0.001$. 43

4.2 Top: Classification results of a six state HMM trained on the data set `step` using baseline EM. Bottom: Classification results of a six state HMM trained on the data set `clar` using baseline EM. "E" in the legend denotes an empty state. 44

6.1 Experimental results for data set `step`. Left: Number of experimentally determined local maxima. Right: Maximum log likelihood among all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta stars with schedule $\Delta\gamma = 0.1$, green circles with schedule $\Delta\gamma = 0.01$; red triangles with schedule $\Delta\gamma = 0.001$. Dashed lines are the results with deterministic annealing only, solid lines are the results of the combined annealing and regularization technique. 62

6.2 Test results for data set `step`. Left: Mean log likelihood across all HMMs. Right: Standard deviation of log likelihood across all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles with schedule $\Delta\beta = 0.01$; red triangles 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. 63

6.3 Left: Classification result from test 1 using deterministic annealing EM only for a seven state HMM trained on data set `step`. Right: Classification result from test 1 using deterministic annealing with regularization for a seven state HMM trained on data set `step`. 64

6.4 Left: Classification result from test 1 using regularized deterministic annealing EM for an eight state HMM trained on data set `step`. Right: Classification result from test 1 using the same method to train a nine state model. 65

6.5 Experimental results for data set `stepstep`. Left: Number of experimentally determined local maxima. Right: Maximum log likelihood among all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta stars with schedule $\Delta\gamma = 0.1$, green circles with schedule $\Delta\gamma = 0.01$; red triangles with schedule $\Delta\gamma = 0.001$. Dashed lines are the results with deterministic annealing only, solid lines are the results of the combined annealing and regularization technique. 66

6.6 Test results for data set `clar`. Left: Number of experimentally determined local maxima. Right: Maximum log likelihood among all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles with schedule $\Delta\beta = 0.01$; red triangles 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. 67

6.7 Test results for data set `clar`. Left: Mean log likelihood across all HMMs. Right: Standard deviation of log likelihood across all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles with schedule $\Delta\beta = 0.01$; red triangles 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. 68

6.8 Classification results for a seven-state HMM applied to the data set `clar`. 69

6.9 Classification results for a two- through six-state HMMs applied to the data set `clar`. 70

6.10 Left: Number of local maxima for different methods applied to `step` with relaxed local maxima criteria Hamming distance ≤ 1. Right: Number of local maxima for different methods applied to `clar` with relaxed local maxima criteria Hamming distance ≤ 1. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles with schedule $\Delta\beta = 0.01$; red triangles 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. 71

7.1 Example of discrete data set with state assignments such that state transitions occur only between different output symbols. 77

8.1 Coincident state changes for six-state HMMs trained on signals from each of 127 SCIGN GPS stations. 88

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. 89

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. . 90

8.4 Earthquakes assigned to state 1 of a hidden Markov model trained on the Southern California seismic record (1960-1999). 92

8.5 Earthquakes assigned to state 19 of a hidden Markov model trained on the Southern California seismic record (1960-1999). 93

8.6 Earthquakes assigned to state 22 of a hidden Markov model trained on the Southern California seismic record (1960-1999). 94

8.7 On the left is a time-spectra plot (spectrogram) showing the Landers earthquake of 1992. Key signals appearing before the earthquake are circled. (I) is a local earthquake, (II) is a teleseismic event, (III) is an aseismic event that may be the result of precursory activity. On the right are the actual TriNet time series which have been processed to form the spectrogram. A suspicious tilt signal appears approximately 20 minutes prior to the main shock. 96

8.8 Results of application of an HMM trained using regularized deterministic annealing EM to an unusual long-duration signal in Pasadena, California. The HMM classifies the background signal as one class (red), noise spikes as two classes (light and dark blue), and the long-duration signal itself as a mixture of the remaining three classes. 98

8.9 Three instances of the results of applying an HMM trained using standard EM to an unusual long-duration signal in Pasadena, California. Note the differences in classification results between the three different initializations. 99

Acknowledgments

First and foremost, I would like to thank my thesis advisor, Dr. Vwani Roychowdhury, for all his help and guidance over the course of this work. In addition I would like to thank the members of my thesis committee, Dr. Lieven Vandenberghe, Dr. Chiara Sabatti, and Dr. Ali Sayed, for their suggestions and input, as well as my former advisor, Dr. Helen Na, who helped me take the first steps.

I would also like to thank numerous colleagues at the Jet Propulsion Laboratory who supplied advice and encouragement in abundance. Dr. Kenneth Hurst, Dr. Andrea Donnellan, and Dr. Sharon Kedar, provided commentary and feedback on the geophysics aspect of the work, as well as support at critical times. Dr. Michael Turmon gave me the benefit of his mathematical and algorithmic insight on numerous occasions, Tim Stough assisted with technical problems of all sorts, large and small. Kacie Shelton assisted with editing.

I am indebted as well to all the friends who helped to remind me of the good things in life in the midst of writing and research. Chief honors in this regard go to Anand Chelian, Colin O'Neal, and Justin Fang – I owe you all for far more than this. My family also deserves credit for their boundless support; to my parents and dearly departed grandparents I give my heartfelt thanks.

Also deserving of thanks are a few individuals, who, however unwittingly, helped me renew my motivation just as it was flagging: Dr. Kiri Wagstaff, Alex Fukunaga, and Jorge Cham. There are also a few individuals who, by providing me with irreplaceable opportunities, helped to set me down the path that led to this work: Dr. Jennifer Sun, Dr. Pietro Perona, and Dr. Michael Burl.

Vita

April 10, 1975	Born, Princeton, New Jersey
1996	B.S., Engineering and Applied Science California Institute of Technology Pasadena, California
1996–1998	Member Technical Staff Jet Propulsion Laboratory Pasadena, California
1997–1998	Research Assistant Department of Electrical Engineering University of California, Los Angeles
1998	M.S., Electrical Engineering University of California, Los Angeles
1998–present	Senior Member Technical Staff Jet Propulsion Laboratory Pasadena, California

Publications

Granat, R. and H. Na, "Estimating dynamic ionospheric changes without a prior models." *Radio Science*, 35(2):341–349, 2000.

Granat, R. and A. Donnellan, "A hidden Markov model based tool for geophysical data exploration." *Pure and Applied Geophysics*, 159(10):2271–2283, 2002.

Turmon, M., R. Granat, D. S. Katz and J. Z. Lou, "Tests and tolerances for high-performance software-implemented fault detection." *IEEE Trans. on Computers*, 52(5):579–591, 2003.

Granat, R., "A method of hidden Markov model optimization for use with geophysical data sets." *Computational Science – ICCS 2003, Pt. III, Proceedings*, 2659:892-901, 2003.

Abstract of the Dissertation

Regularized Deterministic Annealing EM for Hidden Markov Models

by

Robert A. Granat

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2004

Professor Vwani Roychowdhury, Chair

It is well known that maximum likelihood optimization of hidden Markov models (HMMs) suffers from the problem of local maxima. Applications of HMMs to date have focused primarily on areas, such as speech recognition and protein sequence analysis, where a priori knowledge can be used to constrain the problem and thereby avoid local maxima. We tackle the problem of optimizing HMMs in situations where such a priori information is not available. Our motivating problem is that of applying HMMs to analysis of geophysical time series, but these circumstances are common in many types of scientific analysis. We present evidence that the problem is not addressed by existing methodologies, and present an alternative approach based on the use of the deterministic annealing expectation-maximization (EM) algorithm. We address the major weakness of the annealing method by designing statistical priors that regularize the solution away from particular local maxima in which there are redundant states. We present an analysis of the performance of the method on both synthetic and field instrument data, demonstrating the superior ability of the method to avoid local maxima as compared to the standard and deterministic annealing EM algorithms. In addition

we present mathematical analysis showing that for common data types there is an exponential lower bound on the number of HMM local maxima. We conclude by showing results of the method as applied to the analysis of several geophysical data sets.

CHAPTER 1

Introduction

In this work we address problems that occur when using the maximum likelihood criteria for optimization of hidden Markov model (HMM) parameters. In particular, we are concerned with the well known problem of locally maximum solutions. We posit that even though hidden Markov models have enjoyed considerable practical success in a number of application areas, the local maximum problem remains a significant barrier to more general application of HMMs.

Successful applications of HMM technology can be found in areas including speech synthesis and recognition for continuous output HMMs, and protein matching and analysis for discrete output HMMs. In these domains the local maxima problem 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 [JR85, FL89, MWP00], restrictions on the form of the output distributions [EDR89], and parameter tying [BN90, YW94, BM01]. 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 [LH89, Lee90] but also the details of the actual physical process which generates sound waves [JR91].

Our motivating problem is of a different nature. We wish to perform exploratory analysis of scientific data where the physics of the underlying system

is uncertain and a matter of ongoing discovery and debate. This means that not only do we lack any basis for generating constraints, but also that we actively want to avoid restricting the solution, since that may bias it away from discovering interesting science. Nevertheless we are motivated to use HMMs in this kind of analysis because in many cases the physical evidence suggests that the underlying system does in fact undergo distinct state changes that can be modeled by a discrete hidden variable.

As an example of this kind of exploratory research we applied HMMs to analysis of time series measurements of geophysical systems. These measurements are collected by a variety of instruments, for instance global positioning system (GPS) based measurements of ground displacement or seismographic measurements of surface velocity. The data collected by these sensors is only peripheral to the actual physical system itself – we cannot measure displacement and velocity any significant distance beneath the earth's crust, for example.

In this context we explored the number of local maxima encountered by the expectation-maximization (EM) algorithm in the absence of constraints. Empirical testing on our sample data sets revealed that the EM method encounters large numbers of local maxima even for modestly sized data sets and models with few states. Since we were unable to use approaches reliant on domain knowledge, we turned to more general optimization techniques. Our chosen approach was to use the deterministic annealing EM method [UN98], which employs an annealing effect to suppress local maxima early in the optimization process. Our application of the deterministic annealing method to unsupervised training of HMMs resulted in a significant improvement over the baseline EM method. However, we also discovered that the annealing method tends to get stuck in a certain set of systemic local maxima which are characterized by the existence redundant states

with identical output distributions. To address this issue, we designed statistical priors that act to bias the solution away from these systemic local maxima. Use of this regularization scheme along with the deterministic annealing method resulted in a greatly improved ability to avoid local maxima as compared to both the baseline and deterministic annealing EM methods. Since this was done without introducing a bias against any specific characteristic of the data, we were able to perform the optimization without influencing the scientific results.

1.1 Experimental Data

To facilitate our explanations throughout this work we intersperse results of experiments performed using two particular data sets. The first of these, which we designate `clar`, consists of global positioning system (GPS) derived relative displacement measurements in three dimensions (north-south, east-west, and vertical) collected daily over about two years spanning 1998-1999. These GPS measurements were collected by the Claremont, California station of the Southern California Integrated Geodetic Network (SCIGN), which is dedicated to studying the relationship of crustal deformation to earthquake processes. 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 left side of figure 1.1 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). We can also observe several more subtle signals, for instance the increased noise and the beginning and end of the time window and the small east-west shift around

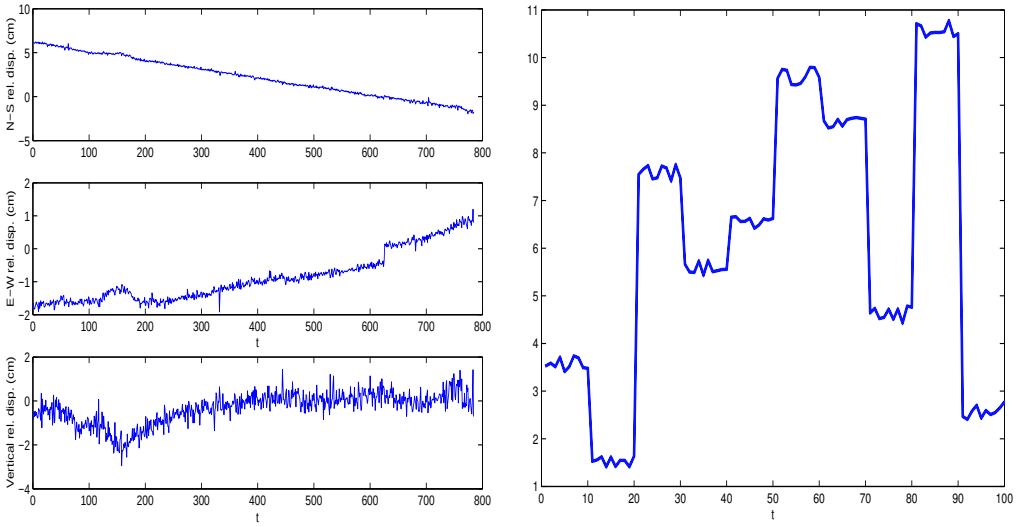


Figure 1.1: Two experimental data sets used in this work. Left: The data set `clar`, collected daily by a GPS station in Claremont, California. Note the unusual ground water pumping signal around days 100-200, and the abrupt shift caused by the 1999 Hector Mine earthquake on day 626. Right: The synthetic data set `step`, with a known ground truth.

day 450.

The `clar` data set is an excellent example of our target scientific data, but it lacks a definative ground truth in terms of both its statistical properties and underlying state assignment. We therefor introduce a synthetic data set which we designate `step`. This one-dimensional data set consists of a series of discrete steps with integer values from 1 to 10 to which has been added uniform random noise with values on $[-0.4, +0.4]$. The order of the steps has been randomly shuffled. This data set can be seen on the right side of figure 1.1.

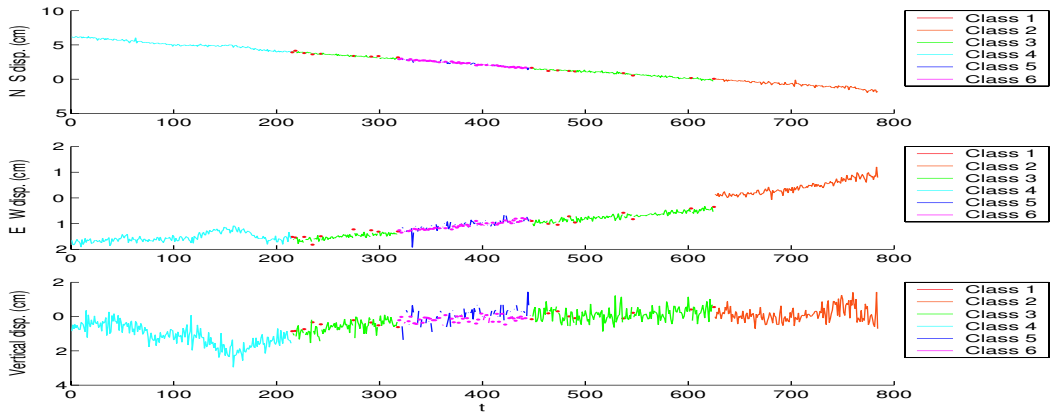


Figure 1.2: Classification results for a six-state HMM trained on the data set `clar` using the standard EM method. Note that classes (states) 1 and 6 are learning noise rather than any actual mode of the signal.

1.2 Preview of Results

We now present a brief preview of the results of our method, highlighting the advantage in performance it enjoys over the basic EM technique. Figures 1.2, 1.3, and 1.4 show several classification results produced by HMMs trained on the Claremont, California GPS data set `clar` using the standard EM method. In each case a different random initialization of the model parameters was used, resulting in a different solution each time. The model solutions differ considerably from one another, resulting in very different state sequence assignments (which are interpreted as observation classifications). We contrast this with the classification result shown in figure 1.5. This was the only solution found in a thousand applications of our regularized deterministic annealing EM training method, each also employing a different random parameter initialization. This solution is not only stable across different initializations, but also of higher quality, correctly identifying both the ground water pumping signal and the Hector Mine earthquake signal.

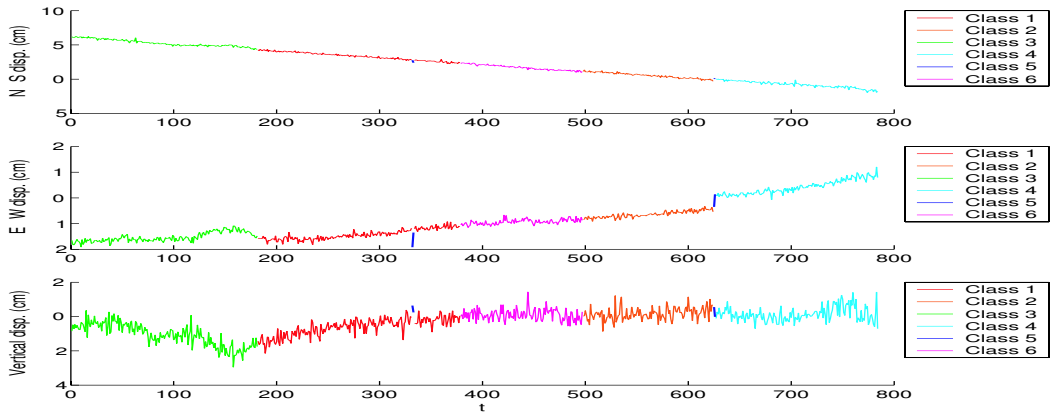


Figure 1.3: Classification results for a six-state HMM trained on the data set `clar` using the standard EM method. Note that the method fails to separate the signal into classes before and after the east-west shift around day 450.

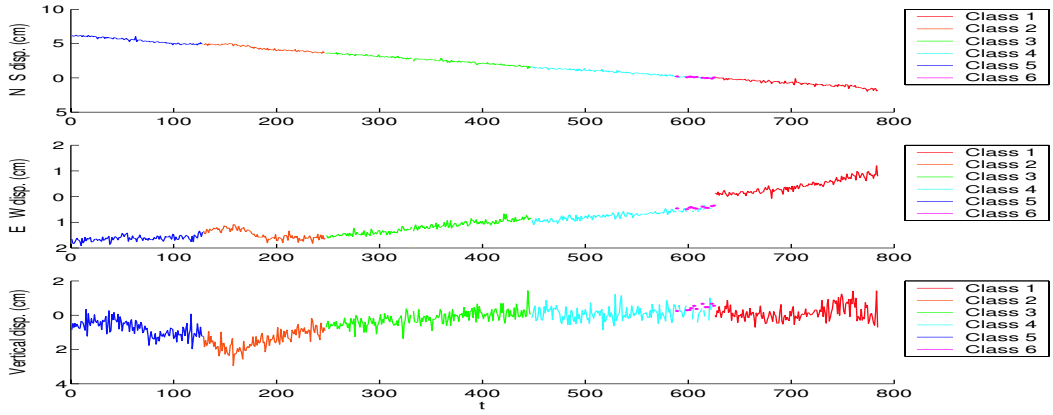


Figure 1.4: Classification results for a six-state HMM trained on the data set `clar` using the standard EM method. Note that class 6 is learning noise rather than any actual mode of the signal.

1.2.1 Science Results

We used the capabilities of our model optimization technique to investigate long-range fault interactions. Long range fault interactions are currently an important area of study in geophysics, but to date evidence of such interactions has been

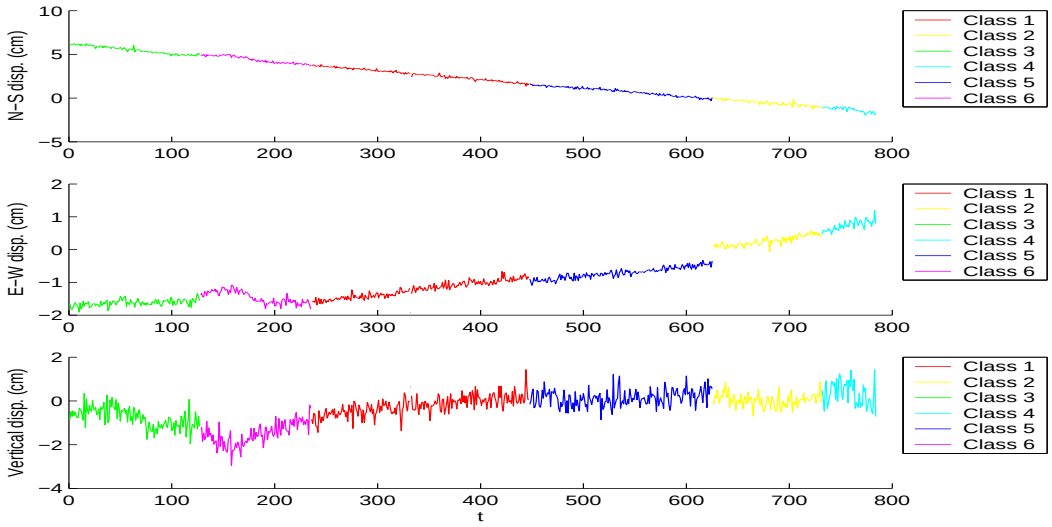


Figure 1.5: Classification results for a seven-state HMM trained on the data set `clar` using the regularized deterministic annealing EM method.

scarce and their nature not understood. Using our method, we are able to find evidence of such interactions in two geophysical data sets. The first of these data sets is a record of seismicity in Southern California, in which the location (latitude, longitude, depth) and magnitude is recorded. We used our method to fit a 25-state hidden Markov model to this data and observed that transition probability between states 1 and 22 was 0.7136, implying a strong correlation between the events. Since the state 1 earthquakes (see figure 1.6) were deep earthquakes along the coast side of the San Andreas fault system and the state 22 earthquakes (shown in figure 1.7) were events over 50km away in the northern part of the Sierre Nevada fault system, long-distance stress transfer was strongly implied.

We also used our method to fit six-state hidden Markov models to 127 time series of surface displacement spanning 1998-1999 collected by a network of geodetic sensors in Southern California. We then looked for correlations between state

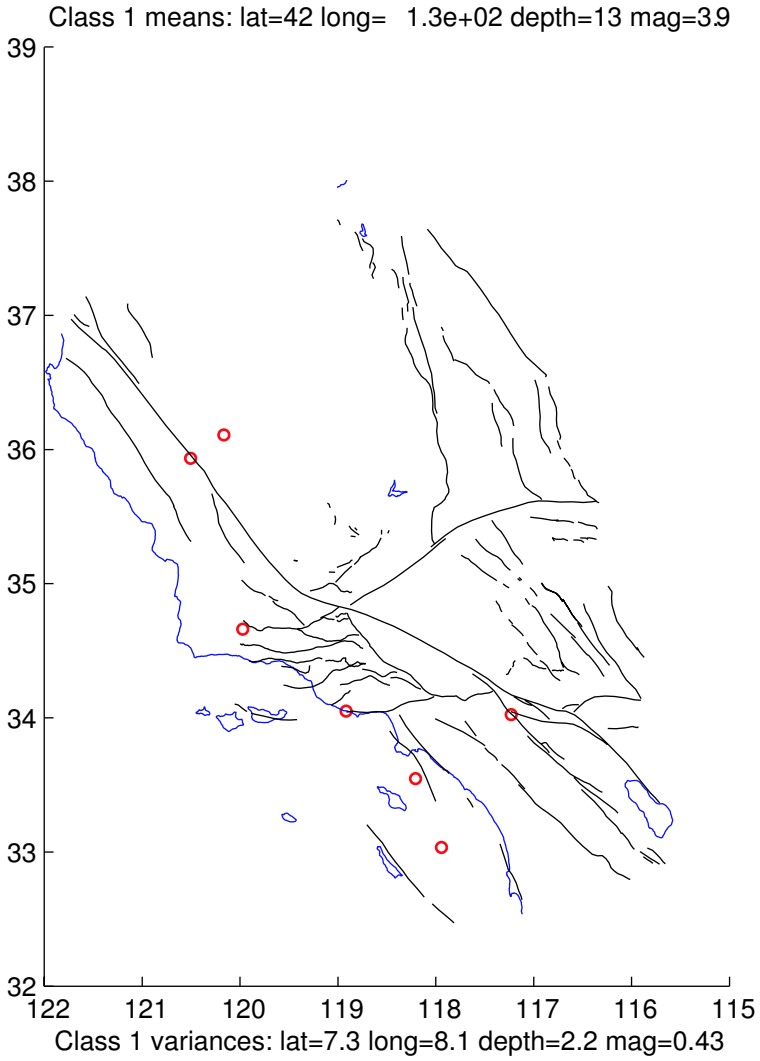


Figure 1.6: Earthquakes assigned to state 1 of a hidden Markov model trained on the Southern California seismic record (1960-1999).

changes across different stations; a large number of correlated state changes would imply the existence of a region-wide event with an effect across multiple fault systems. The result of these correlation measurements can be seen in figure 1.8. The large peak on day 652 corresponds to the strong Hector Mine earthquake of 1999, but there are no other large earthquakes during this time period. The implica-

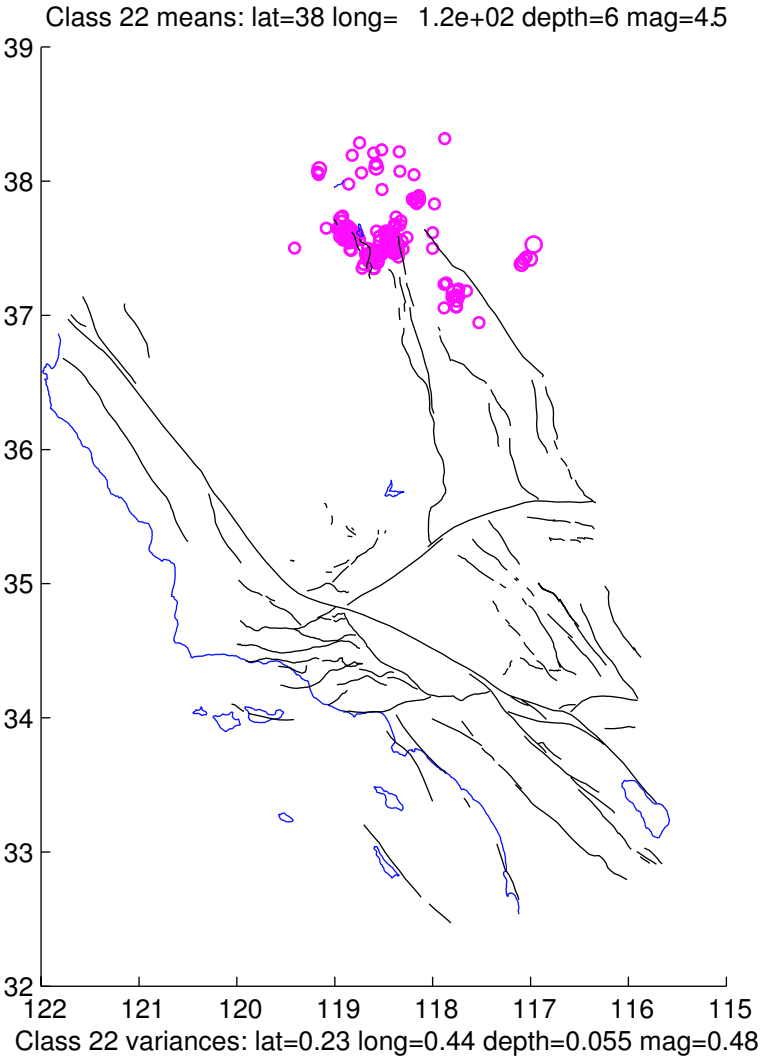


Figure 1.7: Earthquakes assigned to state 22 of a hidden Markov model trained on the Southern California seismic record (1960-1999).

tion is that there are other, aseismic effects that are taking place on a regional, inter-fault scale. We note also the significant reduction of noise obtained by our method (red) as opposed to the standard EM optimization method (blue).

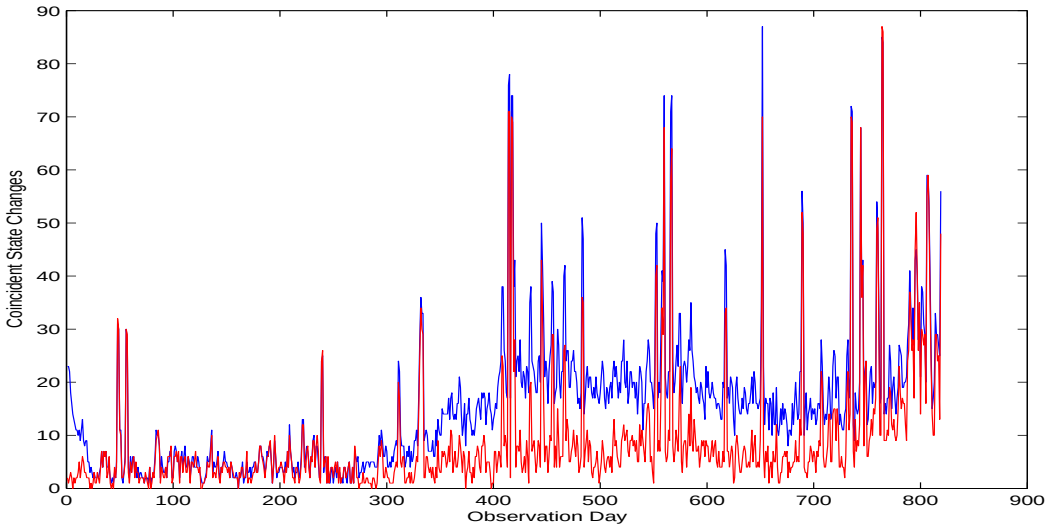


Figure 1.8: Coincident state changes for six-state HMMs trained using standard EM (blue) and regularized deterministic annealing EM (red) on signals from each of 127 SCIGN GPS stations.

1.3 Outline

We begin this work in Chapter 2 with a review of the mathematics of hidden Markov models. We present the maximum likelihood optimization problem for HMMs and describe in detail the derivation of the expectation-maximization procedure for solving that problem. We continue with a similar discussion of finite mixture models and mixed continuous HMMs. This sets up the mathematical framework we will use for the remainder of our discussion, putting in context our modifications to the standard EM approach.

In Chapter 3 we discuss the maximum likelihood objective function, its boundedness, and its applicability to exploratory scientific research. We discuss methods for determining the number of local maxima in the objective function and advocate an empirical approach that uses the Hamming distance between the individually most likely state sequence assignments to determine the number of

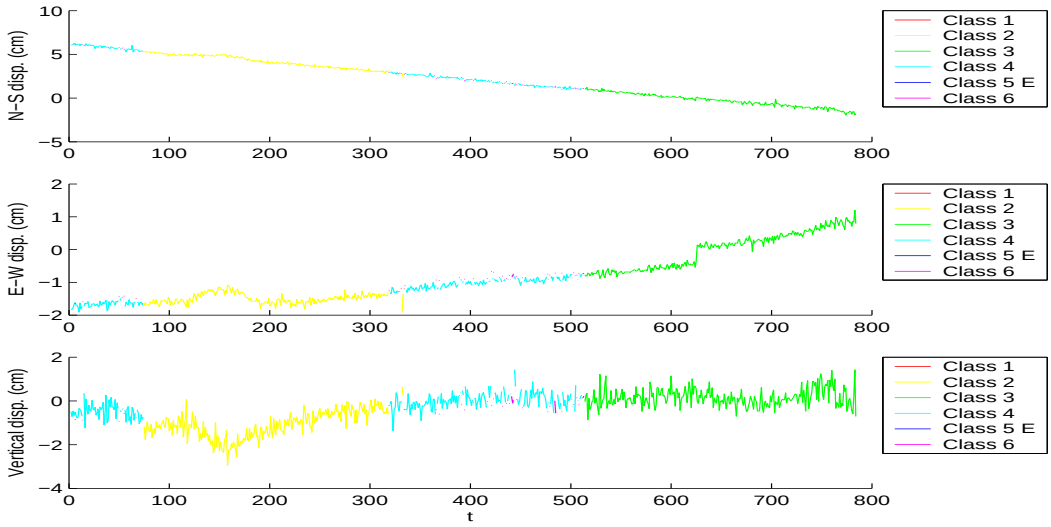


Figure 1.9: Classification result of a six state HMM trained on the data set `clar`. States 4 and 5 are identical, resulting in state 5 becoming an empty state. ("E" in the legend denotes an empty state).

local maxima. Applying this method to our test data sets, we discover that the number of local maxima rises rapidly with model size when using the standard EM approach.

Chapter 4 introduces the deterministic annealing EM method, first presenting the general mathematics of the approach, and then applying it in the specific to the optimization of hidden Markov models. We discuss the performance of the method on our test data sets and analyze its behavior during the optimization process. We conclude that although the method delivers significant improvement over baseline EM optimization, it suffers from a tendency to be stuck in local maxima where there are redundant states with identical output distributions. These identical output distributions usually have the effect of producing an "empty" state to which no observations are assigned, as shown in figure 1.9.

We identify these data-independent systemic local maxima as being the chief

difficulty encountered by the deterministic annealing method (later in Chapter 7 we demonstrate that there are an exponential number of such maxima).

This provides us with the motivation to develop statistical priors that bias the solution away from these systemic local maxima, while still retaining good properties from an optimization perspective. We present these priors in Chapter 5. In practice, the priors appear in the optimization procedure as regularization/penalty terms that modify the so-called Q-function maximized in the "M"-step of the EM algorithm. We present several priors that are applied to the initial state probability distribution, the state-to-state probability distribution, and either discrete or Gaussian output probability distributions. Each prior comes with a weighting term that controls the degree to which it modifies the original objective function. We derive bounds on these weighting parameters that guarantee concavity of the Q-function and thereby the ability to find a unique maximum during the "M"-step.

Results of combining deterministic annealing with our regularization scheme are presented in Chapter 6. We demonstrate that the two approaches together deliver greatly superior performance than either alone. We discuss the strengths and weaknesses of the approach, and suggest some ways in which the procedure could be improved through combination with other complementary techniques.

We return to a more mathematical approach in Chapter 7, in which we present analysis of the local maxima of the likelihood function for HMMs and construct locally maximum solutions with redundant states. Using these constructions, we show that number of locally maximum solutions of this form is provably bounded below by an exponential for common data types.

Concluding this work is a presentation in Chapter 8 of scientific results produced through application of the method to several geophysical data sets. These

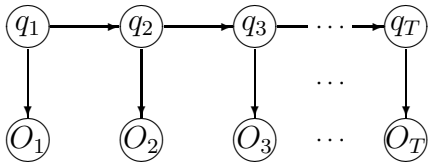
data sets are currently in use for analysis of Southern California seismic activity, and consist of seismic catalogs, GPS-based measurements of surface deformation, and seismographic velocity signals. We show how our approach is able to aid scientific understanding of earthquake systems.

CHAPTER 2

Hidden Markov Models

We begin with a review of hidden Markov models (HMMs). We describe the structure of HMMs and pose finding the maximum likelihood HMM for a given observation sequence as a non-convex optimization problem. We then review the most common optimization method used to solve this problem, the expectation-maximization (EM) algorithm, and detail its application to HMMs, laying the mathematical groundwork for our later improvements. Lastly, we describe the relationship between hidden Markov models and a similar class of models, finite mixture models.

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.



Partially observed Markov chain.

Figure 2.1: A representation of the hidden Markov model, with hidden nodes in underlying system states q, and observable variables O.

2.1 Notation

Our notation is similar to that employed by Rabiner [Rab89] and is as follows: a hidden Markov model λ with N states is composed of a vector of initial state probabilities $\pi = (\pi_1, \ldots, \pi_N)$, a matrix of 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 $b_i(y, \mu_i, \Sigma_i)$. An observation sequence O of length T is denoted $O_1 O_2 \cdots O_T$ and a state sequence Q of the model is denoted $q_1 q_2 \cdots q_T$.

2.2 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 func-

tion. 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 [JR90], maximum mutual information (MMI), [BBS86] minimum discrimination information (MDI) [EDR89], and maximum classification error (MCE) [CLJ94]. Of these, all but the first require labeled training examples on which to train the models, making them inappropriate for our targeted application domains. 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 $O = O_1 O_2 \cdots O_T$, 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

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda). \tag{2.1}$$

Assuming statistical independence of observations,

$$P(O|Q,\lambda) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T). \tag{2.2}$$

The probability of the given state sequence Q is

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}. \tag{2.3}$$

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

$$P(O,Q|\lambda) = P(O|Q,\lambda) P(Q|\lambda), \tag{2.4}$$

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

$$P(O|\lambda) = \sum_{\text{all } Q = q_1 q_2 \cdots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T). \tag{2.5}$$

We can pose the optimization of $P(O|\lambda)$ as a non-convex optimization problem:

$$\begin{aligned}
\text{Maximize}: \quad & P(O|\lambda) \\
\text{Subject to}: \quad & \sum_{i=1}^{N} \pi_i = 1 \\
& \pi_i \geq 0,\ i = 1, \ldots, N \\
& \sum_{j=1}^{N} a_{ij} = 1,\ i = 1, \ldots, N \\
& a_{ij} \geq 0,\ i = 1, \ldots, N,\ j = 1, \ldots, N \\
& \sum_{m=1}^{M} b_i(m) = 1,\ i = 1, \ldots, N \\
& b_i(m) \geq 0,\ i = 1, \ldots, N,\ m = 1, \ldots, M. \qquad (2.6)
\end{aligned}$$

Note that the above is for the discrete output case. In the case of continuous outputs, the last two constraints are replaced by

$$\begin{aligned}
\int_Y b_i(y)dy = 1,\ i = 1, \ldots, N \\
b_i(y) \geq 0,\ i = 1, \ldots, N, y \in Y. \qquad (2.7)
\end{aligned}$$

This problem is 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 [DLR77], although alternative approaches exist, such as those employing genetic algorithms [KCM01] recursive predictive error techniques [CKM94], or gradient projection [HC93].

2.3 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

$$P(\lambda) = E[p(x,\lambda)|\lambda] = \int_{\mathcal{X}} p(x,\lambda) d\mu(x) \tag{2.8}$$

and

$$Q(\lambda,\lambda') = E[\log p(x,\lambda')|\lambda] = \int_{\mathcal{X}} p(x,\lambda) \log p(x,\lambda') d\mu(x), \tag{2.9}$$

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 $\mathcal{X} \to \mathcal{Y}(\mathcal{X})$.

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

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

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.

2.4 Optimization procedure for the HMM

We now present the specific instance of the EM algorithm for calculating the optimal HMM parameters, based on that first suggested by Baum and colleagues [Bau72, BE67, BP66, BPS70, BS68]. For the hidden Markov model, we have the complete data likelihood

$$p(Q, O, \lambda) = \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T), \tag{2.10}$$

with $P(\lambda) = E[p(q, O, \lambda)|\lambda]$ defined as in (2.5). 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 this as

$$\sum_{i=1}^{N} z_{i1} \log \pi_i + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} z_{it} z_{j,t+1} \log a_{ij} + \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} \log b_i(O_t). \tag{2.11}$$

From this we can calculate

$$Q(\lambda^{(k)}, \lambda) = \sum_{i=1}^{N} \tau_{i1}^{(k)} \log \pi_i + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij} + \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \log b_i(O_t) \tag{2.12}$$

where

$$\tau_{ijt} = \quad P(Z_{it} = 1, Z_{j,t+1} = 1|O, \lambda) \quad t = 1, \ldots, T-1, \tag{2.13}$$

$$\tau_{it} = \qquad P(Z_{it} = 1|O, \lambda) \qquad t = 1, \ldots, T, \tag{2.14}$$

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

2.4.1 HMM Q-function Maximization

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

$$Q_1(\lambda^{(k)}, \lambda) = \sum_{i=1}^{N} \tau_{i1}^{(k)} \log \pi_i, \tag{2.15}$$

$$Q_2(\lambda^{(k)}, \lambda) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij}, \tag{2.16}$$

$$Q_3(\lambda^{(k)}, \lambda) = \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \log b_i(O_t). \tag{2.17}$$

Maximization of each component may be pursued separately. However, a direct solution by calculation of the critical points of the first two components is not possible. For instance,

$$\frac{\partial Q_1}{\partial \pi_i} = \frac{\partial \sum_{i=1}^{N} \tau_{i1}^{(k)} \log \pi_i}{\partial \pi_i} = \frac{\tau_{i1}^{(k)}}{\pi_i} = 0 \tag{2.18}$$

is clearly not useful, and derivatives of Q_2 fare similarly.

Instead we solve the general convex optimization problem

$$\begin{aligned} \text{Minimize :} \quad & f(x) = -\sum_{i=1}^{N} c_i \log x_i \\ \text{Subject to :} \quad & \sum_{i=1}^{N} x_i = 1, \end{aligned} \tag{2.19}$$

with constants $c_i \geq 0$. First, we calculate the Lagrangian

$$L(x, \nu) = -\sum_{i=1}^{N} c_i \log x_i + \nu(\sum_{i=1}^{N} x_i - 1), \tag{2.20}$$

which has a maximum in the x_is at $x_i = c_i/\nu$. The dual problem is then

$$\text{Maximize :} \quad g(\nu) = -\sum_{i=1}^{N} c_i \log(c_i/\nu) + \sum_{i=1}^{N} c_i - \nu. \tag{2.21}$$

The maximum of the dual problem can be found by setting the derivative equal to zero and solving. We find that $\nu^* = \sum_{i=1}^N c_i$ is the maximum, with $g(\nu^*) = -\sum_{i=1}^N c_i \log(c_i / \sum_{i=1}^N c_i)$. Since $f(x) = g(\nu^*)$ is feasible with $x_i = c_i / \sum_{i=1}^N c_i$, this is the minimizing solution to the primal problem.

2.4.1.1 Maximization of Initial Probabilities

Since τ_{i1} is dependent only on the first observation, we can calculate using Bayes' Theorem:

$$\tau_{i1} = \frac{\pi_i b_i(O_1)}{\sum_{j=1}^N \pi_j b_j(O_1)}. \tag{2.22}$$

Our solution to (2.19) gives us the values π_i which maximize Q_1:

$$\pi_i = \frac{\tau_{i1}^{(k)}}{\sum_{j=1}^N \tau_{i1}^{(k)}} = \tau_{i1}^{(k)} = \frac{\pi_i^{(k)} b_i^{(k)}(O_1)}{\sum_{j=1}^N \pi_j^{(k)} b_j^{(k)}(O_1)}. \tag{2.23}$$

2.4.1.2 Maximization of Transition Probabilities

Similarly, we can use the solution to (2.19) to give us the values a_{ij} which maximize Q_2:

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}^{(k)}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}^{(k)}}. \tag{2.24}$$

Noting that $\tau_{it} = \sum_{j=1}^N \tau_{ijt}$ for $t = 1, \ldots, T-1$, we have

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}^{(k)}}{\sum_{t=1}^{T-1} \tau_{it}^{(k)}}. \tag{2.25}$$

2.4.1.3 Maximization of Discrete Output Distributions

If the outputs of the model are discrete, we can apply our solution to (2.19) once more by noting that

$$Q_3 = \sum_{i=1}^N \sum_{t=1}^T \sum_{m=1}^M \tau_{it}^{(k)} \delta(O_t - m) \log b_i(m), \tag{2.26}$$

and therefor the output probability distributions that maximize Q_3 are

$$b_i(m) = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{\sum_{t=1}^{T} \tau_{it}^{(k)}}. \tag{2.27}$$

where m is a possible discrete output symbol.

2.4.1.4 Maximization of Continuous Output Distributions

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:

$$\begin{aligned} Q_3 &= \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \left(\log n - \frac{1}{2} \log \det(\Sigma_i) - \frac{1}{2} (O_t - \mu_i)^T \Sigma_i^{-1} (O_t - \mu_i) \right) \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \Bigg(\log n - \frac{1}{2} \log \det(\Sigma_i) - \frac{1}{2} (m_i - \mu_i)^T \Sigma_i^{-1} (m_i - \mu_i) \\ &\qquad - \frac{1}{2} (O_t - m_i)^T \Sigma_i^{-1} (O_t - m_i) \Bigg), \end{aligned} \tag{2.28}$$

where $m_i = \sum_{t=1}^{T} \tau_{it}^{(k)} O_t / \sum_{t=1}^{T} \tau_{it}^{(k)}$. Let

$$S_i = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} (O_t - m_i)(O_t - m_i)^T}{\sum_{t=1}^{T} \tau_{it}^{(k)}}. \tag{2.29}$$

Then

$$\begin{aligned} Q_3 = \sum_{t=1}^{T} \sum_{i=1}^{N} \tau_{it}^{(k)} \Bigg(&\log n + \frac{1}{2} \log \det(\Sigma_i^{-1}) \\ &- \frac{1}{2} (m_i - \mu_i)^T \Sigma_i^{-1} (m_i - \mu_i) - \frac{1}{2} \operatorname{Tr} \Sigma_i^{-1} S_i \Bigg). \end{aligned} \tag{2.30}$$

Since Σ_i is positive definite, we see that Q_3 is maximized in the μ_is when

$$\mu_i = m_i = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} O_t}{\sum_{t=1}^{T} \tau_{it}^{(k)}}. \tag{2.31}$$

Given this maximizing solution for μ_i, we can solve for Σ_i directly by taking the derivative. For a D-by-D matrix A, let the matrix B be such that

$$\{B\}_{ij} = \mathrm{cof}_{ji}(A). \tag{2.32}$$

Then we have

$$\begin{aligned} AB &= \det(A) I \\ B &= \det(A) A^{-1} \\ \{B\}_{ij} &= \det(A) \{A^{-1}\}_{ij}, \end{aligned} \tag{2.33}$$

and therefore

$$\frac{\partial \det(A)}{\partial \{A\}_{ij}} = \frac{\partial}{\partial \{A\}_{ij}} \sum_{i=1}^{D} \{A\}_{ij} \mathrm{cof}_{ij}(A) = \mathrm{cof}_{ij}(A) = \{B\}_{ji} = \det(A) \{A^{-1}\}_{ji}, \tag{2.34}$$

and

$$\frac{\partial \log \det(A)}{\partial \{A\}_{ij}} = \{A^{-1}\}_{ji}. \tag{2.35}$$

Using these relations, we calculate the derivative of Q_3 with respect to each element of the Σ_i^{-1}s (neglecting constant factors) and set the result equal to zero:

$$\frac{\partial Q_3}{\partial \{\Sigma_i^{-1}\}_{ab}} = \{\Sigma_i\}_{ba} - \{S_i\}_{ba} = 0. \tag{2.36}$$

From this we see that Q_3 has a critical point in the Σ_is at

$$\Sigma_i = S_i = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} (O_t - \mu_i^{(k+1)})(O_t - \mu_i^{(k+1)})^T}{\sum_{t=1}^{T} \tau_{it}^{(k)}}. \tag{2.37}$$

Since Q_3 is concave this is a global maximum.

2.4.2 Forward-Backward Procedure

We have seen how to maximize the model parameters given the probabilities τ_{it} and τ_{ijt}, but how do we calculate these quantities at each iteration of the EM algorithm? 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

$$\alpha_t(i) = P(O_1 \cdots O_t, Z_{it} = 1|\lambda). \tag{2.38}$$

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:

$$\alpha_1(i) = \pi_i b_i(O_1), \ i = 1, \ldots, N. \tag{2.39}$$

2. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad t = 1, \ldots, T-1, \quad j = 1, \ldots, N. \tag{2.40}$$

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

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{2.41}$$

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

$$\beta_t(i) = P(O_{t+1} \cdots O_T|Z_{it} = 1, \lambda). \tag{2.42}$$

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:

$$\beta_T(i) = 1, \ i = 1, \ldots, N. \tag{2.43}$$

2. Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \qquad t = T-1, \ldots, 1, \quad i = 1, \ldots, N. \tag{2.44}$$

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

Now we can calculate the probabilities τ_{it} and τ_{ijt} using the forward and backwards variables.

$$\begin{aligned} \tau_{it} &= P(Z_{it} = 1 | O, \lambda) \\ &= \frac{P(Z_{it} = 1 | O, \lambda) P(O|\lambda)}{P(O|\lambda)} \\ &= \frac{P(Z_{it} = 1 | \lambda)}{P(O|\lambda)} \\ &= \frac{P(O_1 \cdots O_t, Z_{it} = 1 | \lambda) P(O_{t+1} \cdots O_T | Z_{it} = 1, \lambda)}{P(O|\lambda)} \\ &= \frac{\alpha_t(i) \beta_t(i)}{P(O|\lambda)} \\ &= \frac{\alpha_t(i) \beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i) \beta_t(i)} \end{aligned} \tag{2.45}$$

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

$$q_t = \underset{1 \le i \le N}{\operatorname{argmax}}(\tau_{it}), \ t = 1, \ldots, T. \tag{2.46}$$

We can also 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

$$\begin{aligned}
\tau_{ijt} &= P(Z_{it}=1, Z_{j,t+1}=1|O,\lambda) \\
&= \frac{P(Z_{it}=1, Z_{j,t+1}=1, O|\lambda)}{P(O|\lambda)} \\
&= \frac{P(O_1\cdots O_t, Z_{it}=1|\lambda)P(O_{t+1}\cdots O_T, Z_{j,t+1}=1|Z_{it}=1,\lambda)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)P(O_{t+1}, Z_{j,t+1}=1|Z_{it}=1,\lambda)P(O_{t+2}\cdots O_T|Z_{j,t+1}=1,\lambda)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)P(Z_{j,t+1}=1|Z_{it},\lambda)P(O_{t+1}|Z_{it}=1, Z_{j,t+1}=1,\lambda)\beta_{t+1}(j)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^N\sum_{j=1}^N \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}. \qquad (2.47)
\end{aligned}$$

2.5 Finite mixture models and HMMs

We now introduce a different class of models, finite mixture models, which are related to hidden Markov models. A finite mixture model is also a statistical for data, but it does not depend on any ordering of the observations (and ignores any ordering should it exist). Finite mixture models are often combined with hidden Markov models and an analysis of both can provide some insight into the problem of local maxima.

2.5.1 Finite Mixture Models

A finite mixture model [TSM85] λ_{fmm} with R components is composed of the mixture parameters $w = (w_1, \ldots, w_R)$ and the observation probability density functions associated with each mixture component, $b_r(m)$ for discrete output symbols, or $b_r(y, \theta_{r1}, \ldots, \theta_{rM})$ for continuous outputs. In general, we wish to

solve the following problem:

$$\begin{aligned}
\text{Maximize :} \quad & \prod_{t=1}^{T} P(O_t|\lambda_{fmm}) \\
\text{Subject to :} \quad & \sum_{r=1}^{R} w_r = 1 \\
& w_r \geq 0,\ r = 1, \ldots, R \\
& \sum_{m=1}^{M} b_r(m) = 1,\ r = 1, \ldots, R \\
& b_r(m) \geq 0,\ r = 1, \ldots, R,\ m = 1, \ldots, M. \qquad (2.48)
\end{aligned}$$

We can express the objective function in terms of the model parameters as follows:

$$\prod_{t=1}^{T} P(O_t|\lambda_{fmm}) = \prod_{t=1}^{T} \sum_{r=1}^{R} w_r b_r(O_t). \qquad (2.49)$$

In the case of continuous outputs, the last constraint is replaced by

$$\begin{aligned}
& \int_Y b_r(y)dy = 1,\ r = 1, \ldots, R \\
& b_r(y) \geq 0,\ r = 1, \ldots, R, y \in Y. \qquad (2.50)
\end{aligned}$$

Once again we use the EM method to solve this optimization problem [RW84]. We can represent the complete data log likelihood for the finite mixture model as

$$\sum_{r=1}^{R} \sum_{t=1}^{T} z_{rt} \log w_r b_r(O_t) \qquad (2.51)$$

where $z = (z_1, \ldots, z_T)$ is a set of component indicator vectors such that $z_{rt} = 1$ if the observation is drawn from the rth mixture component, $z_{rt} = 0$ otherwise. From this we can calculate

$$Q(\lambda_{fmm}, \lambda_{fmm}^{(k)}) = \sum_{r=1}^{R} \sum_{t=1}^{T} \tau_{rt}^{(k)} \log w_r b_r(O_t) \qquad (2.52)$$

where

$$\tau_{rt} = P(Z_{rt} = 1|O_t, \lambda_{fmm})\ t = 1, \ldots, T \qquad (2.53)$$

and Z is a probabilistic component indicator variable analogous to z. We can calculate $\tau_{it}^{(k)}$ via Bayes' Rule:

$$\tau_{rt}^{(k)} = \frac{w_r^{(k)} b_r^{(k)}(O_t)}{\sum_{r=1}^{R} w_r^{(k)} b_r^{(k)}(O_t)}. \tag{2.54}$$

We choose updates of w_r and b_r that maximize Q. We can find the update for w_r using our solution to (2.19):

$$w_r = \frac{\sum_{t=1}^{T} \tau_{rt}^{(k)}}{\sum_{t=1}^{T} \sum_{r=1}^{R} \tau_{rt}^{(k)}} = \frac{1}{T} \sum_{t=1}^{T} \tau_{rt}^{(k)}. \tag{2.55}$$

To find the update rule for b_r we find the maximum directly via the derivative, solving

$$\frac{\partial L_F}{\partial \theta_{rm}} = \sum_{t=1}^{T} \tau_{rt}^{(k)} \frac{\partial}{\partial \theta_{rm}} \log b_r(O_t, \theta_{rm}) = 0, \tag{2.56}$$

which has no general analytical solution. As in the HMM case, for certain forms of the output distribution an analytic solution is available.

2.5.1.1 HMMs as FMMs

The hidden Markov model can be seen as a special case of finite mixture model, one in which there is a single observation O and N^T mixture components, each corresponding to a different state sequence Q. In this view we have

$$w_Q = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}, \tag{2.57}$$

$$b_Q(O) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T). \tag{2.58}$$

Optimization of the model posed in this way is of course difficult given the exponential number of mixture components. Although this formulation is contrived in the sense that it ignores the time ordered structure that allows efficient solution via the forward-backward method, it nevertheless gives us some insight into why the hidden Markov model optimization problem may be inherently more difficult

than the finite mixture model optimization problem. And in fact, there is a very real problem of an exponential number of HMM local maxima that we discuss in detail in Chapter 7.

2.5.2 Mixture Hidden Markov Models

We can also construct a hidden Markov model whose state outputs are themselves finite mixture models. Such models have found particularly widespread use in the field of speech processing (see, for example, [JLS86]). We can formulate the model as follows: if each finite state of the model has R mixture components, then the model is $\lambda = (\pi, A, w, B)$ where $w = (w_{11}, \ldots, w_{NR})$, $B = (b_{11}, \ldots, b_{NR})$ and π and A retain their original meanings. Let $W = (w_{1r_1}, \ldots, w_{Nr_N})$ be some choice of mixture components for each model state. Then for this model we have

$$\begin{aligned} P(O|\lambda) &= \sum_{\text{all } W,Q} \pi_{q_1} w_{q_1 r_{q_1}} b_{q_1 r_{q_1}}(O_1) a_{q_1 q_2} w_{q_2 r_{q_2}} b_{q_2 r_{q_2}}(O_2) \cdots \\ &\quad \cdots a_{q_{T-1} q_T} w_{q_T r_{q_T}} b_{q_T r_{q_T}}(O_T). \end{aligned} \tag{2.59}$$

Calculation of the forward and backward parameters proceeds as follows:

1. Initialization:

$$\alpha_t(i) = \pi_i \sum_{r=1}^{R} w_{ir} b_{ir}(O_1), \ i = 1, \ldots, N. \tag{2.60}$$

$$\beta_T(i) = 1, \ i = 1, \ldots, N. \tag{2.61}$$

2. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] \sum_{r=1}^{R} w_{jr} b_{jr}(O_{t+1}), \qquad t = 1, \ldots, T-1, \quad j = 1, \ldots, N. \tag{2.62}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} \left(\sum_{r=1}^{R} w_{jr} b_{jr}(O_{t+1}) \right) \beta_{t+1}(j), \qquad t = T-1, \ldots, 1, \quad i = 1, \ldots, N. \tag{2.63}$$

Derivation of this procedure follows that of the forward-backward procedure for the standard hidden Markov model. Once the forward and backward variables have been calculated, we can derive τ_{it} and τ_{ijt} according to (2.45) and (2.47), with the difference that

$$\tau_{ijt} = \frac{\alpha_t(i)a_{ij}\left(\sum_{r=1}^{R} w_{jr}b_{jr}(O_{t+1})\right)\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}. \tag{2.64}$$

This allows us to re-estimate at π and A at each iteration using (2.23) and (2.24). Now we have

$$\begin{aligned}\tau_{irt} &= P(Z_{irt}=1|O,\lambda) \\ &= P(Z_{irt}=1|Z_{it}=1,O,\lambda)P(Z_{it}|O,\lambda) \\ &= \frac{w_{ir}b_{ir}(O_t)}{\sum_{r=1}^{R} w_{ir}b_{ir}(O_t)}\tau_{it}\end{aligned} \tag{2.65}$$

We can re-estimate the mixture weights according to:

$$w_{ir} = \frac{\sum_{t=1}^{T}\tau_{irt}^{(k)}}{\sum_{r=1}^{R}\sum_{t=1}^{T}\tau_{irt}^{(k)}} = \frac{\sum_{t=1}^{T}\tau_{irt}^{(k)}}{\sum_{t=1}^{T}\tau_{it}^{(k)}}. \tag{2.66}$$

Again there is no general form for the output distributions, but in the special case of Gaussian outputs for each mixture model component we have,

$$\mu_{ir} = \frac{\sum_{t=1}^{T}\tau_{irt}^{(k)}O_t}{\sum_{t=1}^{T}\tau_{irt}^{(k)}}, \tag{2.67}$$

$$\Sigma_{ir} = \frac{\sum_{t=1}^{T}\tau_{irt}^{(k)}(O_t-\mu_i^{(k+1)})(O_t-\mu_i^{(k+1)})^T}{\sum_{t=1}^{T}\tau_{irt}^{(k)}}. \tag{2.68}$$

It is worth noting that this model reduces to a simple finite mixture model in the case that the hidden Markov model has but one state. This means that hidden Markov models and finite mixture models can be seen a special cases of the each other.

CHAPTER 3

Maximum Likelihood

It is well known that the maximum likelihood function for hidden Markov models is unbounded above for many common continuous output distributions, including members of the exponential family. Unbounded solutions occur when the probability mass of the output distributions becomes concentrated on one or several observations.

Our first reaction to this property is pragmatic – we simply want to find a way to avoid this kind of overfitting and make sure that we can find a non-degenerate, bounded solution. One way to do this is to develop a modified, or penalized, maximum likelihood function which is bounded. A generalized analysis of this approach for the Gaussian mixture model case can be found in [CRI03]; the necessary properties required of a penalty function presented therein extend to hidden Markov models as well. In this work we employ HMMs with multivariate Gaussian output distributions and use a conjugate prior (penalty function) as described in [OT96]. This prior is a simplified Wishart density [Bun94] with the form $\prod_{i=1}^{N} \exp(-\frac{\omega_\Sigma}{2} \operatorname{Tr} \Sigma_i^{-1})$. When employed in the expectation-maximization (EM) algorithm, this term has the practical effect of adding the quantity ω_Σ to the diagonal elements of the covariance matrices at every iteration. We note that without the use of this prior, the experiments and results presented later in this work would not be possible. (Henceforth, we assume the use of this prior at all times.)

Our second, more considered, reaction is to ask whether we really do want to be using the maximum likelihood objective function. This may seem strange, but since in fact we desire a local, not a global maximum (since the global maximum is unbounded), what makes us sure that the highest likelihood bounded solution is really the "correct" solution? Consider the case in which we use constraints to bound the solution. It's certainly possible that the maximum likelihood solution exists right on the constraint boundary as the optimization procedure pushes the solution towards the unbounded global maximum, but one doubts that this solution is desirable.

We suggest that the answer to this questions is application and paradigm specific. For example, in the case of an automatic speech recognizer, we demand definite classification results from the system and we have quantitative ways to analyze performance. In the case of scientific exploration, however, the user (a scientist) may wish to examine several different locally maximum model solutions. From the scientist's perspective, each may represent a potential science result worthy of followup through experiment, field observation, and/or physics-based analysis. We point to previous work applying HMMs to seismicity data [GD02], as well as work applying finite mixture models to studies of solar physics [TPM02] and other fields [SIG99, SC98] as examples of work conducted under this paradigm. In such cases we don't necessarily want to find the single solution with the highest likelihood, but rather a set of quality solutions (see [CL99] for an example of such in the case of DNA sequence analysis). If we merely reduce the number of candidate solutions enough, we will have achieved our goals. In this case if the number of local maxima found by the optimization method is $O(10)$ that is usually sufficient. Of course, we do not actively avoid finding a single solution, in such cases it is merely necessary that we have sufficient confidence in the result.

Having established our objective when modeling such data, we now turn to the actual problem of model optimization. Optimizing the HMM model parameters is a non-linear, non-convex problem requiring an iterative solution. The most common method for solving this problem is the expectation-maximization (EM) algorithm [DLR77], although alternative approaches exist, such as those employing genetic algorithms [KCM01], recursive predictive error techniques [CKM94], or gradient projection [HC93]. Our question is, if we only wish to find set of reasonable local maxima solutions, is this basic EM approach sufficient? To answer this we need to be able to determine the number of locally maximum solutions found by EM. Simple comparison of model parameters to determine equivalence is ineffective in practice due to variations caused by differing initializations and the limitations of finite precision arithmetic. Previous work has emphasized the use of likelihood to determine model distance and thereby whether different solutions correspond to the same local maximum [RJL85, HKM99, HKM00]. While this can provide some insight, experimentation reveals that there are often considerable differences between models with very close likelihoods. Figure 3.1 shows the classification results for two ten state models trained on the data set `step`. Although the model solutions have similar log likelihoods, the classification results differ considerably: in the left hand classification, the third and eighth steps have been split into two classes, while the sixth, seventh, and ninth steps have been grouped into a single class; in the right hand classifaction, the fourth and seventh steps have each been split into two classes and the first, second, and tenth steps grouped into a single class.

The implication of such results is that we need an alternative method for determining the distance between model solutions. Our approach is to use the Hamming distance between the individually most likely state assignments for the observation sequence (i.e., the classification results) as defined in equation (2.46).

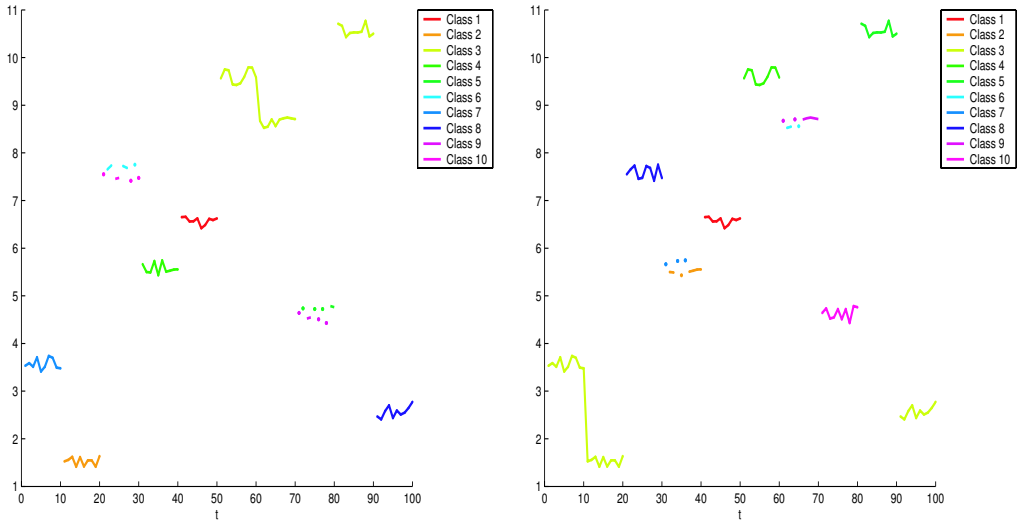


Figure 3.1: Left: Classification results for a ten state HMM for data set `step`. Log likelihood = 234.371. Right: Classification results for a ten state HMM for data set `step`. Log likelihood = 233.369.

We use a linear assignment method based on bipartite graph matching [FF56] 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. At this point it is necessary to note that while the EM algorithm only guarantees convergence to a critical point of the objective function and can even converge to local minima for some objective functions [ACK93], in practice convergence to other than a local maxima is extremely rare. A simple check of relative log likelihoods across test results should be sufficient to identify such unusual cases. In any event in practice it is the fixed points of the optimization method that are at issue in the

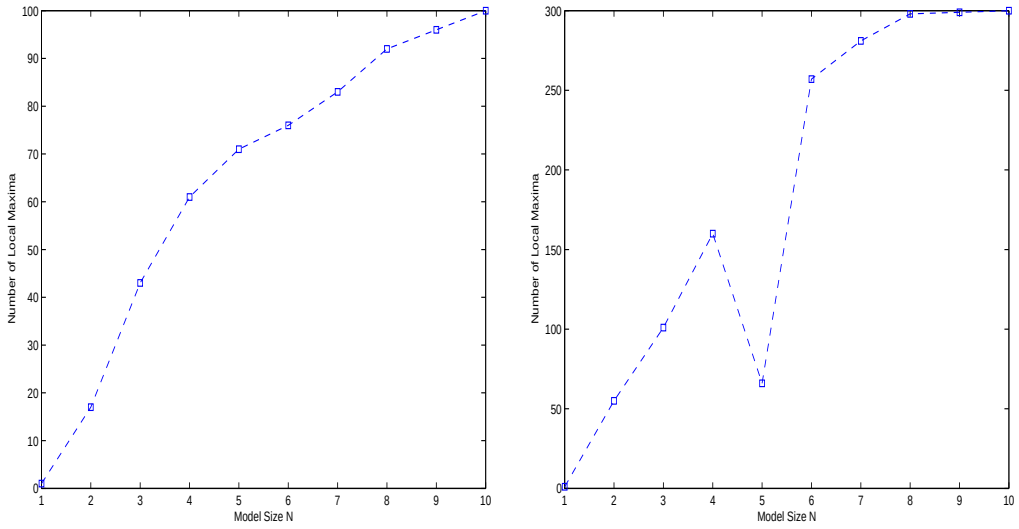


Figure 3.2: Left: Number of local maxima for the data set `step` for HMMs with up to ten states. Right: Number of local maxima for the data set `clar` for HMMs with up to ten states.

absence of a local maxima test that can be applied upon convergence.

While this approach 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. Figure 3.2 shows the results of such tests using the standard EM method for the `step` and `clar` data sets. While these results are for 1000 trial applications of the method, the number of local maxima ceased increasing after about 200 trials in both cases. We observe that there are a large number of local maxima, even for the two state model, and that the number of local maxima increases rapidly with model size. This indicates that the EM algorithm alone is insufficient for our needs. To address this problem we turn to the method of deterministic annealing, which offers a non-problem specific way to avoid local maxima.

CHAPTER 4

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; this new cost function 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 [KGV83], 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 [RGF92] and for clustering problems [BK93, Won93]. Yuille and colleagues [YSU94] 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 [Ros98]. The particular framework we present here was first applied by Ueda and Nakano to mixture density estimation problems [UN94] and then extended to the general case [UN98], 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 at the previous temperature, 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 [RR01] but differs 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.

In the introduction of the deterministic annealing method that follows, we parallel the presentation of the material in [UN98], with some modifications for clarity. Recall from our discussion of the EM algorithm that we have

$$\begin{aligned} P(\lambda, y) &= \int_{\mathcal{X}} p(x, y, \lambda) d\mu(x) \\ Q(\lambda, \lambda') &= \int_{\mathcal{X}} p(x|y, \lambda) \log p(x, y, \lambda') d\mu(x). \end{aligned}$$

Now we define a new function using the posterior $f(x|y)$:

$$E(\lambda) = -\int_{\mathcal{X}} f(x|y) \log p(x, y, \lambda) d\mu(x), \tag{4.1}$$

so that if $f(x|y) = p(x|y, \lambda)$ then minimization of E is equivalent to maximization of Q. Since we lack prior knowledge about f we use the principle of maximum

entropy to specify the probability. That is, we wish to solve the maximization problem

$$\begin{aligned} \text{Maximize :} \quad & S = -\int_{\mathcal{X}} f(x|y)\log f(x|y)d\mu(x) \\ \text{Subject to :} \quad & \int_{\mathcal{X}} f(x|y)d\mu(x) = 1 \\ & f(x|y) \geq 0 \\ & \int_{\mathcal{X}} f(x|y)\log p(x,y,\lambda)d\mu(x) = -E. \end{aligned} \tag{4.2}$$

To solve this problem we construct the Lagrangian

$$\begin{aligned} L_S(f,\lambda) = & -\int_{\mathcal{X}} f(x|y)\log f(x|y)d\mu(x) \\ & + \nu\left(\int_{\mathcal{X}} f(x|y)d\mu(x) - 1\right) + \gamma\left(\int_{\mathcal{X}} f(x|y)\log p(x,y,\lambda)d\mu(x) + E\right). \end{aligned} \tag{4.3}$$

The variation of $L_S, \delta L_S$, due to the variation of $f, \delta f$, is

$$\delta L_S = \int_{\mathcal{X}} \big(-1 - \log f(x|y) + \nu + \gamma \log p(x,y,\lambda)\big)\delta f d\mu(x). \tag{4.4}$$

Since at the maximum $\delta L_S = 0$ regardless of the value of δf, we have

$$-1 - \log f(x|y) + \nu + \gamma \log p(x,y,\lambda) = 0, \tag{4.5}$$

and so

$$f(x|y) = \exp\big(1 - \nu - \gamma \log p(x,y,\lambda)\big). \tag{4.6}$$

Integrating both sides over $\mathcal{X}$ and noting the problem constraints yields

$$\exp(1-\nu) = \frac{1}{\int_{\mathcal{X}} \exp\big(\gamma \log p(x,y,\lambda)\big)}. \tag{4.7}$$

Eliminating the variable ν we obtain

$$f(x|y) = \frac{\exp\big(\gamma \log p(x,y,\lambda)\big)}{\int_{\mathcal{X}} \exp\big(\gamma \log p(x,y,\lambda)\big)}, \tag{4.8}$$

which is a Gibbs distribution with partition function

$$Z = \int_{\mathcal{X}} \exp\big(\gamma \log p(x, y, \lambda)\big) d\mu(x). \tag{4.9}$$

Given this partition function we can define the free energy as a cost function depending on the temperature:

$$\begin{aligned} F(\gamma, \lambda) &= -\frac{1}{\gamma} \log Z \\ &= -\frac{1}{\gamma} \log \exp\big(\gamma \log p(x, y, \lambda)\big) d\mu(x). \end{aligned} \tag{4.10}$$

We note from (4.8) that

$$-\frac{1}{\gamma} \log \int_{\mathcal{X}} p^{\gamma}(x, y, \lambda) d\mu(x) = -\log p(x, y, \lambda) + \frac{1}{\gamma} \log f(x|y, \lambda). \tag{4.11}$$

Taking the conditional expectation with respect to the distribution f, we have

$$-\frac{1}{\gamma} \log \int_{\mathcal{X}} p^{\gamma}(x, y, \lambda) d\mu(x) = U(\gamma, \lambda) - \frac{1}{\gamma} S(\gamma, \lambda), \tag{4.12}$$

where

$$U(\gamma, \lambda) = E_f[-\log p(x, y, \lambda)|y] = -\int_{\mathcal{X}} f(x|y) \log p(x, y, \lambda) d\mu(x), \tag{4.13}$$

$$S(\gamma, \lambda) = E_f[-\log f(x|y, \lambda)|y] = -\int_{\mathcal{X}} f(x|y) \log f(x|y) d\mu(x). \tag{4.14}$$

So we can write

$$F(\gamma, \lambda) = U(\gamma, \lambda) - \frac{1}{\gamma} S(\gamma, \lambda). \tag{4.15}$$

It is known that at equilibrium a thermodynamic system settles into a configuration that minimizes its free energy. Therefore we consider the problem of minimizing F at a fixed temperature $\gamma > 0$. To perform the minimization, we perform an iterative algorithm very similar to the EM algorithm. Let λ' be an estimate of λ. Then taking the conditional expectation given y and λ' we have

$$F(\gamma, \lambda) = U(\gamma, \lambda|\lambda') - \frac{1}{\gamma} S(\gamma, \lambda|\lambda'), \tag{4.16}$$

where

$$U(\gamma, \lambda|\lambda') = E_f[-\log p(x, y, \lambda)|y, \lambda'] = -\int_{\mathcal{X}} f(x|y, \lambda') \log p(x, y, \lambda) d\mu(x), \tag{4.17}$$

and

$$S(\gamma, \lambda|\lambda') = E_f[-\log f(x|y, \lambda)|y, \lambda'] = -\int_{\mathcal{X}} f(x|y, \lambda') \log f(x|y) d\mu(x). \tag{4.18}$$

Then if $\lambda = \overline{\lambda}$ minimizes $U(\gamma, \lambda|\lambda')$, then $F(\gamma, \overline{\lambda}) \leq F(\gamma, \lambda')$, where equality holds if and only if both $U(\gamma, \overline{\lambda}|\lambda') = U(\gamma, \lambda'|\lambda')$ and $f(x|y, \overline{\lambda}) = f(x|y, \lambda')$. Proof:

$$F(\gamma, \overline{\lambda}) - F(\gamma, \lambda') = \big(U(\gamma, \overline{\lambda}|\lambda') - U(\gamma, \lambda'|\lambda')\big) + \frac{1}{\gamma}\big(S(\gamma, \lambda'|\lambda') - S(\gamma, \overline{\lambda}|\lambda')\big). \tag{4.19}$$

Since $U(\gamma, \overline{\lambda}|\lambda')$ is a minimum, we have

$$U(\gamma, \overline{\lambda}|\lambda') - U(\gamma, \lambda'|\lambda') \leq 0. \tag{4.20}$$

And since

$$\begin{aligned} S(\gamma, \lambda'|\lambda') - S(\gamma, \overline{\lambda}|\lambda') &= \int_{\mathcal{X}} f(x|y, \lambda') \log\left(\frac{f(x|y, \overline{\lambda})}{f(x|y, \lambda')}\right) d\mu(x) \\ &\leq \log \int_{\mathcal{X}} f(x|y, \lambda') \left(\frac{f(x|y, \overline{\lambda})}{f(x|y, \lambda')}\right) d\mu(x) \\ &= \log \int_{\mathcal{X}} f(x|y, \overline{\lambda}) d\mu(x) = \log 1 = 0, \end{aligned} \tag{4.21}$$

we therefore have

$$F(\gamma, \overline{\lambda}) - F(\gamma, \lambda') \leq 0. \tag{4.22}$$

From this we can show that for a transformation $\mathcal{F}$ that if $\mathcal{F}(\lambda)$ is a critical point of $U(\gamma, \lambda|\lambda')$ as a function of λ', then the fixed points of $\mathcal{F}$ are critical points of F. This suggests that we can minimize the free energy using an iterative algorithm in which we minimize the function $U(\gamma, \lambda, \lambda')$ at each step. The method

is identical to the EM algorithm, except that the function Q is replaced by U. A deterministic annealing EM algorithm uses this method as an internal step, and proceeds as follows:

1. Set $\gamma = \gamma_{min} > 0$.
2. Start with $k = 0$ and pick a starting $\lambda^{(k)}$.
3. Calculate
$$U(\gamma, \lambda|\lambda^{(k)}) = -\int_{\mathcal{X}} \log p(x, y, \lambda) \frac{p^\gamma(x, y, \lambda^{(k)})}{\int_{\mathcal{X}} p^\gamma(x, y, \lambda^{(k)}) d\mu(x)} d\mu(x)$$
4. Minimize $U(\gamma, \lambda|\lambda^{(k)})$ over λ. This gives us the transformation $\mathcal{F}$.
5. Set $\lambda^{(k+1)} = \mathcal{F}(\lambda^{(k)})$. If $U(\gamma, \lambda|\lambda^{(k+1)}) - U(\gamma, \lambda|\lambda^{(k)})$ is below some threshold, go on to step 6. Otherwise, go to step 3.
6. Increase γ. If $\gamma < 1$, go to step 3. Otherwise, stop.

4.1 Deterministic annealing for HMMs

We can apply the deterministic annealing algorithm to hidden Markov models in a straightforward way. At each temperature we optimize over the function

$$U(\gamma, \lambda|\lambda^{(k)}) = \sum_{i=1}^{N} \tau_{i1}^{(k)}(\gamma) \log \pi_i + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \tau_{ijt}^{(k)}(\gamma) \log a_{ij} + \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)}(\gamma) \log b_i(O_t), \quad (4.23)$$

where

$$\tau_{it}(\gamma) = \frac{\alpha_t(i, \gamma)\beta_t(i, \gamma)}{\sum_{i=1}^{N} \alpha_t(i, \gamma)\beta_t(i, \gamma)} \quad (4.24)$$

$$\tau_{ijt}(\gamma) = \frac{\alpha_t(i, \gamma)a_{ij}^\gamma b_j^\gamma(O_{t+1})\beta_{t+1}(j, \gamma)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i, \gamma)a_{ij}^\gamma b_j^\gamma(O_{t+1})\beta_{t+1}(j, \gamma)}. \quad (4.25)$$

The modified forward and backward variables $\alpha(\gamma)$ and $\beta(\gamma)$ are calculated by a modification of the standard iteratitive procedure:

1. Initialization:

$$\alpha_t(i,\gamma) = \pi_i^\gamma b_i^\gamma(O_1),\ i = 1,\ldots,N. \tag{4.26}$$

$$\beta_T(i,\gamma) = 1,\ i = 1,\ldots,N. \tag{4.27}$$

2. Induction:

$$\alpha_{t+1}(j,\gamma) = \left[\sum_{i=1}^{N} \alpha_t(i,\gamma) a_{ij}^\gamma\right] b_j^\gamma(O_{t+1}), \qquad t = 1,\ldots,T-1,\ j = 1,\ldots,N. \tag{4.28}$$

$$\beta_t(i,\gamma) = \sum_{j=1}^{N} a_{ij}^\gamma b_j^\gamma(O_{t+1})\beta_{t+1}(j,\gamma), \qquad t = T-1,\ldots,1,\ i = 1,\ldots,N. \tag{4.29}$$

Figure 4.1 shows the results of 1000 tests of the deterministic annealing method applied to the synthetic data set `step`. For these tests we used three different annealing schedules, each with a fixed rate $\Delta\gamma$. We observe that the application of the method results in a large reduction in the number of local maxima. However, at first glance it appears that, counter to our intuition, the faster annealing schedules are producing better results than our slowest schedule. On closer examination, however, we see that in fact these slower schedules are producing extremely low likelihood solutions. In these cases, almost the entire time series is being assigned to a single state. This points to a fundamental problem with the deterministic annealing method, despite the encouraging results produced by the slow annealing schedule.

The core of this problem is caused by the fact that the deterministic annealing method is subject to being caught in certain types of local maxima where the

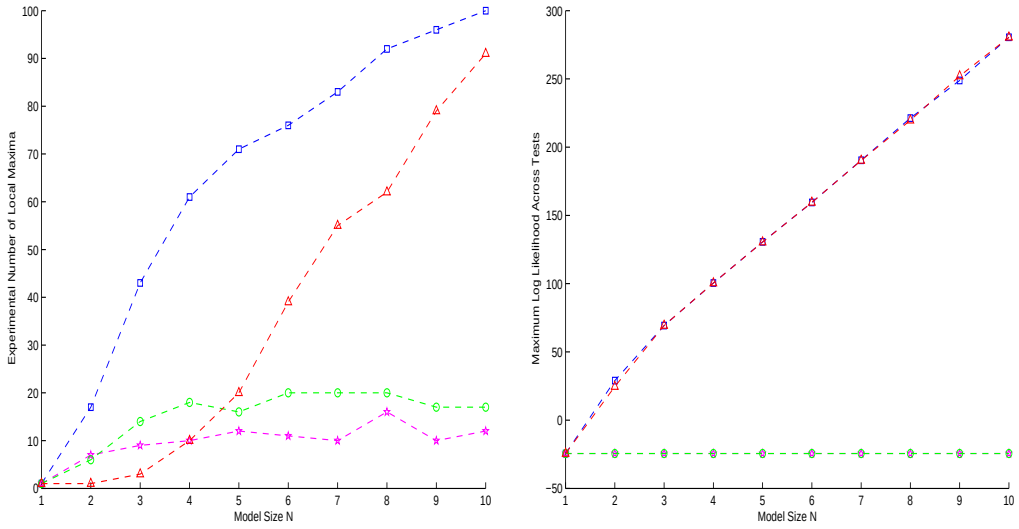


Figure 4.1: Left: Number of experimentally determined local maxima for HMMs with varying numbers of hidden states applied to the data set `step`. Right: Maximum log likelihood among all experiments for HMMs with varying numbers of hidden states applied to the data set `step`. Blue squares show results for the baseline HMM with standard EM optimization; magenta stars results with schedule $\Delta\gamma = 0.1$, green circles results with schedule $\Delta\gamma = 0.01$; red triangles results with schedule $\Delta\gamma = 0.001$.

output distributions are close or identical. These redundant states are often "empty," that is, no observations are assigned according to equation (2.46). This effect occurs with the standard EM approach (examples of which are shown in figure 4.2) but the problem is more pronounced with deterministic annealing.

This issue was discussed in detail for the case of naive Bayes networks by Whiley and Titterington [WT02]; that analysis extends straightforwardly to the HMM case as well. Intuitively, consider what happens if one starts with $\gamma_{min} = 0$. Then we have $\tau_{it}(\gamma) = 1/N$ and $\tau_{ijt}(\gamma) = 1/N^2$ and so $\pi_i = 1/N$, $a_{ij} = 1/N$, and $b_i = b_j$ for all i, j. Since this is a local maxima of the objective function for $0 \leq \gamma \leq 1$, this will then be the (undesirable) final solution for the model

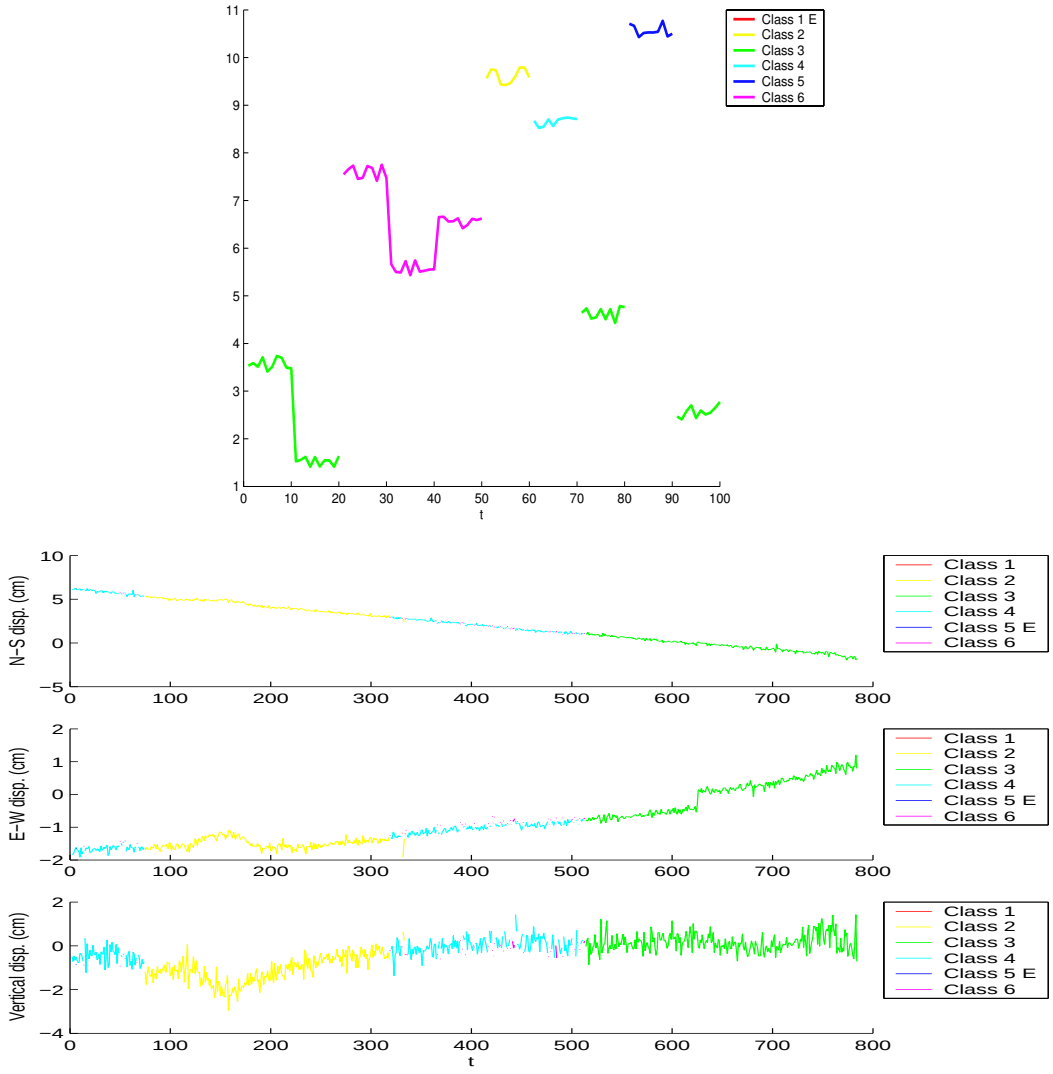


Figure 4.2: Top: Classification results of a six state HMM trained on the data set `step` using baseline EM. Bottom: Classification results of a six state HMM trained on the data set `clar` using baseline EM. "E" in the legend denotes an empty state.

parameters. (This solution is analogous to the equally weighted independence model for naive Bayes networks.) While it may it appear that this problem can be circumvented by restricting $\gamma_{min} > 0$, this in turn implies that the final optimization solution is once again sensitive to the initial values of the model

parameters. One might think that simply setting γ_{min} close to zero would be sufficient to address this issue, but experiment has shown that if γ_{min} is too small, the solution still converges undesirably as if $\gamma_{min} = 0$. Our observations lead us to suspect that for low γ where the HMM objective function is concave, the global maximum is in fact the equally weighted independence model. Ueda and Nakano suggest addressing local maxima issues by calculating an estimate of the Hessian at stationary points and perturbing the solution via line search in the direction of the eigenvectors corresponding to the negative eigenvalues of the Hessian. Since this is computationally expensive they suggest as an alternative a random perturbation of the solution; it is this approach we use in our implementation. Our experiments indicate that this perturbation does help to some extent, but is not sufficient to solve the problem. This agrees with the results presented by Whiley and Titterington for naive Bayes networks. We therefor seek to modify the basic deterministic annealing approach to address this weakness.

CHAPTER 5

Regularization

The analysis of the preceding section indicates that 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 response to this is to design regularization terms that act to push the optimization procedure away from these parts of the parameter space.

While the use of statistical priors has seen widespread use in many areas (for instance, in the optimization of neural networks) their use in the optimization of hidden Markov models has been somewhat less common. Statistical priors for hidden Markov models have been used most frequently in applications to DNA sequencing and analysis. Training data in such cases is often limited, and so to discourage overtraining a prior is applied to the discrete output distributions. These priors are most often Dirichlet type or similar, and act to smooth and flatten the distributions, pushing the solution away from simply copying the distribution of observed symbols in the training sequences. For an example of such, see [CL99]. Statistical priors have also been used to bias the initial and state transition parameters of the model. McGuire et. al. [MWP00] apply HMMs to phylogenetic analysis and claim to use a prior that represents a priori knowledge about the difficulty of changing topology during a recombination event. The prior targets therefor targets the state to state transition probabilities of the model. Only the revised update rule for the transition probabilities is presented, however,

and the originating prior is somewhat unclear. This update rule is

$$a_{ij} = \omega\delta_{ij} + (1-\omega)\sum_{t=1}^{T}\tau_{it}^{(k)}/T, \tag{5.1}$$

where δ_{ij} is the Kronecker delta function and ω is a weighting factor. Brand [Bra99] proposed applying an entropic prior to the initial and state transition probabilities of continuous output HMMs with the goal of driving those probabilities towards 0 or 1 and thereby facilitating the trimming of states. Brand observed good results from this method in experiment, but our own application of the method to the data sets `step`, `clar`, and others were much less successful. In particular, we observed that use of the prior resulted in the elimination of states even when N was less than the ground truth for the system. An even greater difficulty, however, was the fact that use of the entropic prior resulted in a mixed concave/convex Q-function in π and A, making the optimization procedure as a whole problematic.

This last difficulty encountered in employing the method of Brand highlights the fact that it is important not only to design priors that accurately reflect the available a priori knowledge about the system, but also to make sure that the resulting regularized optimization problem remains tractable. This is the guiding philosophy of this work and we take particular care to make sure that the regularized Q-function is concave in all cases. This takes on particular importance because the priors we use generate positive terms which are added to the Q-function. This contrasts with the more common regularization approach in which the terms introduce bias by penalizing the solution (i.e., negative terms would be added to a maximization problem). The concavity condition and resulting weight constraints ensure that the solution, while biased, is still bounded.

Recall that for the HMM, the Q-function is constructed by taking the expec-

tation

$$E[\log p(q, O, \lambda')|\lambda] = \int_{\mathcal{Q}\to\mathcal{O}(\mathcal{Q})} p(q, O, \lambda) \log p(q, O, \lambda') d\mu(q) \tag{5.2}$$

where to avoid confusion of notation q is the state sequence hidden variable with domain $\mathcal{Q}$ and $\mathcal{O}$ is the domain of the observations O. If we apply the prior $p(\lambda)$ to $p(q, O, \lambda)$ we generate a new Q-function

$$\begin{aligned} Q'(\lambda, \lambda') &= E[\log p(q, O, \lambda')p(\lambda')|\lambda] \\ &= \int_{\mathcal{Q}\to\mathcal{O}(\mathcal{Q})} p(q, O, \lambda)[\log p(q, O, \lambda') + \log p(\lambda')] d\mu(q) \\ &= Q(\lambda, \lambda') + \int_{\mathcal{Q}\to\mathcal{O}(\mathcal{Q})} p(q, O, \lambda) \log p(\lambda') d\mu(q) \\ &= Q(\lambda, \lambda') + [\log p(\lambda')] P(\lambda|O). \end{aligned} \tag{5.3}$$

Since we maximize $Q'(\lambda, \lambda')$ with respect to λ' at each EM iteration, the value of the objective function $P(\lambda|O)$ here can be treated as a multiplicative constant factor on the regularization term $\log p(\lambda')$. In practice, due to the difficulty of calculating $P(\lambda|O)$ at each iteration we abstract it into a constant weighting factor ω which appears as part of the prior itself.

To review, the Q-function for the HMM which is maximized during each EM iteration is

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

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 $Q_3(B)$. We can regularize each of these separable functions in turn.

5.1 Initial and state transition terms

We note that when output distributions are identical, often related transition probabilities are zero or unity. In other words, if $b_i = b_j$, it is likely that $a_{kj} =$

$0, k = 1, \ldots, N$ (or vise versa). Although we consider the output distributions to be the dominating factor promoting state redundancy (a supposition supported by an early study on the relative parameter sensitivity of HMMs by Rabiner et. al. [RJL85]), we cannot neglect the initial and state-to-state transition probabilities as a factor. (Related analysis can be found in Chapter 7.) Since in any case biasing these probabilities away from zero and unity can be done with at little cost, we propose the Dirichlet-type unnormalized priors

$$P_1(\pi) = \prod_{i=1}^{N} \pi_i^{\omega_{Q_1}}, \tag{5.4}$$

$$P_2(A) = \prod_{i=1}^{N} \prod_{j=1}^{N} a_{ij}^{\omega_{Q_2}}, \tag{5.5}$$

where $\omega_{Q_1}, \omega_{Q_2} > 0$. This has the effect of adding log barrier regularization terms to Q_1 and Q_2 so that

$$Q_1'(\lambda, \lambda^{(k)}) = \sum_{i=1}^{N} \tau_{i1}^{(k)} \log \pi_i + \omega_{Q_1} \sum_{i=1}^{N} \log \pi_i, \tag{5.6}$$

$$Q_2'(\lambda, \lambda^{(k)}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij} + \omega_{Q_2} \sum_{i=1}^{N} \sum_{i=1}^{N} \log a_{ij}. \tag{5.7}$$

We see from this that $\omega_{Q_1}, \omega_{Q_2} > 0$ can be viewed as weighting terms. Our update rules for the EM iteration are then

$$\pi_i = \frac{\pi_i^{(k)} b_i^{(k)}(O_1) + \omega_{Q_1}}{\sum_{j=1}^{N} \pi_j^{(k)} b_j^{(k)}(O_1) + N\omega_{Q_1}}, \tag{5.8}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}^{(k)} + \omega_{Q_2}}{\sum_{t=1}^{T-1} \tau_{it}^{(k)} + N\omega_{Q_2}}, \tag{5.9}$$

and so the elements of π and A cannot lie in $\{0, 1\}$.

5.2 Output distribution terms

As we note in section 2.4.1.4, there is no general optimization procedure for output distributions. Likewise, there is also no general regularization term that 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. We present three candidate regularization terms, one for discrete output distributions based on the inner product, and two for Gaussian output distributions based, respectively, on the Mahalanobis and Euclidean distance metrics.

5.2.1 Discrete output terms

For discrete output distributions we propose a regularization term based on the inner product:

$$Q'_3 = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \tau_{it}^{(k)} \delta(O_t - m) \log b_i(m) - \omega_{Q_3} \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{m=1}^{M} b_i(m) b_j(m) \quad (5.10)$$

for which the associated prior is

$$P_3(B) = \prod_{i=1}^{N}\prod_{j=1}^{N}\prod_{m=1}^{M} e^{-\omega_{Q_3} b_i(m) b_j(m)} \quad (5.11)$$

and where $\omega_{Q_3} > 0$ is a weighting factor. We can find the update rule for this modified function by constructing the Lagrangian

$$\begin{aligned} L_{Q'_3}(b, \nu) = &\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \tau_{it}^{(k)} \delta(O_t - m) \log b_i(m) \\ &- \omega_{Q_3} \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{m=1}^{M} b_i(m) b_j(m) + \sum_{i=1}^{N} \nu_i \left(\sum_{m=1}^{M} b_i(m) - 1 \right), \end{aligned} \quad (5.12)$$

and taking its derivative:

$$\frac{\partial L_{Q'_3}}{\partial b_i(m)} = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{b_i(m)} - 2\omega_{Q_3} \sum_{j=1}^{N} b_j(m) + \nu_i. \quad (5.13)$$

Setting this equal to zero we have

$$b_i(m) = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{-\nu_i + 2\omega_{Q_3} \sum_{j=1}^{N} b_j(m)}, \tag{5.14}$$

so by summing both sides over m we get

$$\sum_{m=1}^{M} \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{-\nu_i + 2\omega_{Q_3} \sum_{j=1}^{N} b_j(m)} = 1 \tag{5.15}$$

and so we have $NM + N$ equations for a like number of unknowns. We note that

$$\frac{\partial^2 Q_3'}{\partial b_i(m) \partial b_j(k)} = \begin{cases} -\frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{b_i^2(m)} - 2\omega_{Q_3} & \text{if } i = j \text{ and } m = k \\ -2\omega_{Q_3} & \text{if } i \neq j \text{ and } m = k \\ 0 & \text{otherwise} \end{cases} . \tag{5.16}$$

For the function to be concave, the Hessian H with partial derivative elements presented in equation (5.16) must be negative definite. In other words,

$$\sum_{m=1}^{M} \sum_{i=1}^{N} \left(-x_{i,m}^2 \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{b_i^2(m)} - 2\omega_{Q_3} x_{i,m} \sum_{j=1}^{N} x_{j,m} \right) < 0, \text{for all } x \neq 0 \tag{5.17}$$

We note that $\sum_{i=1}^{N} x_{i,m} \sum_{j=1}^{N} x_{j,m} \leq N \sum_{i=1}^{N} x_{i,m}^2$. In addition, we also note that for $\sum_{i=1}^{N} x_{i,m} \sum_{j=1}^{N} x_{j,m} < 0$, $|\sum_{i=1}^{N} x_{i,m} \sum_{j=1}^{N} x_{j,m}| < N \sum_{i=1}^{N} x_{i,m}^2$. So we can reduce the condition (5.17) to

$$\sum_{m=1}^{M} \sum_{i=1}^{N} x_{i,m}^2 \left(2N\omega_{Q_3} - \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{b_i^2(m)} \right) \leq 0, \text{for all } x \neq 0, \tag{5.18}$$

which implies that the function is concave for

$$\omega_{Q_3} \leq \min_{i,m} \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{2N}, \tag{5.19}$$

as $0 \leq b_i(m) \leq 1$ for all i, m. This value is easily calculated at each EM iteration.

However, even given the concavity of Q_3', simultaneous solution of equations (5.14) and (5.15) is analytically difficult. From a practical implementation perspective, it is advantageous to use an iterative solver to find the maximizing

values. In our implementation, we alternately solve for the ν_is and $b_i(m)$s using a Newton method to find the roots of (5.15). We note that because we do not calculate the maximum directly that the resulting optimization is not in fact a true EM approach. In this manner it is similar to the work presented by Noumeir et. al. [NML95], in which an iterative Newton-Raphson method is used to find the maximum at each iteration. As in that work, we stop our iteration if the error is less than some limit and if $Q(\lambda^{(k+1)}|\lambda^{(k)}) > Q(\lambda^{(k)}|\lambda^{(k)})$. This latter condition ensures that the method is at least a generalized EM (GEM) method which guarantees convergence to a local maximum.

5.2.2 Gaussian output terms: Mahalanobis distance

We now present some candidate regularization terms designed for Gaussian output distributions. Our first regularization term is based on the Mahalanobis distance, and leads to the modified Q-function

$$\begin{aligned} Q'_3 &= \sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{it}^{(k)}\Bigg(\log n - \frac{1}{2}\log\det(\Sigma_i) - \frac{1}{2}(m_i-\mu_i)^T\Sigma_i^{-1}(m_i-\mu_i) \\ &\qquad - \frac{1}{2}(O_t-m_i)^T\Sigma_i^{-1}(O_t-m_i) \\ &\qquad + \frac{\omega_{Q_3}}{2}\sum_{j=1}^{N}(\mu_i-\mu_j)^T\Sigma_i^{-1}(\mu_i-\mu_j)\Bigg), \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\tau_{it}^{(k)}\Bigg(\log n - \frac{1}{2}\log\det(\Sigma_i) - \frac{1}{2}(O_t-m_i)^T\Sigma_i^{-1}(O_t-m_i) \\ &\qquad + \sum_{j=1}^{N}\frac{1}{2}\Big((\omega_{Q_3}-\frac{1}{N})\mu_i - \omega_{Q_3}\mu_j + \frac{1}{N}m_i\Big)^T\Sigma_i^{-1} \\ &\qquad\qquad \Big((\omega_{Q_3}-\frac{1}{N})\mu_i - \omega_{Q_3}\mu_j + \frac{1}{N}m_i\Big)\Bigg), \end{aligned} \tag{5.20}$$

where $\omega_{Q_3} > 0$ is a weighting factor and $m_i = \sum_{t=1}^{T} \tau_{it}^{(k)} O_t / \sum_{t=1}^{T} \tau_{it}^{(k)}$ as introduced in section 2.4.1.4. The prior in this case is

$$P_3(B) = \prod_{i=1}^{N} \prod_{j=1}^{N} \exp\Big(\frac{\omega_{Q_3}}{2}(\mu_i - \mu_j)^T \Sigma_i^{-1} (\mu_i - \mu_j)\Big). \tag{5.21}$$

Note that for the ease of subsequent manipulation and computation, we do not properly account for the independence of the regularizing prior from the hidden state variable. In theory this means that systems with highly skewed populations of observations from each of the output distributions will be misrepresented in this regularization scheme. In practice we have observed no evidence that this effect has serious consequence. Nevertheless, for completeness we present the modified Q-function resulting from the prior (5.21):

$$\begin{aligned} Q_3' = \sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \Bigg(& \log n - \frac{1}{2} \log \det(\Sigma_i) - \frac{1}{2}(m_i - \mu_i)^T \Sigma_i^{-1} (m_i - \mu_i) \\ & - \frac{1}{2}(O_t - m_i)^T \Sigma_i^{-1} (O_t - m_i) \\ & + \frac{\omega_{Q_3}}{2 \sum_{t=1}^{T} \tau_{it}^{(k)}} \sum_{j=1}^{N} (\mu_i - \mu_j)^T \Sigma_i^{-1} (\mu_i - \mu_j) \Bigg). \end{aligned} \tag{5.22}$$

We therefor keep in mind that the regularized Q-function (5.20) is only an approximation.

To find the update rule for the means, we perform direct maximization via the derivative. For ease of representation, we take vector derivatives, noting that for a real valued matrix A and column vector x,

$$\begin{aligned} \frac{\partial}{\partial x}(Ax) &= A, \\ \frac{\partial}{\partial x^T}(Ax) &= A^T, \\ \frac{\partial}{\partial x}(x^T A x) &= x^T (A + A^T), \\ \frac{\partial}{\partial x^T}(x^T A x) &= (A + A^T) x. \end{aligned} \tag{5.23}$$

We then have

$$\frac{\partial Q_3'}{\partial \mu_i^T} = \sum_{j=1}^{N} \Sigma_i^{-1}\Big(\omega_{Q_3} - \frac{1}{N}\Big)\Big((\omega_{Q_3} - \frac{1}{N})\mu_i - \omega_{Q_3}\mu_j + \frac{1}{N}m_i\Big), \tag{5.24}$$

and so we solve

$$(N\omega_{Q_3} - 1)\mu_i + m_i - \omega_{Q_3}\sum_{j=1}^{N}\mu_j = 0 \tag{5.25}$$

simultaneously for $i = 1, \ldots, N$. Let $U = (\mu_1 \cdots \mu_N)$ and $M = (m_1 \cdots m_N)$ be matrices formed from the column vectors μ_i and m_i, and let $1_{N\times N}$ be an N-by-N matrix of ones. The resulting system

$$U(\omega_{Q_3} 1_{N\times N} + (1 - N\omega_{Q_3})I) = M \tag{5.26}$$

can be solved by any standard linear method. Note that solving this system of equations gives us a critical point, which will only be the global maximum if the function is concave in the means μ_i. Taking the second derivative of the modified Q-function, we have

$$\frac{\partial^2 Q_3'}{\partial \mu_i^T \partial \mu_j} = \begin{cases} \Sigma_i^{-1}((N-1)\omega_{Q_3} - 1) & \text{if } i = j \\ -\Sigma_i^{-1}\omega_{Q_3} & \text{otherwise} \end{cases}. \tag{5.27}$$

So the Hessian is

$$H = (N\omega_{Q_3} - 1)I - \omega_{Q_3} 1_{N\times N}. \tag{5.28}$$

For the function to be concave, the Hessian must be negative definite. In other words,

$$(N\omega_{Q_3} - 1)\sum_{i=1}^{N} x_i^2 - \omega_{Q_3}\sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j < 0, \text{for all } x \neq 0. \tag{5.29}$$

We note that $\sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j \leq N\sum_{i=1}^{N} x_i^2$ and that for $\sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j < 0$, $|\sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j| < N\sum_{i=1}^{N} x_i^2$. So we can reduce the condition (5.29) to

$$\begin{aligned} (N\omega_{Q_3} - 1)\sum_{i=1}^{N} x_i^2 + N\omega_{Q_3}\sum_{i=1}^{N} x_i^2 &\leq 0, \\ (2N\omega_{Q_3} - 1)\sum_{i=1}^{N} x_i^2 &\leq 0, \text{for all } x \neq 0, \end{aligned} \tag{5.30}$$

and therefore the function is concave and has a global maximum for

$$\omega_{Q_3} \leq \frac{1}{2N}. \tag{5.31}$$

Now that we have an EM update rule (5.26) for the means, we can determine the update rule for the covariance matrices given the new means. Let

$$S_i' = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} (O_t - \mu_i^{(k+1)})(O_t - \mu_i^{(k+1)})^T}{\sum_{t=1}^{T} \tau_{it}^{(k)}}, \tag{5.32}$$

$$R_{ij} = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} (\mu_i^{(k+1)} - \mu_j^{(k+1)})(\mu_i^{(k+1)} - \mu_j^{(k+1)})^T}{\sum_{t=1}^{T} \tau_{it}^{(k)}}. \tag{5.33}$$

Then we can write

$$Q_3' = \sum_{t=1}^{T} \sum_{i=1}^{N} \tau_{it}^{(k)} \left(\log n + \frac{1}{2} \log \det(\Sigma_i^{-1}) - \frac{1}{2} \operatorname{Tr} \Sigma_i^{-1} S_i' + \frac{\omega_{Q_3}}{2} \sum_{j=1}^{N} \operatorname{Tr} \Sigma_i^{-1} R_{ij} \right). \tag{5.34}$$

Once again we can find the maximum directly via the derivative:

$$\frac{\partial Q_3}{\partial \{\Sigma_i^{-1}\}_{ab}} = \{\Sigma_i\}_{ba} - \left(\{S_i'\}_{ba} - \omega_{Q_3} \sum_{j=1}^{N} \{R_{ij}\}_{ba} \right) = 0. \tag{5.35}$$

From this we see that Q_3' has a critical point at

$$\Sigma_i = S_i' - \omega_{Q_3} \sum_{j=1}^{N} R_{ij}. \tag{5.36}$$

Since both the S_i's and the R_{ij}s are positive definite without restrictions, we cannot pick an $\omega_{Q_3} > 0$ such that the maximization problem is concave in the covariances. However, we also have the constraint that the covariance matrices must be positive definite. In the unmodified EM method, this constraint is implicit because the update step guarantees positive definiteness of the covariances. However, we can add it explicitly here to address this problem. We note that setting $\omega_{Q_3} = 0$ when updating the covariances is a reasonable compromise in practice. Nevertheless, this difficulty in maintaining positive definite covariance matrices suggests that using an alternate regularization function would be preferable.

5.2.3 Gaussian output terms: Euclidean distance

As a more practical alternative to the Mahalanobis distance based regularization term, we devise on based on the squared Euclidean distance. This form of regularization has a number of theoretical and computational advantages over using the Mahalanobis distance, and is what we use in our software implementation of the method. Nevertheless, it does lack the intuitive appeal of using the Mahalanobis distance, in which tight, low-variance output distributions exert less pressure on their neighbors than more diffuse ones. The modified Q-function in this case is

$$\begin{aligned} Q_3' = \sum_{i=1}^{N}\sum_{t=1}^{T} \tau_{it}^{(k)} \Bigg(& \log n - \frac{1}{2}\log\det(\Sigma_i) - \frac{1}{2}(m_i - \mu_i)^T \Sigma_i^{-1}(m_i - \mu_i) \\ & - \frac{1}{2}(O_t - m_i)^T \Sigma_i^{-1}(O_t - m_i) \\ & + \frac{\omega_{Q_3}}{2}\sum_{j=1}^{N}(\mu_i - \mu_j)^T(\mu_i - \mu_j)\Bigg). \end{aligned} \tag{5.37}$$

For which the associated prior is

$$P_3(B) = \prod_{i=1}^{N}\prod_{j=1}^{N} \exp\Big(\frac{\omega_{Q_3}}{2}(\mu_i - \mu_j)^T(\mu_i - \mu_j)\Big). \tag{5.38}$$

As in the case of the Mahalobis distance based prior, the regularized Q-function (5.37) is an approximation which ignores the fact that the prior is independent of the hidden variable. Once again, we found in practice that this theoretical inaccuracy had little impact in practice.

To find the update rule for the means we once again take the derivative in

the means:

$$\begin{aligned}\frac{\partial Q_3'}{\partial \mu_i^T} &= \Sigma_i^{-1}(m_i - \mu_i) + \sum_{j=1}^{N} \omega_{Q_3}\Big(\mu_i - \mu_j\Big) \\ &= \Sigma_i^{-1}(m_i - \mu_i) - \omega_{Q_3}\sum_{j=1}^{N}\mu_j + N\omega_{Q_3}\mu_i. \end{aligned} \tag{5.39}$$

Setting the derivative to zero we have

$$\Sigma_i^{-1}m_i + (N\omega_{Q_3}I - \Sigma_i^{-1})\mu_i = \omega_{Q_3}\sum_{j=1}^{N}\mu_j \quad \text{for } i = 1, \ldots, N, \tag{5.40}$$

which leads us to the system of equations

$$\left(\begin{bmatrix} \Sigma_1^{-1} & & \\ & \ddots & \\ & & \Sigma_N^{-1} \end{bmatrix} + \omega_{Q_3}\begin{bmatrix} I_{D\times D} & \cdots & I_{D\times D} \\ \vdots & \ddots & \vdots \\ I_{D\times D} & \cdots & I_{D\times D} \end{bmatrix} - N\omega_{Q_3}I_{ND\times ND}\right) U = \begin{bmatrix} \Sigma_1^{-1} & & \\ & \ddots & \\ & & \Sigma_N^{-1} \end{bmatrix} M, \tag{5.41}$$

which can be solved by any standard linear method given the covariances Σ_i^{-1}. We evaluate the conditions under which this solution is the global maximum by calculating the Hessian:

$$\frac{\partial^2 Q_3'}{\partial \mu_i^T \partial \mu_j} = \begin{cases} \omega_{Q_3}(N-1)I - \Sigma_i^{-1}((N-1) & \text{if } i = j \\ -\Sigma_i^{-1}\omega_{Q_3} & \text{otherwise} \end{cases}, \tag{5.42}$$

so

$$H = \begin{bmatrix} N\omega_{Q_3}I - \Sigma_1^{-1} & & \\ & \ddots & \\ & & N\omega_{Q_3}I - \Sigma_N^{-1} \end{bmatrix} - \begin{bmatrix} I_{N\times N} & \cdots & I_{N\times N} \\ \vdots & \ddots & \vdots \\ I_{N\times N} & \cdots & I_{N\times N} \end{bmatrix}. \tag{5.43}$$

If the Q-function is concave, then the Hessian H is negative definite, and so for all nonzero column vectors x composed of stacked $N \times 1$ vectors x_i (noting that Σ_i is positive definite for all i),

$$\begin{aligned} x^T H x &< 0 \\ \sum_{i=1}^{N} x_i^T \Big(N\omega_{Q_3} I - \Sigma_i^{-1} \Big) x_i - \omega_{Q_3} \sum_{i=1}^{N} x_i^T \sum_{j=1}^{N} x_j &< 0 \\ 2N\omega_{Q_3} \sum_{i=1}^{N} x_i^T x_i - \sum_{i=1}^{N} x_i^T \Sigma_i^{-1} x_i &\leq 0 \\ 2N\omega_{Q_3} \sum_{i=1}^{N} x_i^T x_i - \sum_{i=1}^{N} x_i^T x_i ||\Sigma_i^{-1}|| &\leq 0 \\ \omega_{Q_3} &\leq \frac{||\Sigma_i^{-1}||}{2N}. \end{aligned} \tag{5.44}$$

This gives us a condition on ω_{Q_3} for the Q-function to have a global maxima in the means. To find the maximum in the covariances, we take the derivatives of (5.37) in the components of Σ_i^{-1} and set them equal to zero, which gives us

$$\Sigma_i = \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} (O_t - \mu_i)(O_t - \mu_i)^T}{\sum_{t=1}^{T} \tau_{it}^{(k)}}. \tag{5.45}$$

This is a global maximum since the Q-function is concave as a function of the covariances Σ_i. To find the means and covariances we need to solve equations (5.41) and (5.45) simultaneously. This can be done by using the approximation $\Sigma_i = S_i$ in equation (5.41) as an initial guess and then iterating between equations (5.41) and (5.45) until the solution converges. In practice, it is often sufficient merely to approximate Σ_i as S_i when calculating the means without any attempt at iterative convergence whatsoever. As in the case of our discrete output regularization, this iterative rather than direct maximization forces us to characterize the method as a generalized EM algorithm rather than a pure EM approach. Nevertheless, provided we ensure that the Q-function never decreases in value, we can guarantee convergence to a local maxima.

We summarize various priors, their associated regularization terms and the corresponding constraints on the weighting terms necessary to guarantee concavity of the Q-function in tables 5.1 and 5.2. Note that for the Mahalanobis distance the constraint only guarantees concavity in the means.

Prior	$\prod_{i=1}^{N} \pi_i^{\omega_{Q_1}}$
Regularization term	$\omega_{Q_1} \sum_{i=1}^{N} \log \pi_i$
Prior	$\prod_{i=1}^{N} \prod_{j=1}^{N} a_{ij}^{\omega_{Q_2}}$
Regularization term	$\omega_{Q_2} \sum_{i=1}^{N} \sum_{i=1}^{N} \log a_{ij}$
Prior	$\prod_{i=1}^{N} \prod_{j=1}^{N} \prod_{m=1}^{M} e^{-\omega_{Q_3} b_i(m) b_j(m)}$
Regularization term	$-\omega_{Q_3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{m=1}^{M} b_i(m) b_j(m)$
Prior	$\prod_{i=1}^{N} \prod_{j=1}^{N} \exp\Big(\frac{\omega_{Q_3}}{2} (\mu_i - \mu_j)^T \Sigma_i^{-1} (\mu_i - \mu_j)\Big)$
Regularization term	$\sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \big(\frac{\omega_{Q_3}}{2} \sum_{j=1}^{N} (\mu_i - \mu_j)^T \Sigma_i^{-1} (\mu_i - \mu_j)\big)$
Prior	$\prod_{i=1}^{N} \prod_{j=1}^{N} \exp\Big(\frac{\omega_{Q_3}}{2} (\mu_i - \mu_j)^T (\mu_i - \mu_j)\Big)$
Regularization term	$\sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \big(\frac{\omega_{Q_3}}{2} \sum_{j=1}^{N} (\mu_i - \mu_j)^T (\mu_i - \mu_j)\big)$

Table 5.1: Priors and associated regularization terms.

Regularization Term	Weight Constraint
$\omega_{Q_1} \sum_{i=1}^{N} \log \pi_i$	$\omega_{Q_1} > 0$
$\omega_{Q_2} \sum_{i=1}^{N} \sum_{i=1}^{N} \log a_{ij}$	$\omega_{Q_2} > 0$
$-\omega_{Q_3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{m=1}^{M} b_i(m) b_j(m)$	$\omega_{Q_3} \leq \min_{i,m} \frac{\sum_{t=1}^{T} \tau_{it}^{(k)} \delta(O_t - m)}{2N}$
$\sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \big(\frac{\omega_{Q_3}}{2} \sum_{j=1}^{N} (\mu_i - \mu_j)^T \Sigma_i^{-1} (\mu_i - \mu_j)\big)$	$\omega_{Q_3} \leq 1/2N$
$\sum_{i=1}^{N} \sum_{t=1}^{T} \tau_{it}^{(k)} \big(\frac{\omega_{Q_3}}{2} \sum_{j=1}^{N} (\mu_i - \mu_j)^T (\mu_i - \mu_j)\big)$	$\omega_{Q_3} \leq \|\|\Sigma_i^{-1}\|\|/2N$

Table 5.2: Weight parameter constraints for various regularization terms.

CHAPTER 6

Results and Discussion

Having established the theoretical basis for our approach, we now present some results of using the combined deterministic annealing and regularization techniques to train hidden Markov models. Once again, we use our two test data sets as the basis for our discussion.

The results we present are for using the squared Euclidean distance based regularization described in the preceding chapter to train Gaussian output hidden Markov models. In our experiments, we do not fix the value of ω_{Q_3}, but instead set the value to the upper bound on convexity 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 not in fact a true EM optimization. However, our implementation does require that the log likelihood function decrease at every iteration and satisfies the requirements of a generalized expectation-maximization (GEM) method, guaranteeing convergence to a local maxima. To bound the log likelihood function above, we use $\omega_\Sigma = 10^{-6}$.

6.1 Experiment Design

For our experiments we ran each of seven training methods 1000 times on the test data for every model size $N = 1, \ldots, 10$. These training methods are (1) baseline EM, (2) deterministic annealing EM (DAEM) alone with annealing schedule

$\Delta\gamma = 0.1$, (3) DAEM alone with annealing schedule $\Delta\gamma = 0.01$, (4) DAEM alone with annealing schedule $\Delta\gamma = 0.001$, (5-7) DAEM with regularization with the same respective schedules as training methods 2-4. For each model size and training method, we counted the number of local maxima via the Hamming distance between state assignments. As well, we determined the maximum log likelihood across all 1000 training runs and gathered statistics of the log likelihood results, calculating the mean and standard deviation. For all experiments, we initialized the initial and state-to-state transition probabilities randomly from a uniform distribution. Gaussian means were also initialized randomly from a uniform distribution, but covariances were generated according to $\Sigma = Q^T W Q$, where W was a diagonal matrix with uniform random diagonal elements between zero and one, and Q was the "Q" portion of a QR transformation of a square matrix also with uniform random elements between zero and one. (If the random matrix was singular, new random matrices would be generated until a non-singular matrix was created.) In preparation for training, the test data sets were shifted and normalized along each dimension so that all observation values lay between zero and one; this was to prevent relative dimension scaling effects from dominating the training processes.

6.2 Synthetic Data Results

Figure 6.1 shows the results of the combined method on the data set `step`. We note that the combined method has fewer local maxima than the annealing method alone for all three annealing schedules – in fact, the combined method has superior performance with schedule $\Delta\beta = 0.01$ than the annealing method alone has with $\Delta\beta = 0.001$. In addition, the combined method allows for the use of more rapid annealing schedules than are possible with deterministic annealing

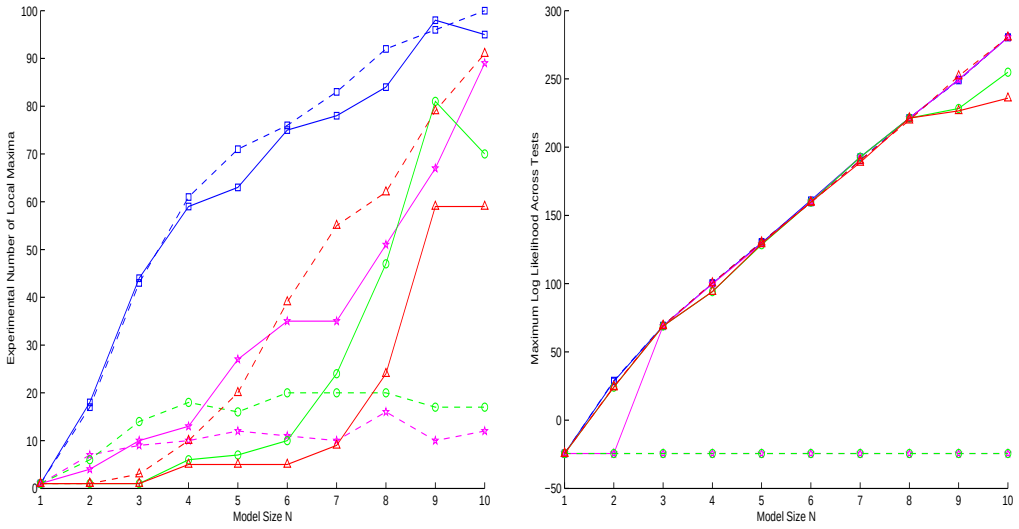


Figure 6.1: Experimental results for data set `step`. Left: Number of experimentally determined local maxima. Right: Maximum log likelihood among all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta stars with schedule $\Delta\gamma = 0.1$, green circles with schedule $\Delta\gamma = 0.01$; red triangles with schedule $\Delta\gamma = 0.001$. Dashed lines are the results with deterministic annealing only, solid lines are the results of the combined annealing and regularization technique.

alone, since the two slower schedules failed to produce useful results at all when used in isolation. We observe also that the maximum log likelihood across all trials of the combined method is very close to that found by the baseline EM method. This gives us some assurance that the method is producing reasonable solutions in the aggregate. Although the maximum log likelihood is lower for the combined method for nine and ten state models, this is unsurprising considering that the method does in fact optimize over a different objective function from either the EM or deterministic annealing EM methods. Figure 6.2 shows the mean log likelihood and the standard deviation of the log likelihood across all tests for each model size. These results confirm the advantages of the combined approach at the slowest annealing step, but also bring to light some difficulties

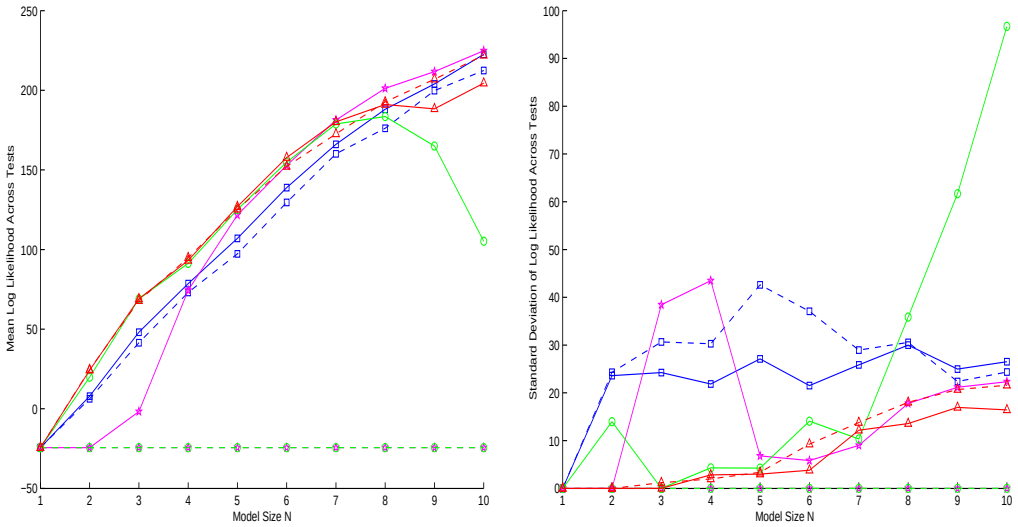


Figure 6.2: Test results for data set `step`. Left: Mean log likelihood across all HMMs. Right: Standard deviation of log likelihood across all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles with schedule $\Delta\beta = 0.01$; red triangles with schedule $\Delta\beta = 0.001$. Dashed lines are the results with deterministic annealing only, solid lines are the results of the combined annealing and regularization technique.

encountered by the combined method when $\Delta\gamma = 0.01$. This, combined with the poor results of deterministic annealing alone with the faster schedules, suggests that in general a conservative approach to annealing schedules is advisable.

Figure 6.3 compares classification results for a seven state model trained on `step` using the deterministic annealing method only (left) and the annealing with regularization (right). Each of these results is taken from the first of the 1000 tests performed on this data set. We see that with the annealing approach, the method is suffering from a local maxima in which class outputs are identical, with classes 1 and 4 unable to separate (what separation they do experience is due to the forced perturbation of identical output distributions that we enforce in our

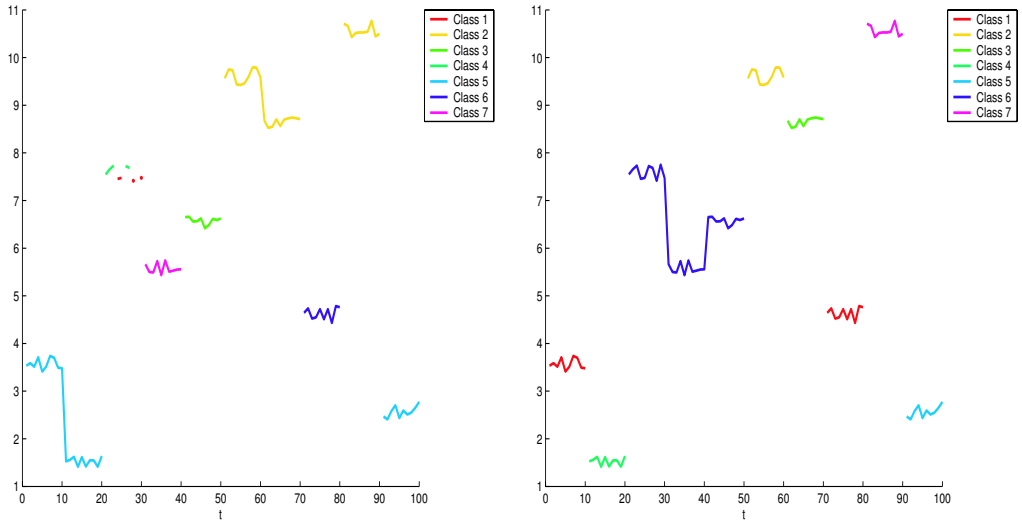


Figure 6.3: Left: Classification result from test 1 using deterministic annealing EM only for a seven state HMM trained on data set `step`. Right: Classification result from test 1 using deterministic annealing with regularization for a seven state HMM trained on data set `step`.

implementation).

We notice that as the number of model states rises above seven that even for the regularized deterministic annealing method at the slowest schedule the number of local maxima found rises rapidly. In figure 6.4 we present a sample local maxima for $N = 8$ (left) and one for $N = 9$ (right). What we see is that as in the example shown for $N = 8$ the local maxima are typically just the result of various assignments of the steps to states. We have $\binom{10}{8} = 45$ possible assignments, which is greater than the number of local maxima found. This is unsurprising, given that certain combinations of steps are likely to deprecated (for instance when a very low and very high step are in sequence). However, for $N = 9$, we have $\binom{10}{9} = 10$, so the large number of local maxima cannot be attributed to mere combinatorics. The example classification in figure 6.4 shows

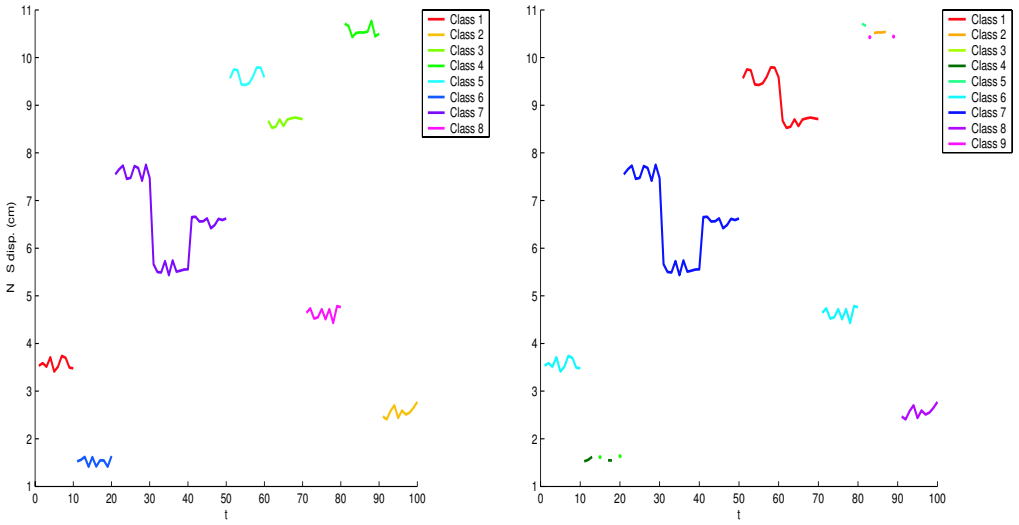


Figure 6.4: Left: Classification result from test 1 using regularized deterministic annealing EM for an eight state HMM trained on data set `step`. Right: Classification result from test 1 using the same method to train a nine state model.

a result very similar to that produced by the deterministic annealing method alone and indicates that the regularization term is becoming ineffective at this larger model size.

In order to address concerns that the EM method might be having difficulties because the data contains insufficient transition probability statistics, we repeated the tests described above on an extended data set of similar form. The data set, which we designate `stepstep`, was created by repeating the same pattern of discrete steps as in the data set `step` ten times, forming 1000 observations. Noise with the same statistics was then added to the signal. We present the results of our experiments in figure 6.5. We see that with this longer series that the performance of the regularized deterministic annealing EM method actually improves relative to standard EM, particularly with larger model sizes. Once again, the annealing method fails completely for the two fastest annealing schedules.

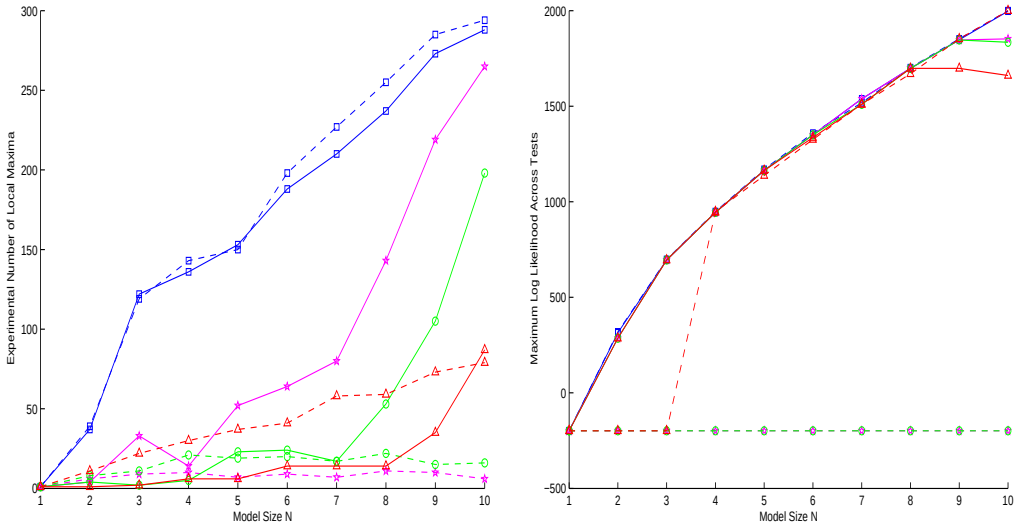


Figure 6.5: Experimental results for data set `stepstep`. Left: Number of experimentally determined local maxima. Right: Maximum log likelihood among all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta stars with schedule $\Delta\gamma = 0.1$, green circles with schedule $\Delta\gamma = 0.01$; red triangles with schedule $\Delta\gamma = 0.001$. Dashed lines are the results with deterministic annealing only, solid lines are the results of the combined annealing and regularization technique.

6.3 Field Data Results

In figure 6.6 we present results of the method as applied to the data set `clar`. We see for this data set that the results are similar in trend to those from the data set `step`. However, there are some key differences. In general, the number of local maxima for both the annealing alone and the combined method are lower than in the `step` case. In fact, we observe that for the combined method at the slowest annealing schedule there are three or fewer maxima for $N = 1, \ldots, 6$ and only eleven solutions for $N = 7$, three of which comprise 92% of the experimental results. 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

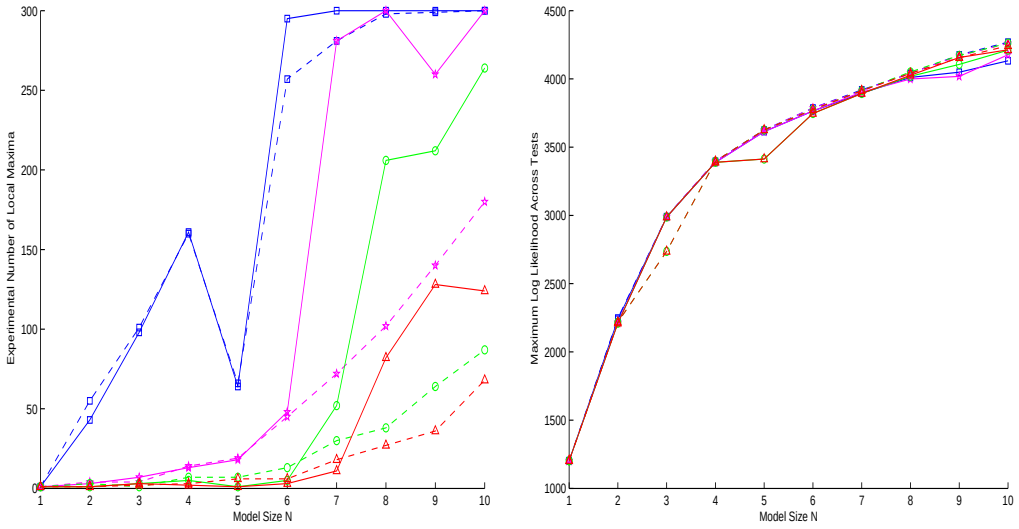


Figure 6.6: Test results for data set `clar`. Left: Number of experimentally determined local maxima. Right: Maximum log likelihood among all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles with schedule $\Delta\beta = 0.01$; red triangles with schedule $\Delta\beta = 0.001$. Dashed lines are the results with deterministic annealing only, solid lines are the results of the combined annealing and regularization technique.

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 6.7 shows the mean and standard deviation of the log likelihood results for these tests; these confirm our initial observations and parallel our observations on the performance of the method for the test data set `step`.

Figure 6.8 displays a classification result of the combined method for $N = 7$ on the slowest annealing schedule. We see that the method has identified all the major modes of the system including not only the before and after Hector

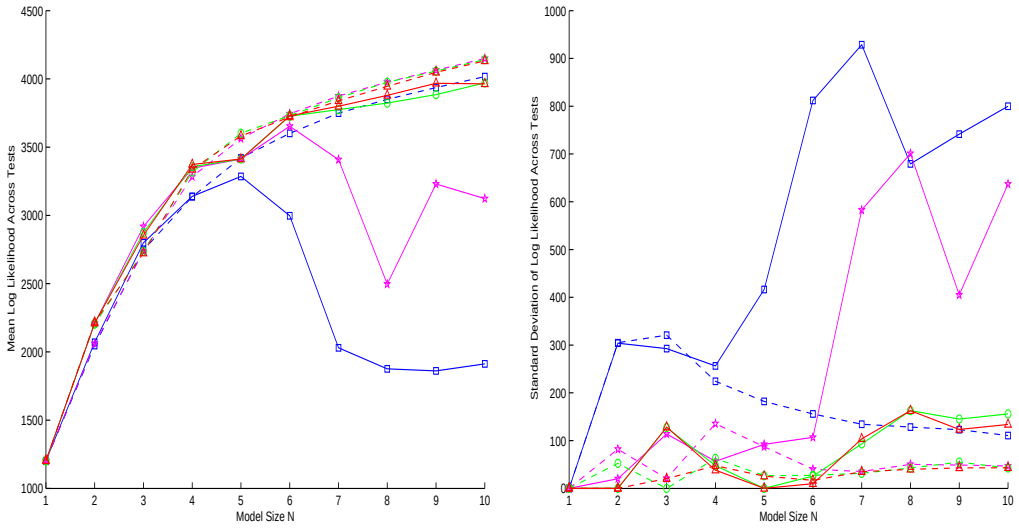


Figure 6.7: Test results for data set `clar`. Left: Mean log likelihood across all HMMs. Right: Standard deviation of log likelihood across all HMMs. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles with schedule $\Delta\beta = 0.01$; red triangles with schedule $\Delta\beta = 0.001$. Dashed lines are the results with deterministic annealing only, solid lines are the results of the combined annealing and regularization technique.

Mine earthquake states and the water pumping signal but also a number of more subtle signals. Figure 6.9 shows classification results for $N = 2, \ldots, 6$, again learned using the slowest annealing schedule. Each classification is the only one found by the method, since in each instance there was only a single maxima. Note the steady progression of subdivision of the various features that can be observed as the size of the model increases.

6.4 Altering Local Maxima Criteria

We also explored the effect of using a slightly less strict criteria for local maxima. In this scheme, we consider models with a Hamming distance between individually

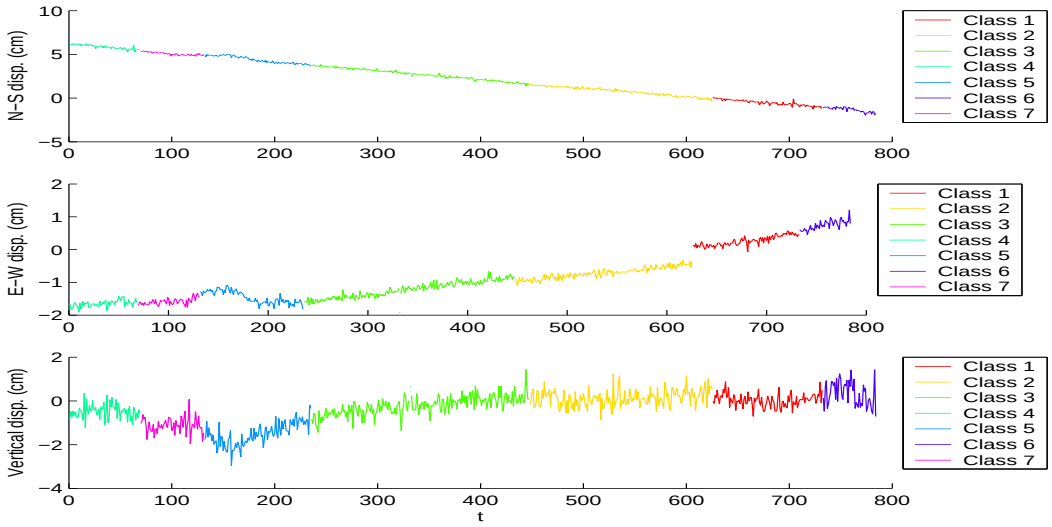


Figure 6.8: Classification results for a seven-state HMM applied to the data set `clar`.

most likely state assignment sequences less than or equal to one to be at the same local maxima. In other words, we make an exception for sequences that differ at a single time point. The results of these experiments are summarized in figure 6.10; we see that the results are essentially the same as those calculated by requiring strict equality.

6.5 Discussion

Our combined deterministic annealing and regularization scheme appears to offer significant advantages over both the baseline EM method and the deterministic annealing method alone. Nevertheless, there are some significant limitations to the method.

The deterministic annealing method tends to fail in cases where the high temperature maxima is significantly separated in model parameter space from the maxima of the original problem; the method tracks the early maxima smoothly

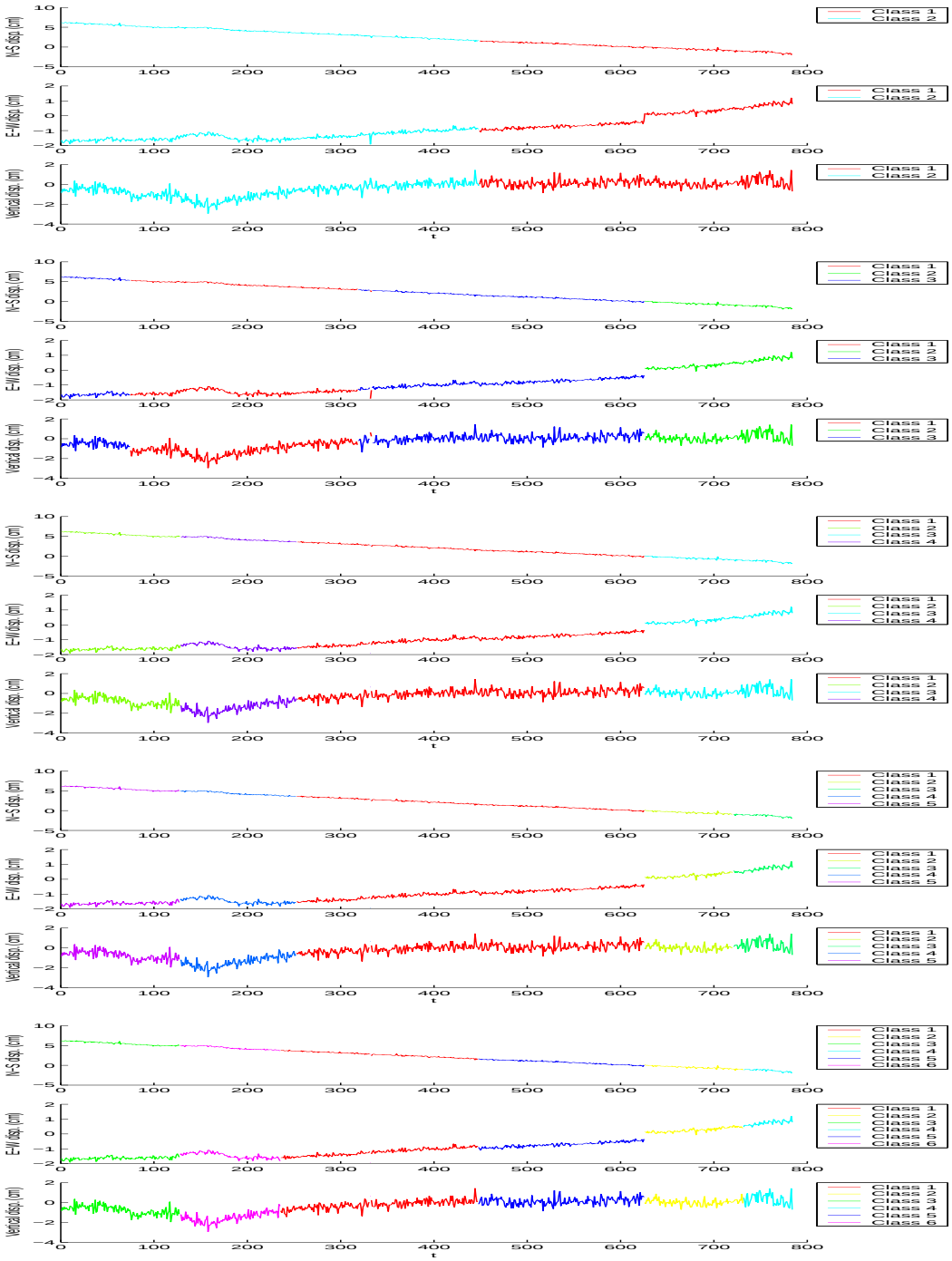


Figure 6.9: Classification results for a two- through six-state HMMs applied to the data set `clar`.

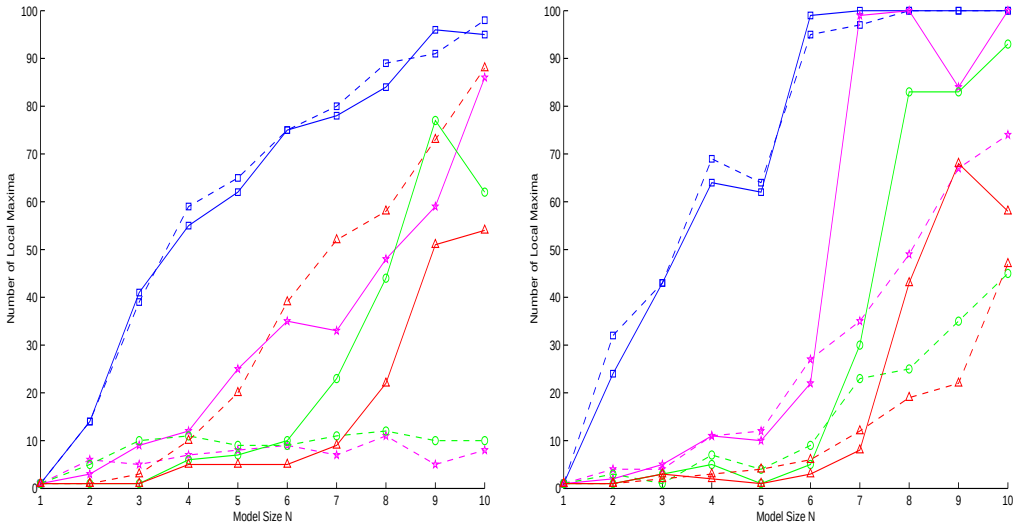


Figure 6.10: Left: Number of local maxima for different methods applied to `step` with relaxed local maxima criteria Hamming distance ≤ 1. Right: Number of local maxima for different methods applied to `clar` with relaxed local maxima criteria Hamming distance ≤ 1. Blue squares show results for the baseline HMM with standard EM optimization; magenta circles with schedule $\Delta\beta = 0.01$; red triangles 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.

and cannot jump to the more favorable solution. Our combined method shares this weakness, as the regularization term does nothing to address this. It is possible that methods designed to escape local maxima, such as weight annealing [ENF02] or split-and-merge methods [UNG00] could be combined with our approach to address this problem. However, these techniques remain untested on hidden Markov models and their interactions with deterministic annealing and regularization are uncertain.

We observe that regularization term does not actually solve the problem of identical output distributions, but merely discourages them. When identical out-

put distributions do occur, for instance in the early steps of the annealing process, we are obliged to perturb the distributions in order to allow the regularization term to work. To address this, in our implementation we check for identical outputs at each iteration and then randomly perturb the mean of one. Although these are small perturbations (by default, each element of the perturbation vector is of order 10^{-3}; observation data is dimension-wise shifted and normalized to lie between 0 and 1), they nevertheless violate the claim that the annealing procedure is in fact deterministic.

Our regularization scheme also carries with it the risk that it will modify the objective function too much and push the solution away from the true optimum of the original. Using small weighting terms for the regularization can reduce this risk but reduces the efficacy of the regularization itself. Our preference is to avoid fine-tuning of the weighting term on an application by application basis by instead setting the weight to the maximum at each iteration, as this promotes ease-of-use for non-expert users of the method. However, we recognize that there are likely to be cases in which this produces inferior results. As we remark in the preceding section in our discussion of the results for the data set `clar`, however, this seeming problem may in fact have the unexpected benefit of indicating the minimally representative number of states that describe the data. Systematic study of this phenomena will be necessary before we can come to any definite conclusions. Certainly, this phenomenon is not universal; we see that the number of local maxima found by the method for the data set `step` starts increasing rapidly well before the true number of states. These local maxima come about for a number if reasons. Some of them are the result of different subsets of the steps being combinatorially assigned to various states, while others are the result of idiosyncrasies of data.

Finally, we note that the method consumes a great deal more time than standard EM optimization. Although the regularization carries with it only a small amount of overhead, the annealing procedure can lengthen the computation time considerably. Since the deterministic annealing EM performs a standard EM maximization of the model parameters at each temperature, one might expect that the computation time would increase inversely proportional to the annealing schedule temperature step. However, the annealed cost function tends to be much smoother than the original, particularly at higher temperatures. This results in regions of the objective function that are quite flat; traversing such regions using EM can take a great deal of time. When not in such flat regions, the optimization procedure at a particular temperature tends to be much faster, but in general the difficult temperatures dominate the computation time, often requiring thousands of EM iterations to escape flat regions. As a result, the deterministic annealing EM and regularized deterministic annealing EM methods can easily take several orders of magnitude more time than the baseline EM method. Fortunately, there several solutions that present themselves. Although a cap on the number of EM iterations is not advisable, since that could prevent the optimization procedure from reaching the true maximum for a given temperature, variants and generalizations of the EM method designed to speed up convergence instead. These include SAGE [FH94], AECM [LR94], and generalized conjugate-gradient acceleration [JJ93]. In particular, the last of these seems well matched to our problem escaping flat regions of the parameters space. However, the effect of incorporating these methods lies beyond the scope of this work.

CHAPTER 7

Local Maxima

In Chapter 3 we discussed the problem of local maxima in the likelihood function of hidden Markov models and evaluated the extent of the problem on an experimental basis. In this chapter, we approach the problem of local maxima on an analytical basis. We construct local maxima of the HMM objective function in both the initial and state-to-state transition probabilities as well as in the output distributions. On result of this analysis is a confirmation that duplicated states are potentially a significant cause of problems in optimization, independent of the deterministic annealing EM algorithm. We do this by demonstrating that a lower bound on the number of locally maximum solutions with redundant states is exponential given certain assumptions about the observation sequence.

We begin our analysis by discussing the less pressing problem of local maxima in the initial and state-to-state transition probabilities, and then move on the discuss local maxima in the output distributions for first discrete and then continuous observation sequences.

7.1 Initial and Transition Probabilities

Our experimental observations of locally maximum solutions lead us to suspect that solutions with initial or state-to-state transition probabilities of zero or unity are associated with undesirable local maxima. We therefor 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 resultant state sequence determined by some particular π^*, A^* chosen from this set. Then

$$\alpha(i) = \begin{cases} b_{q_1^*}(O_1) \cdots b_{q_t^*}(O_t) & \text{if } i = q_t^* \\ 0 & \text{otherwise} \end{cases}, \tag{7.1}$$

and

$$\beta(i) = \begin{cases} b_{q_T^*}(O_T) \cdots b_{q_{t+1}^*}(O_{t+1}) & \text{if } i = q_t^* \\ 0 & \text{otherwise} \end{cases}, \tag{7.2}$$

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

$$\tau_{it} = \begin{cases} 1 & \text{if } i = q_t^* \\ 0 & \text{otherwise} \end{cases}, \tag{7.3}$$

and that

$$\tau_{ijt} = \begin{cases} 1 & \text{if } i = q_t^* \text{ and } j = q_{t+1}^* \\ 0 & \text{otherwise} \end{cases}. \tag{7.4}$$

From this we can derive the updates:

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

$$a_{ij}^{(k+1)} = \begin{cases} 1 & \text{if } a_{ij}^{(k)} = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{7.5}$$

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. However, many of these points in the parameter space will result in the same output sequence (for example, all sequences generated by $\pi_1 = 1$, $a_{11} = 1$ will be identical, regardless of the values of the other elements of A). The distinguishable solutions each generate a deterministic sequence of states, which

transitions from state to state until reaching a final repeating state i for which $a_{ii} = 1$; if no such state is reached, then the pattern simply repeats until the end of the observation sequence. We can count the number of sequences by noting that for every number of utilized states greater than one there are two possibilities: that the states will repeat or that the sequence will reach a final repeating state. For example, with two states used, we can either repeat (e.g., $Q = 121212\ldots$) or move immediately into the repeating state (Q = 122222...). Assuming that $T > N$, then we have $1 + \sum_{n=2}^{N}(n! + (n-1)!)$ possible sequences. However, if we ignore solutions which are permutations, we only have $2n - 1$ unique sequences. Our conclusion is then that these types of local maxima are not a serious problem for HMM optimization, and move on to consider local maxima in the output distributions.

7.2 Output Distribution Functions

7.2.1 Discrete Output

We first consider the case in which the observation sequence is discrete. For this data set let us propose a sequence of candidate values of the hidden underlying state variable, $Q^* = q_1^* \cdots q_T^*$, such that (1) $O_{t_1} = O_{t_2}$ implies that $q_{t_1}^* = q_{t_2}^*$ and (2) $q_t^* \neq q_{t+1}^*$ if and only if $O_t \neq O_{t+1}$. In other words, the state changes only when the observation sequence changes, and each state is associated with a unique set of observations. Figure 7.1 shows a simple example of such an observation sequence, with the state transitions marked by vertical lines.

We claim that a model $\lambda^* = \underset{\lambda}{\text{argmax}}\, P(O|\lambda, Q^*)$ is then a critical point of the likelihood function. Let S_i be the set of observation symbols associated with the state i, i.e. $q_t = i$ implies $O_t \in S_i$. Furthermore, let $L_m = \sum_{t=1}^{T} \delta(O_t - m)$ be the

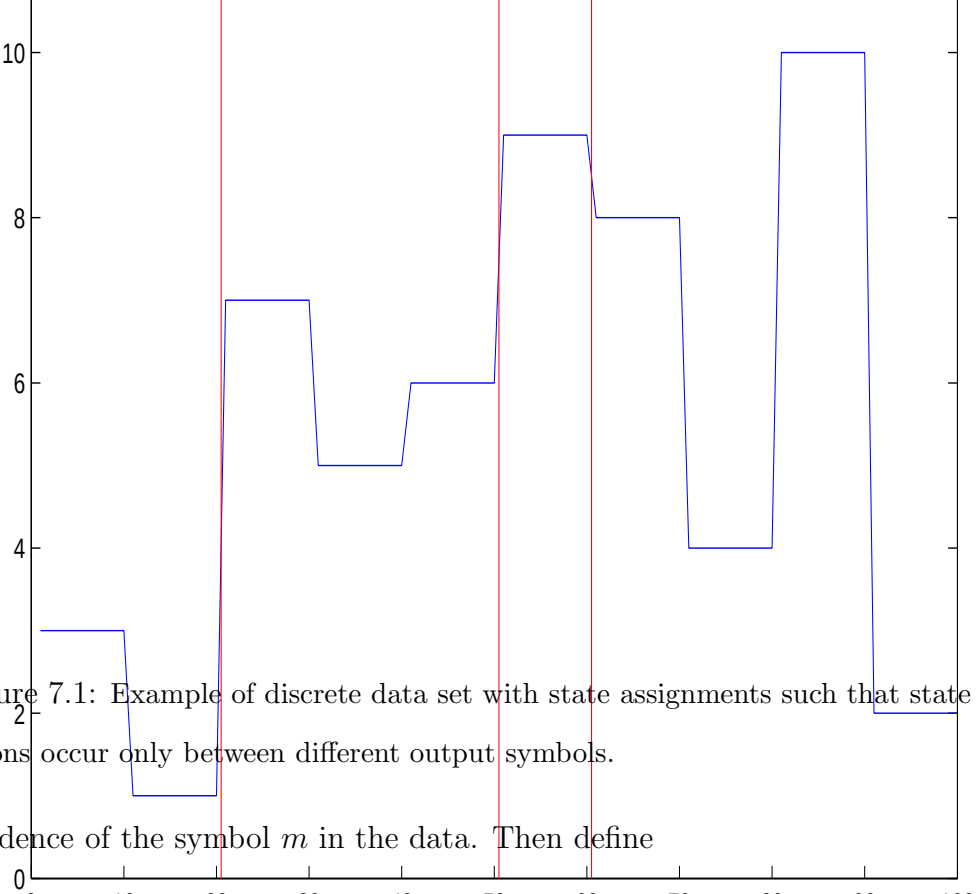


Figure 7.1: Example of discrete data set with state assignments such that state transitions occur only between different output symbols.

incidence of the symbol m in the data. Then define

$$\mathcal{L}_i = \sum_{m \in S_i} L_m. \tag{7.6}$$

For ease of presentation but without loss of generality, assume that the observations $O_t \in S_i$ are sequential for each state i. Then the locally maximum model λ^* is such that

$$\pi_i^* = \begin{cases} 1 & \text{if } O_1 \in S_i \\ 0 & \text{otherwise} \end{cases}, \qquad b_i^*(m) = \begin{cases} \frac{L_m}{\mathcal{L}_i} & \text{if } m \in S_i \\ 0 & \text{otherwise} \end{cases},$$
$$a_{ij}^* = \begin{cases} \frac{\mathcal{L}_i - 1}{\mathcal{L}_i} & \text{if } i = j \text{ and } O_T \notin S_i \\ \frac{1}{\mathcal{L}_i} & \text{if } O_t \in S_i \text{ and } O_{t+1} \in S_j \text{ for some } t \\ 1 & \text{if } i = j \text{ and } O_T \in S_i \\ 0 & \text{otherwise} \end{cases}. \tag{7.7}$$

For λ^* we have forward-backward parameters

$$\begin{aligned}
\alpha_1(i) &= \begin{cases} \frac{L_{O_1}}{\mathcal{L}_i} & \text{if } O_1 \in S_i \\ 0 & \text{otherwise} \end{cases}, \\
\alpha_{t+1}(i) &= \begin{cases} \alpha_t(i)\frac{\mathcal{L}_i-1}{\mathcal{L}_i}\frac{L_{O_{t+1}}}{\mathcal{L}_i} & \text{if } O_t, O_{t+1} \in S_i \\ \alpha_t(j)\frac{1}{\mathcal{L}_j}\frac{L_{O_{t+1}}}{\mathcal{L}_i} & \text{if } O_t \in S_j, O_{t+1} \in S_i \text{ for } i \neq j \\ 0 & \text{otherwise} \end{cases}, \\
\beta_T(i) &= 1, \\
\beta_t(i) &= \begin{cases} \beta_{t+1}(i)\frac{\mathcal{L}_i-1}{\mathcal{L}_i}\frac{L_{O_{t+1}}}{\mathcal{L}_i} & \text{if } O_t, O_{t+1} \in S_i \\ \beta_{t+1}(j)\frac{1}{\mathcal{L}_i}\frac{L_{O_{t+1}}}{\mathcal{L}_j} & \text{if } O_t \in S_i, O_{t+1} \in S_j \text{ for } i \neq j \\ 0 & \text{otherwise} \end{cases}. \qquad (7.8)
\end{aligned}$$

From these we calculate

$$\tau_{it} = \begin{cases} 1 & \text{if } O_t \in S_i \\ 0 & \text{otherwise} \end{cases}, \quad \tau_{ijt} = \begin{cases} 1 & \text{if } O_t \in S_i, O_{t+1} \in S_j \\ 0 & \text{otherwise} \end{cases}. \qquad (7.9)$$

Since this implies that our update rules are

$$\begin{gathered}
\pi_i^{(k+1)} = \frac{\tau_{i1}}{\sum_{j=1}^N \tau_{i1}} = \pi_i^*, \quad a_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}} = a_{ij}^*, \\
b_i(m)^{(k+1)} = \frac{\sum_{t=1}^{T} \tau_{it}\delta(O_t-m)}{\sum_{t=1}^{T} \tau_{it}} = b_i^*(m),
\end{gathered} \qquad (7.10)$$

we see that the model λ^* is a critical point of the EM method.

We show that this fixed point is a true local maximum of the log likelihood function by calculating the Hessian. We have

$$\begin{aligned}
\frac{\partial^2 \log P(O|\lambda^*)}{\partial \pi_i^2} &= -\frac{1}{{\pi_i^*}^2}, \\
\frac{\partial^2 \log P(O|\lambda^*)}{\partial a_{ii}^2} &= -\frac{\mathcal{L}_i}{{a_{ii}^*}^2}, \\
\frac{\partial^2 \log P(O|\lambda^*)}{\partial a_{ij}^2} &= -\frac{1}{{a_{ij}^*}^2} \text{ if } O_t \in S_i \text{ and } O_{t+1} \in S_j \text{ for some } t, \\
\frac{\partial^2 \log P(O|\lambda^*)}{\partial b_i(m)^2} &= -\frac{L_m}{b_i^*(m)^2} \text{ if } m \in S_i, \qquad (7.11)
\end{aligned}$$

and all the 2nd-order cross derivative terms are equal to zero. Since the Hessian is diagonal with exclusively negative elements, it is negative definite and the solution a local maximum.

As an illustrative example, consider the sequence $O = 112233$ of length $T = 6$, on which we train a model of size $N = 2$. Consider

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

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

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

which correspond respectively to the optimal model parameters for presumed underlying state sequences $Q_1 = 122222$, $Q_2 = 112222$, and $Q_3 = 111222$. 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 with model parameters corresponding to $Q = 111122$, as well as a third local maximum corresponding to the entire sequence being in the same state. We ignore an additional three local maxima which are morphologically equivalent to these.

We note that for S unique segments there are $\binom{S-1}{N-1}$ local maxima of the form λ^* utilizing all N states, since we choose $N-1$ of the $S-1$ possible transitions between segments as our state transition points. We futher note that this same analysis holds true for all models for which less than the full number of states are utilized. In these cases we have duplicate output distributions so that $b_i = b_j$ for

certain $i \neq j$. Including these there are $\sum_{n=1}^{N} \binom{S-1}{n-1}$ local maxima of this form 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.

7.2.2 Continuous Output

In our discussion of hidden Markov models with continuous output distributions, we first consider a relatively simple case in which the underlying signal is generated by an R-state HMM $\lambda^{\dagger}$ with output distributions $B^{\dagger}$ such that if $b_i^{\dagger}(O_t, \theta_i^{\dagger}) > 0$ then $b_j^{\dagger}(O_t, \theta_j^{\dagger} + \delta\theta_j^{\dagger}) = 0$ for $i \neq j$, where $\delta\theta_j^{\dagger}, ||\delta\theta_j^{\dagger}|| \geq 0$ is any small perturbation of the output distribution parameters. In other words, the output distributions have distinct and separated domains. We will see that our analysis of this case parallels that of the discrete output distribution case.

For the series of observations generated by this HMM, we consider as our solution an N-state HMM for which the output distribution of a given state b_i is a finite mixture of the output distributions of the generating HMM $\lambda^{\dagger}$. That is,

$$b_i = \sum_{r=1}^{R} w_{ir} b_r^{\dagger}. \tag{7.12}$$

For $N = R$, we propose a sequence of candidate values of the underlying state variable $Q^* = q_1^* \cdots q_T^*$ such that $b_i^{\dagger}(O_t) > 0$ implies that $q_t^* = i$. We then claim that a model $\lambda^* = \underset{\lambda}{\operatorname{argmax}}\ P(O|\lambda, Q^*)$ is then a critical point of the likelihood function. Let S_i be the set of observations associated with the state i, i.e. $q_t = i$ implies $O_t \in S_i$. Furthermore, let $\mathcal{L}_i = |S_i|$. For ease of presentation but without loss of generality, assume that the observations $O_t \in S_i$ are sequential for each

state i. Then the locally maximum model λ^* is such that

$$\pi_i^* = \begin{cases} 1 & \text{if } O_1 \in S_i \\ 0 & \text{otherwise} \end{cases}, \qquad w_{ir}^* = \begin{cases} 1 & \text{if } i = r \\ 0 & \text{otherwise} \end{cases},$$
$$a_{ij}^* = \begin{cases} \frac{\mathcal{L}_i - 1}{\mathcal{L}_i} & \text{if } i = j \text{ and } O_T \notin S_i \\ \frac{1}{\mathcal{L}_i} & \text{if } O_t \in S_i \text{ and } O_{t+1} \in S_j \text{ for some } t \\ 1 & \text{if } i = j \text{ and } O_T \in S_i \\ 0 & \text{otherwise} \end{cases}. \tag{7.13}$$

For λ^* we have forward-backward parameters

$$\begin{aligned} \alpha_1(i) &= \begin{cases} b_i^\dagger(O_1) & \text{if } O_1 \in S_i \\ 0 & \text{otherwise} \end{cases}, \\ \alpha_{t+1}(i) &= \begin{cases} \alpha_t(i)\frac{\mathcal{L}_i - 1}{\mathcal{L}_i} b_i^\dagger(O_{t+1}) & \text{if } O_t, O_{t+1} \in S_i \\ \alpha_t(j)\frac{1}{\mathcal{L}_j} b_i^\dagger(O_{t+1}) & \text{if } O_t \in S_j, O_{t+1} \in S_i \text{ for } i \neq j \\ 0 & \text{otherwise} \end{cases}, \\ \beta_T(i) &= 1, \\ \beta_t(i) &= \begin{cases} \beta_{t+1}(i)\frac{\mathcal{L}_i - 1}{\mathcal{L}_i} b_i^\dagger(O_{t+1}) & \text{if } O_t, O_{t+1} \in S_i \\ \beta_{t+1}(j)\frac{1}{\mathcal{L}_i} b_j^\dagger(O_{t+1}) & \text{if } O_t \in S_i, O_{t+1} \in S_j \text{ for } i \neq j \\ 0 & \text{otherwise} \end{cases}. \end{aligned} \tag{7.14}$$

From these we calculate (following section 2.5.2)

$$\tau_{it} = \begin{cases} 1 & \text{if } O_t \in S_i \\ 0 & \text{otherwise} \end{cases}, \quad \tau_{ijt} = \begin{cases} 1 & \text{if } O_t \in S_i, O_{t+1} \in S_j \\ 0 & \text{otherwise} \end{cases},$$
$$\tau_{irt} = \begin{cases} 1 & \text{if } i = r \text{ and } O_t \in S_i \\ 0 & \text{otherwise} \end{cases}. \tag{7.15}$$

This implies that our update rules are

$$\pi_i^{(k+1)} = \frac{\tau_{i1}}{\sum_{j=1}^N \tau_{i1}} = \pi_i^*, \quad a_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}} = a_{ij}^*,$$
$$w_{ir}^{(k+1)} = \frac{\sum_{t=1}^T \tau_{irt}}{\sum_{t=1}^T \tau_{it}} = w_{ir}^*, \tag{7.16}$$

and so we see that the model λ^* is a critical point of the EM method. We can show that this is a local maximum by calculating the Hession as in section 7.2.1.

Now consider an alternate sequence of candidate values of the underlying state variable Q^* such that $b_i^\dagger(O_t) > 0$ implies that $q_t^* = i$ for $i = 1, \ldots, N-1$ and $b_R^\dagger(O_t) > 0$ implies that $q_t^* = N-1$. In other words, we are attributing observed outputs produced by states $R-1$ and R of the generating model to a single class. We then claim that a model $\lambda^* = \underset{\lambda}{\text{argmax}}\ P(O|\lambda, Q^*)$ is then a critical point of the likelihood function. Let S_i and $\mathcal{L}_i$ retain their definitions based on the preceding section. For ease of presentation but without loss of generality, assume that the observations $O_t \in S_i$ are sequential for each state i and that for some O_t, $O_t \in S_{N-1}$ and $O_{t+1} \in S_N$ (that is, states $N-1$ and N of the generating model are sequential). Then the locally maximum model λ^* is such that

$$\pi_i^* = \begin{cases} 1 & \text{if } O_1 \in S_i, 1 \leq i < N \\ 1 & \text{if } i = N-1 \text{ and } O_1 \in S_N \\ 0 & \text{otherwise} \end{cases},$$

$$w_{ir}^* = \begin{cases} 1 & \text{if } i = r, 1 \leq i < N-1 \\ \frac{\mathcal{L}_{N-1}}{\mathcal{L}_{N-1}+\mathcal{L}_N} & \text{if } i = r = N-1 \\ \frac{\mathcal{L}_N}{\mathcal{L}_{N-1}+\mathcal{L}_N} & \text{if } i = N-1, r = N \\ \frac{\mathcal{L}_{N-1}}{\mathcal{L}_{N-1}+\mathcal{L}_N} & \text{if } i = N, r = N-1 \\ \frac{\mathcal{L}_N}{\mathcal{L}_{N-1}+\mathcal{L}_N} & \text{if } i = r = N \\ 0 & \text{otherwise} \end{cases},$$

$$a^*_{ij} = \begin{cases} \frac{\mathcal{L}_i - 1}{\mathcal{L}_i} & \text{if } i = j, 1 \le i < N-1 \text{ and } O_T \notin S_i \\ \frac{\mathcal{L}_{N-1}+\mathcal{L}_N-1}{\mathcal{L}_{N-1}+\mathcal{L}_N} & \text{if } i = j = N-1 \text{ and } O_T \notin S_N \\ \frac{1}{\mathcal{L}_i} & \text{if } O_t \in S_i \text{ and } O_{t+1} \in S_j \text{ for some } t \\ \frac{1}{\mathcal{L}_{N-1}+\mathcal{L}_N} & \text{if } O_t \in S_N \text{ and } O_{t+1} \in S_j \text{ for some } t, j < N-1 \\ 1 & \text{if } i = j, 1 \le i < N \text{ and } O_T \in S_i \\ 1 & \text{if } i = j = N-1 \text{ and } O_T \in S_N \\ 1 & \text{if } i = j = N \\ 0 & \text{otherwise} \end{cases} . \quad (7.17)$$

For λ^* we have forward-backward parameters

$$\alpha_1(i) = \begin{cases} b^\dagger_i(O_1) & \text{if } O_1 \in S_i \text{ and } 1 \le i < N-1 \\ \frac{\mathcal{L}_{N-1}}{\mathcal{L}_{N-1}+\mathcal{L}_N} b^\dagger_{N-1}(O_1) & \text{if } O_1 \in S_{N-1} \text{ and } i = N-1 \\ \frac{\mathcal{L}_N}{\mathcal{L}_{N-1}+\mathcal{L}_N} b^\dagger_N(O_1) & \text{if } O_1 \in S_N \text{ and } i = N-1 \\ 0 & \text{otherwise} \end{cases} ,$$

$$\alpha_{t+1}(i) = \begin{cases} \alpha_t(i)\frac{\mathcal{L}_i-1}{\mathcal{L}_i} b^\dagger_i(O_{t+1}) & \text{if } O_t, O_{t+1} \in S_i \text{ and } 1 \le i < N-1 \\ \alpha_t(i)\frac{\mathcal{L}_{N-1}+\mathcal{L}_N-1}{\mathcal{L}_{N-1}+\mathcal{L}_N}\frac{\mathcal{L}_{N-1}}{\mathcal{L}_{N-1}+\mathcal{L}_N} b^\dagger_{N-1}(O_{t+1}) & \text{if } O_t, O_{t+1} \in S_{N-1} \text{ and } i = N-1 \\ \alpha_t(i)\frac{\mathcal{L}_{N-1}+\mathcal{L}_N-1}{\mathcal{L}_{N-1}+\mathcal{L}_N}\frac{\mathcal{L}_N}{\mathcal{L}_{N-1}+\mathcal{L}_N} b^\dagger_N(O_{t+1}) & \text{if } O_t \in S_{N-1} \cup S_N, O_{t+1} \in S_N \text{ and } i = N-1 \\ \alpha_t(N-1)\frac{\mathcal{L}_{N-1}+\mathcal{L}_N-1}{\mathcal{L}_{N-1}+\mathcal{L}_N} b^\dagger_i(O_{t+1}) & \text{if } O_t \in S_N, O_{t+1} \in S_i \text{ for } 1 \le i < N-1 \\ \alpha_t(j)\frac{1}{\mathcal{L}_j}\frac{\mathcal{L}_{N-1}}{\mathcal{L}_{N-1}+\mathcal{L}_N} b^\dagger_{N-1}(O_{t+1}) & \text{if } O_t \in S_j, O_{t+1} \in S_{N-1} \text{ for } i = N-1, 1 \le j < N-1 \\ \alpha_t(j)\frac{1}{\mathcal{L}_j} b^\dagger_i(O_{t+1}) & \text{if } O_t \in S_j, O_{t+1} \in S_i \text{ for } i \ne j \\ 0 & \text{otherwise} \end{cases} ,$$

$$\beta_T(i) = 1,$$

$$\beta_t(i) = \begin{cases} \beta_{t+1}(i)\frac{\mathcal{L}_i - 1}{\mathcal{L}_i} b_i^\dagger(O_{t+1}) & \text{if } O_t, O_{t+1} \in S_i \\ & \text{and } 1 \leq i < N-1 \\ \beta_{t+1}(i)\frac{\mathcal{L}_{N-1}+\mathcal{L}_N-1}{\mathcal{L}_{N-1}+\mathcal{L}_N}\frac{\mathcal{L}_{N-1}}{\mathcal{L}_{N-1}+\mathcal{L}_N} b_{N-1}^\dagger(O_{t+1}) & \text{if } O_t, O_{t+1} \in S_{N-1} \\ & \text{and } i = N-1 \\ \beta_{t+1}(i)\frac{\mathcal{L}_{N-1}+\mathcal{L}_N-1}{\mathcal{L}_{N-1}+\mathcal{L}_N}\frac{\mathcal{L}_{N}}{\mathcal{L}_{N-1}+\mathcal{L}_N} b_{N}^\dagger(O_{t+1}) & \text{if } O_t \in S_{N-1} \cup S_N, O_{t+1} \in S_N \\ & \text{and } i = N-1 \\ \beta_{t+1}(j)\frac{\mathcal{L}_{N-1}+\mathcal{L}_N-1}{\mathcal{L}_{N-1}+\mathcal{L}_N} b_j^\dagger(O_{t+1}) & \text{if } O_t \in S_N, O_{t+1} \in S_j \\ & \text{for } 1 \leq j < N-1 \\ \beta_{t+1}(N-1)\frac{1}{\mathcal{L}_j}\frac{\mathcal{L}_{N-1}}{\mathcal{L}_{N-1}+\mathcal{L}_N} b_{N-1}^\dagger(O_{t+1}) & \text{if } O_t \in S_i, O_{t+1} \in S_{N-1} \\ & \text{for } j = N-1, 1 \leq i < N-1 \\ \beta_{t+1}(j)\frac{1}{\mathcal{L}_i} b_j^\dagger(O_{t+1}) & \text{if } O_t \in S_i, O_{t+1} \in S_j \text{ for } i \neq j \\ 0 & \text{otherwise} \end{cases}. \tag{7.18}$$

From these we calculate

$$\begin{aligned} \tau_{it} &= \begin{cases} 1 & \text{if } O_t \in S_i \text{ and } 1 \leq i < N \\ 0 & \text{otherwise} \end{cases}, \\ \tau_{ijt} &= \begin{cases} 1 & \text{if } O_t \in S_i, O_{t+1} \in S_j \text{ and } 1 \leq i < N, 1 \leq j < N \\ 0 & \text{otherwise} \end{cases}, \\ \tau_{irt} &= \begin{cases} 1 & \text{if } i = r,\ O_t \in S_i, \text{ and } 1 \leq i < N \\ 0 & \text{otherwise} \end{cases}. \end{aligned} \tag{7.19}$$

This implies that our update rules are once again

$$\begin{gathered} \pi_i^{(k+1)} = \frac{\tau_{i1}}{\sum_{j=1}^N \tau_{i1}} = \pi_i^*, \quad a_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}} = a_{ij}^*, \\ w_{ir}^{(k+1)} = \frac{\sum_{t=1}^{T} \tau_{irt}}{\sum_{t=1}^{T} \tau_{it}} = w_{ir}^*, \end{gathered} \tag{7.20}$$

and so we see that the model λ^* is also a critical point of the EM method. (Once again, we can show this is a local maximum as in section 7.2.1.) We also see that we can continue this process of attributing observations produced by multiple

states of the generating model to a single state of a trained model to generate more locally maximum solutions. As in the discrete output case, we therefor have $\sum_{n=1}^{N} \binom{R-1}{n-1}$ local maxima of this form for this data and model size N. If $R \geq N$, then $\sum_{n=1}^{N} \binom{R-1}{n-1} \geq 2^{N-1}$, so the lower bound on the number of local maxima is exponential in the model size.

CHAPTER 8

Science Applications

In this chapter we show the results of the regularized deterministic annealing EM method applied to the analysis of several geophysical data sets. We demonstrate that the method allows for innovative investigative approaches and that it offers significant advantages over the conventional EM algorithm. Our three application data sets all relate to the study of the Southern California earthquake fault system, but the general techniques we present are generally applicable to similar data sources.

8.1 SCIGN GPS

The Southern California Integrated Geodetic Network (SCIGN) is composed of over 250 global positioning system (GPS) stations that measure crustal deformation. In the future, it is expected that SCIGN will become part of the Plate Boundary Observatory (PBO), which will consist of over 1000 GPS stations spread across the western United States. The SCIGN stations integrate GPS position measurements daily in order to calculate displacement relative to the beginning of some epoch. Installation of SCIGN was completed in mid-2001, however our analysis focuses on data collected during a period somewhat over two years spanning 1998-2000, and so data from only 127 stations is available.

In Chapter 6 we presented results of our method applied to a single SCIGN

GPS station in Claremont, California. In this study, we are interested in detecting geophysical events with geographically disperse signatures and therefor 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 [MW03, RD03, MW02, MWS02, MMJ02, HHK99, HMT97], 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 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 were less than three local maxima for $N < 7$, eleven maxima for $N = 7$, and that the number of maxima rises 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.

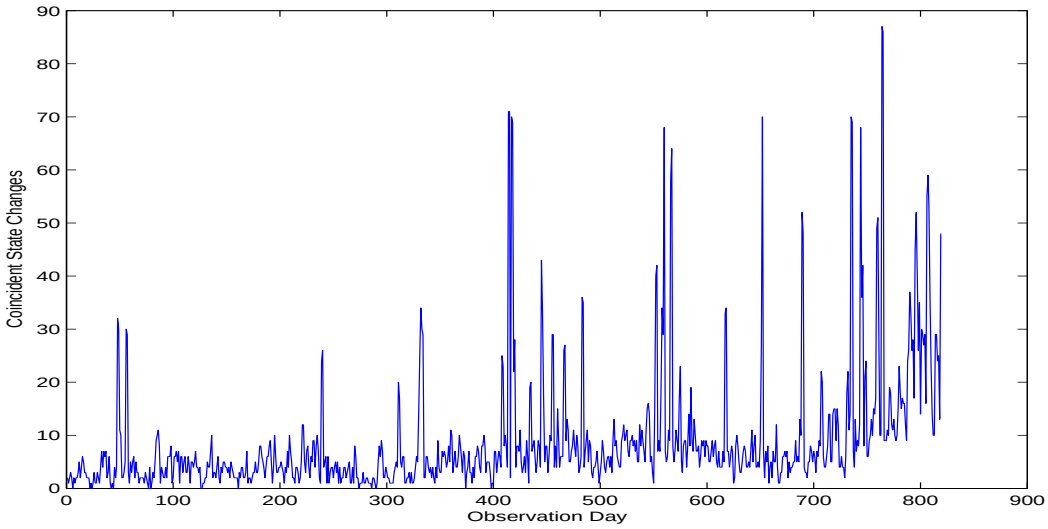


Figure 8.1: Coincident state changes for six-state HMMs trained on signals from each of 127 SCIGN GPS stations.

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 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)

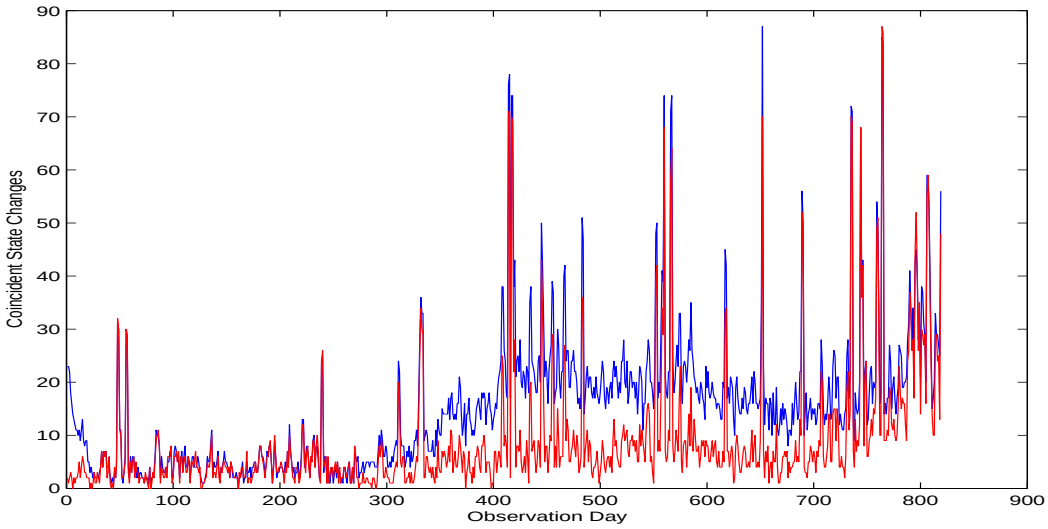


Figure 8.2: Coincident state changes for six-state HMMs trained using standard EM (blue) and regularized deterministic annealing EM (red) on signals from each of 127 SCIGN GPS stations.

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.

8.2 Seismicity

Our next application of the method is to Southern California seismicity data. This data is available from the Southern California Earthquake Center (SCEC) and consists of a record of earthquake times and locations from 1932 to the present, along with annotations describing the quality of event locations and magnitude estimates. Because of limitations in technology, early events are often unreliable and the catalog incomplete and so we only use data collected from

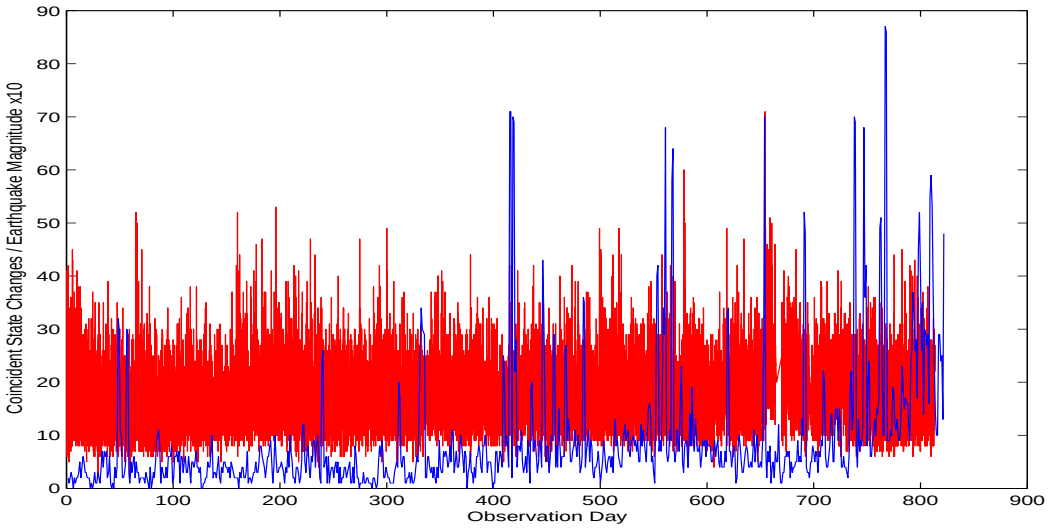


Figure 8.3: Comparison of coincident state changes for six-state HMMs trained using the regularized deterministic annealing EM (blue) and the Southern California earthquake record (red). Earthquake magnitudes, exaggerated by a factor of 10 for visibility, are presented on the vertical axis.

1960 onward. Even during this time period earthquake recording technology still limits the reliability of information and so we only consider events of magnitude 3.0 or above.

For the purpose of our investigation, we consider each earthquake to be a time series observation. Since events are not evenly spaced in time, we consider the series to be happening in "event time" and use the actual event time in calculating two derived attributes, the time since the previous event and the time to the next event. For convenience we use a time window of January 1st, 1960 to December 31st, 1999 and so the values of these derived time attributes are readily available for the first and last observations in the series. The other four attributes we consider for each observation are the x and y positions (calculated with respect to a standard reference latitude and longitude), the depth, and the magnitude.

We wish to use a hidden Markov model to classify the earthquake events into distinct classes of self-similar events and investigate relationships between classes indicated by the state-to-state transition probabilities. For this kind of work it is often useful to restrict analysis to events of magnitude four or above, since relationships between large events can otherwise be lost due to intervening smaller events resulting from background fault activity.

As an example of this kind of analysis, we present some results of a 25-state HMM applied to the time series of seismic events magnitude four and above in Southern California, neglecting time interval information. Figures 8.4-8.6 shows earthquake observations classified as belonging to three states of the HMM. For each state, the locations of earthquakes assigned to that state are shown as circles overlaid on a fault map of Southern California; the California coastline and the borders of the Salton Sea are shown in blue. Circle size corresponds to earthquake magnitude (rounded up). In figure 8.4 are shown earthquakes assigned to state 1. These are small, deep events associated with activity along the San Andreas fault network but biased towards the California coast. In figure 8.5 are shown earthquakes assigned to state 19. These are larger, shallower events readily identified as aftershocks of the San Simeon earthquake of 1952 (magnitude 6.0). In figure 8.6 are shown earthquakes assigned to state 22. These are earthquakes scattered around the northern part of the Sierra Nevada fault system, often occurring in sequence. Some may be aftershocks of the Owens Valley earthquake of 1980 (magnitude 6.2). Our particular interest in these three states lies in the fact that $a_{1,19} = 0.2864$ and $a_{1,22} = 0.7136$. In other words, events in state 1 are precursors to events in states 19 and 22. While this is natural for the two northernmost events in state 1, which precede events in state 19 and may be part of the Parkfield quake aftershock series, it is unusual that events occurring in the coastal part of the San Andreas fault system should precede events in the north-

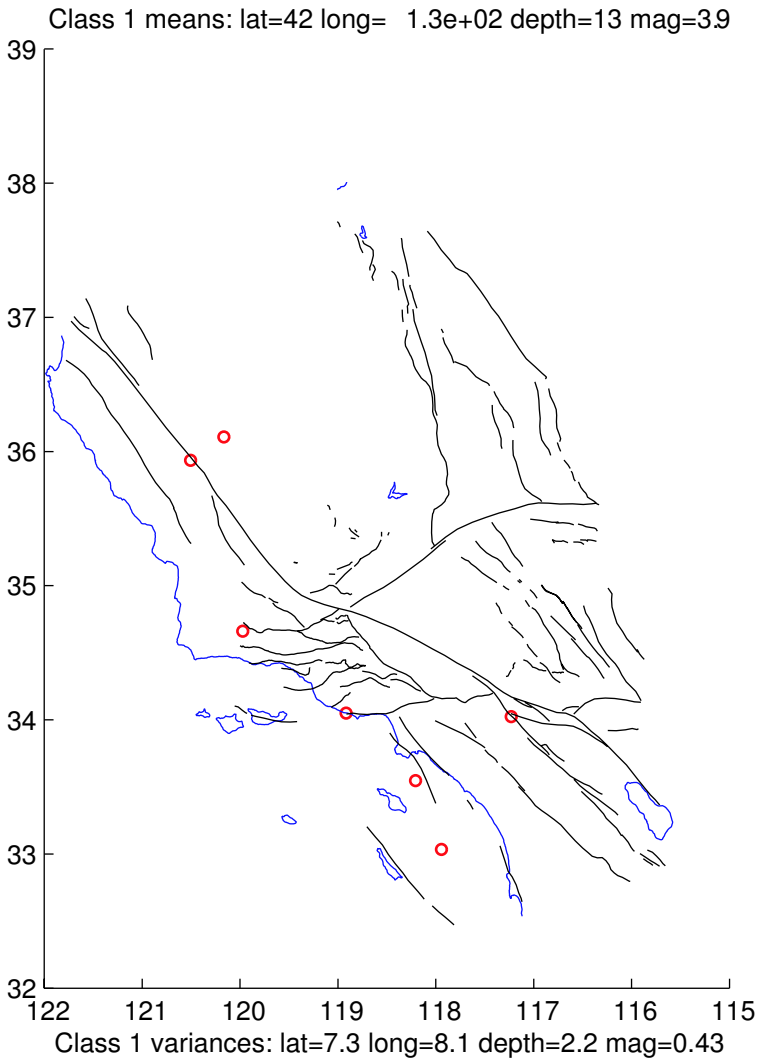


Figure 8.4: Earthquakes assigned to state 1 of a hidden Markov model trained on the Southern California seismic record (1960-1999).

ern Sierra Nevadas. This implies that there may be long range stress transfer between the two fault systems.

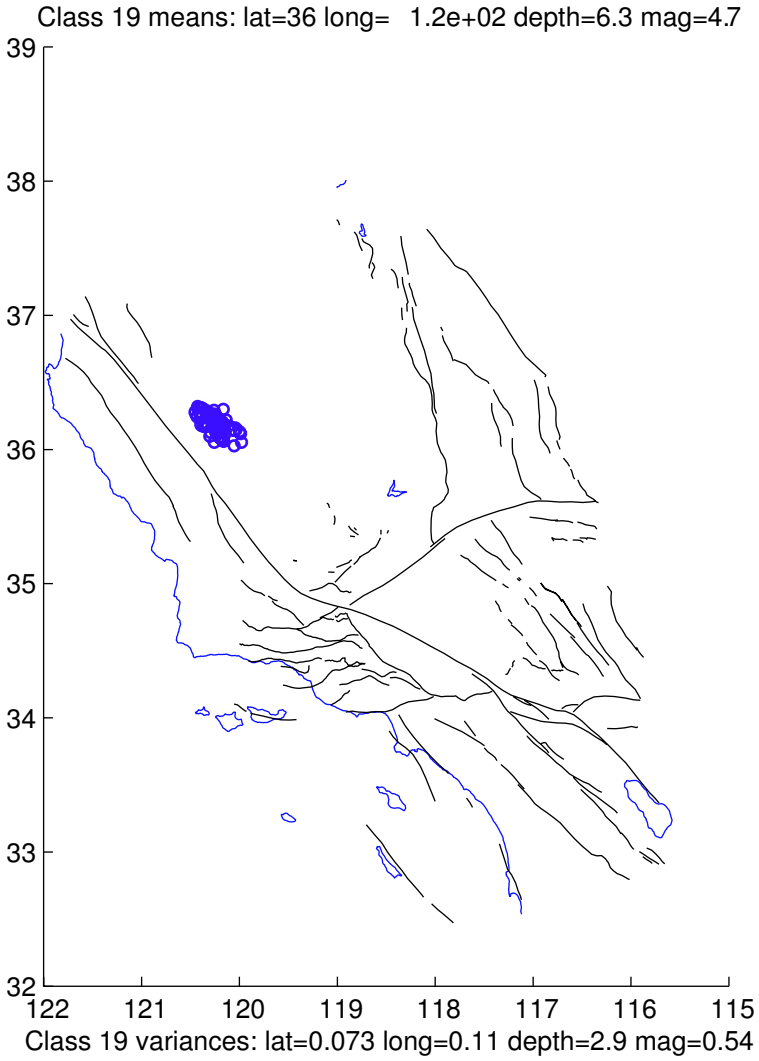


Figure 8.5: Earthquakes assigned to state 19 of a hidden Markov model trained on the Southern California seismic record (1960-1999).

8.3 Seismic Waveforms

The final geophysics application of the method we present is to waveform data collected by the Southern California TriNet broadband seismic network. The instruments in this array measure device "counts" corresponding to surface velocity

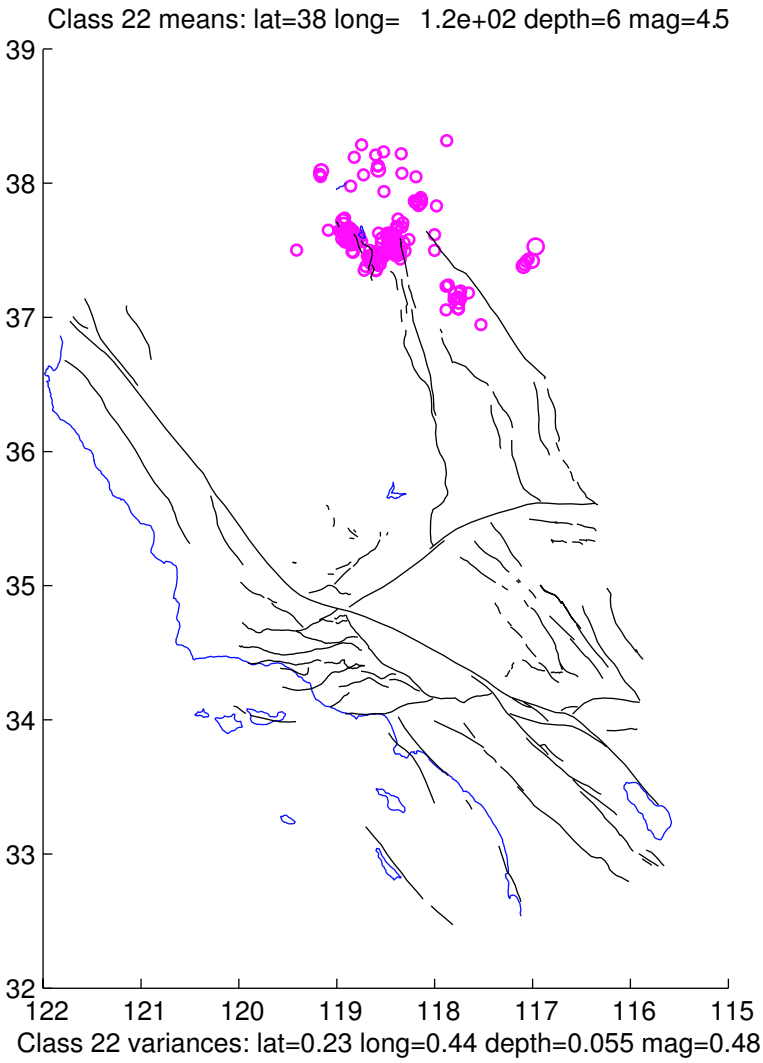


Figure 8.6: Earthquakes assigned to state 22 of a hidden Markov model trained on the Southern California seismic record (1960-1999).

at the instrument location in each of three directions. During an earthquake these sensors record velocity waveforms, typically characterized by a strong peak followed by ringing harmonics and secondary peaks caused by different seismic wave types or by reflected and/or diffracted versions of the waves. However, even in the absence of a seismic event these instruments constantly record ground mo-

tion. Because of the large volume of data, which is collected at up to 20Hz, the data collected between earthquake events has been largely unexplored. Our study focuses on examining this part of the data, with the intent being to identify and catalog aseismic events that occur on a minutes to hours time scale. These events can be seen most clearly in the frequency spectrum of the seismic signal. Figure 8.7, reproduced from [KK96], shows an example of the kind of signal in which we are interested. The low-frequency signal, labeled III, appears immediately before the Landers earthquake of 1992. This example also provides us with additional motivation in this study since it implies that some aseismic events may be earthquake precursors (see also [MW02]). We note that other, non-geophysical events of interest can also be observed in this data. For example in [TKA04] evidence is presented of slow traveling atmospheric pressure wave that can be observed in data from the TriNet seismic array.

Our approach is to find examples of unusual events manually and use them to train a hidden Markov model, which is in turn used to search through the stored data archives for similar events. This will enable the creation of an event catalog that can be used for the study of aseismic events and their role in the earthquake cycle. As the first step in this endeavor, we test the HMM training method on selected example events.

For this work we used TriNet instrument data sampled at 10Hz and low-pass filtered at 1Hz. For a given window of interest, a spectrogram of each dimension of the data was produced using a window 1024 samples in length, stepped through the time series at 128 sample intervals. The result of this procedure was a set of three time series with 1/128th the number of observations of the original signal and values at each of 512 frequencies for each observation. To reduce the dimensionality of the data, the singular value decomposition of each of the three time

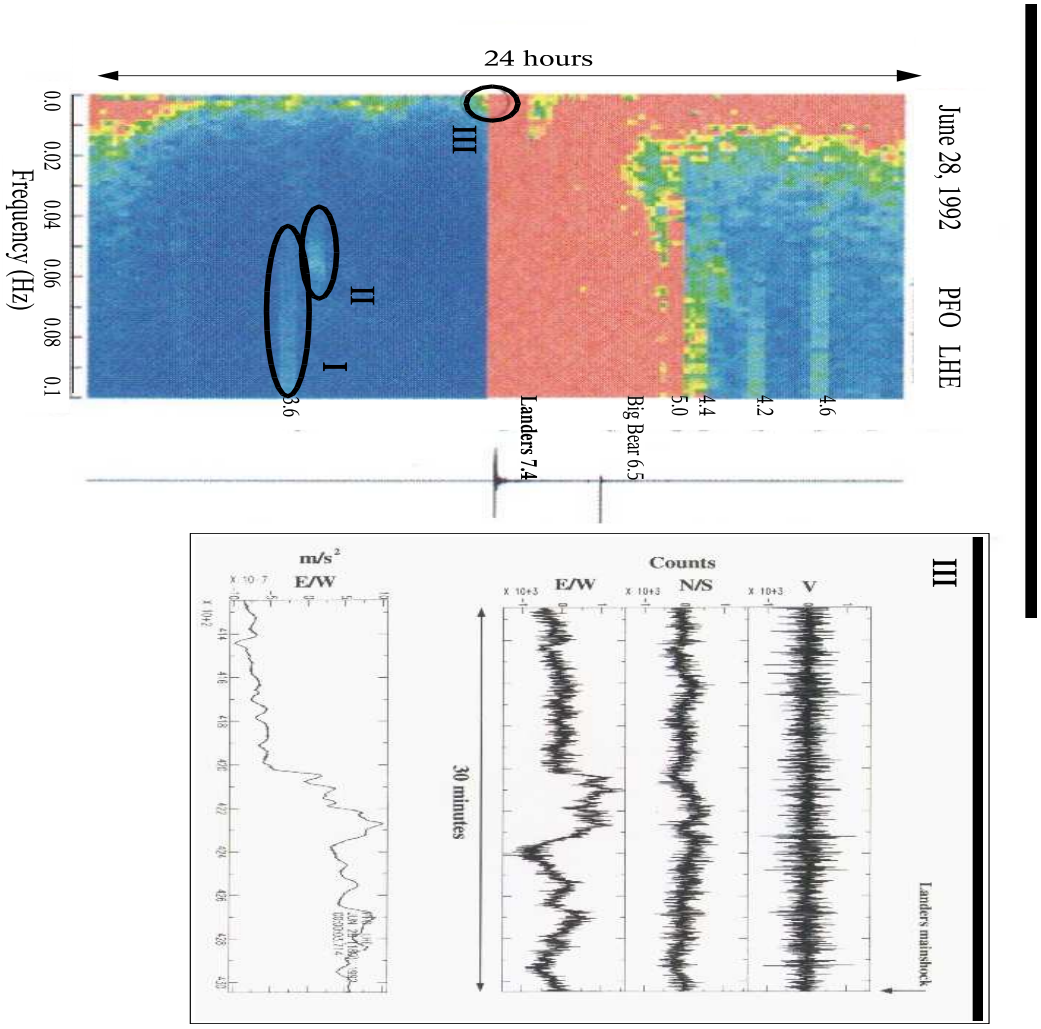


Figure 8.7: On the left is a time-spectra plot (spectrogram) showing the Landers earthquake of 1992. Key signals appearing before the earthquake are circled. (I) is a local earthquake, (II) is a teleseismic event, (III) is an aseismic event that may be the result of precursory activity. On the right are the actual TriNet time series which have been processed to form the spectrogram. A suspicious tilt signal appears approximately 20 minutes prior to the main shock.

series was calculated and the top 15 most significant components of each were extracted (the number 15 was selected according the knee in the singular values). These 15 component series were then combined to form a single 45 dimensional time series. This was the training data for the HMM.

In figure 8.8 we show individually most likely state assignment results for a hidden Markov model trained on data collected at a TriNet station in Pasadena, California in January 2001. The data, collected over approximately two days, was processed in the manner described in the preceding paragraph to provide a principle component spectrogram for training (45 dimensions, 12000+ observations). In the figure he state classes have been projected onto a 128:1 downsampled version of the original signal. The usual signal of interest is a period of low-frequency "fuzz" in the North-South and East-West dimensions that persists in the approximate period between samples 3000 and 7000 (about fourteen hours). Also visible are noise spikes in the vicinity of samples 5000, 7500, and 10000. The first and last of these spikes are the result of a known man-made noise source that appears in the recordings for this instrument on Monday through Friday mornings. The oscillation in the vertical component is the ocean tide. We see that the HMM, trained using our regularized deterministic annealing EM method, is able to successfully classify the background signal (red), the noise spikes (light and dark blue), and the signal of interest (as a combination of three states labeled with green, yellow, and purple). We contrast this result with the results of HMMs trained using the standard EM method and three different random parameter initializations presented in figure 8.9. We see that the standard EM method has a great deal of difficulty identifying the various signals in the time series and that the classification results for different initializations are quite different from one another.

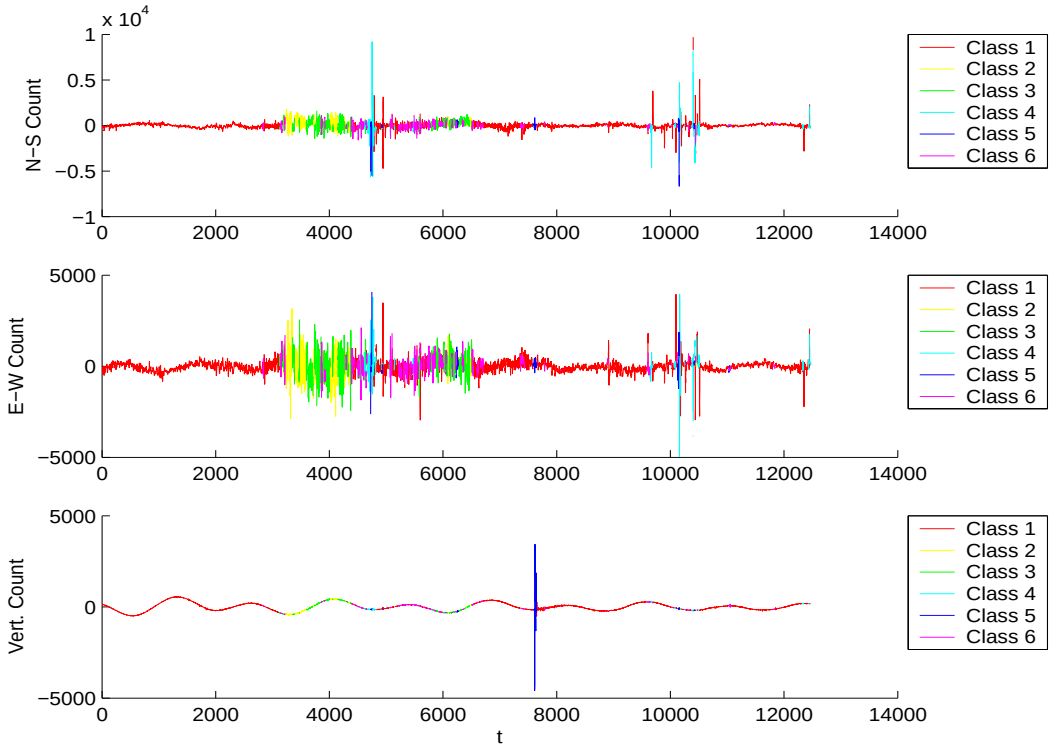


Figure 8.8: Results of application of an HMM trained using regularized deterministic annealing EM to an unusual long-duration signal in Pasadena, California. The HMM classifies the background signal as one class (red), noise spikes as two classes (light and dark blue), and the long-duration signal itself as a mixture of the remaining three classes.

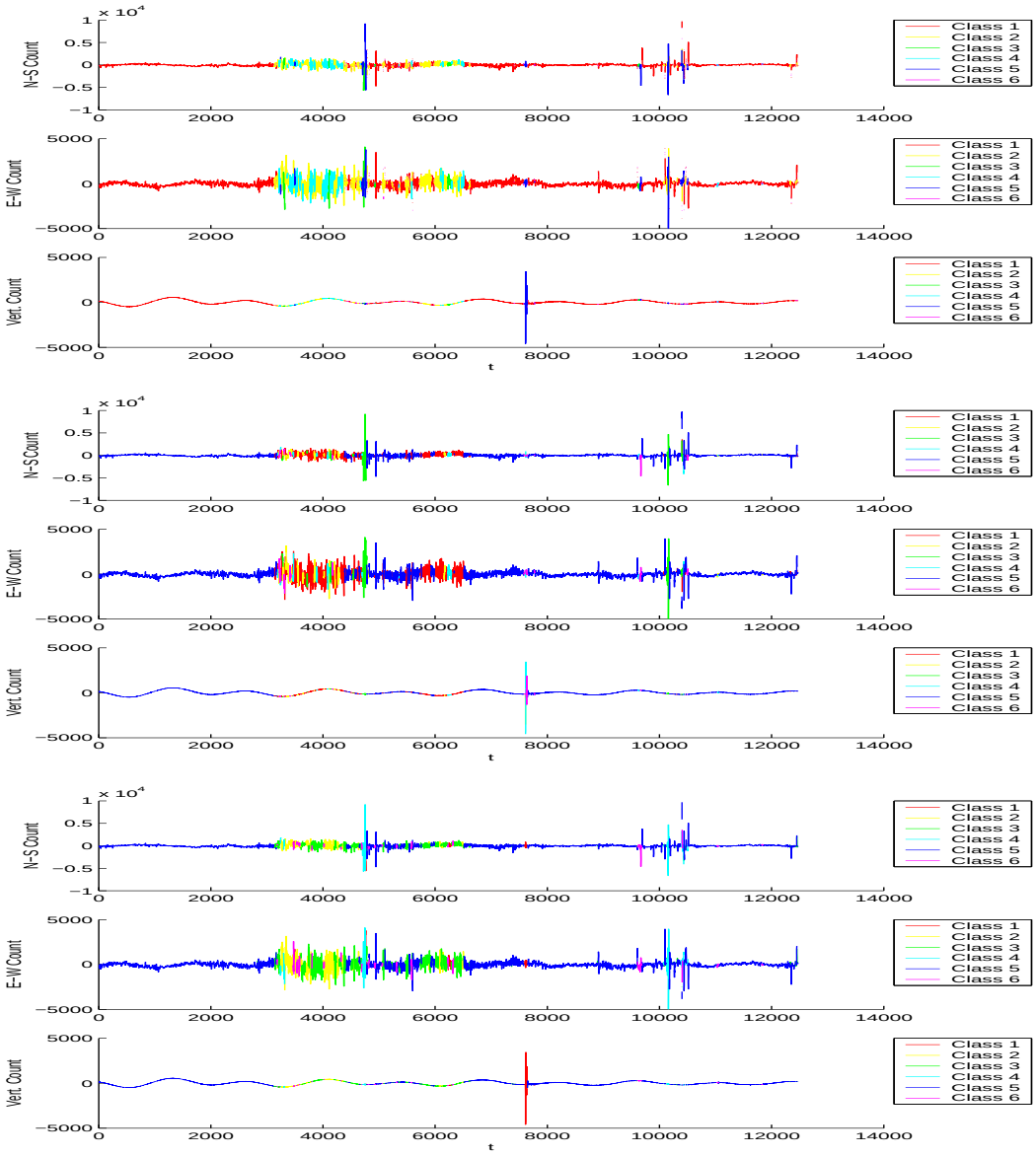


Figure 8.9: Three instances of the results of applying an HMM trained using standard EM to an unusual long-duration signal in Pasadena, California. Note the differences in classification results between the three different initializations.

CHAPTER 9

Conclusions

This work focused on the problem of applying hidden Markov models (HMMs) to exploratory analysis of scientific data sets. Unlike previous applications of HMMs, for instance to speech recognition or protein analysis, exploratory data analysis is by its very nature antithetical to the application of knowledge based constraints to the HMM optimization problem. In other application domains, such constraints are used to overcome the well-known local maxima problem of model optimization. In their absence, we devised an alternative optimization approach based on fundamental statistics and principles of efficient optimization.

Our first step was to verify the extent of the local maxima problem for hidden Markov models in the absence of constraints. To do this, we developed an empirical approach to estimating the number of local maxima for a given problem. Unlike previous approaches which employed log likelihood measures to differentiate between solutions, we focused on using the Hamming distance between solution state assignments to distinguish models and identify local maxima. This allowed us to avoid confusion in cases in which different models had similar likelihoods. Using this approach, we were able to show that even for simple problems the number of local maxima rises rapidly with the number of model states. In addition, we were able to perform theoretic analysis demonstrating that the number of HMM local maxima is exponential in the number of states for certain common data types.

To address this problem, we proposed the use of the deterministic annealing expectation-maximization (EM) method for the HMM optimization problem. This general optimization approach breaks down the original optimization problem into a series of sub-problems, each occurring at a successively cooler computational temperature. The computational temperature alters the original problem objective functions so that at the hottest temperature the objective function is flat, while at the coldest temperature it is equivalent to the original objective function. Gradual lowering of the temperature allows the solution to track important features while ignoring noise and surface complexity present in the original objective function. We showed that applying this method to HMMs yielded significant improvement over the basic EM method, but that deterministic annealing EM had a tendency to find locally maximum solutions in which there were redundant states.

In response to this difficulty, we developed several statistical priors designed to discourage solutions with redundant states. These worked by creating barriers in the objective function around such local maxima. In the modified EM algorithm, these priors manifested themselves as regularization terms added to the so-called Q-function maximized during the M-step. The effect of each regularization term was controlled by a tunable weighting parameter. Upper bounds on the weighting parameters guaranteeing concavity of the modified Q-function were calculated in each case.

Tests of the regularization method showed that while it had only minor effect used on its own, when combined with deterministic annealing EM its impact was very significant. Performance of the combined method on both our real and synthetic test data sets, as measured by empirical count of the number of locally maximum solutions, was considerably improved over the standard EM method.

Having demonstrated the success of the approach on our test data, we then applied the method to three geophysics data sets currently used in scientific investigation of earthquake fault interactions and the earthquake cycle. These three data sets were (1) surface displacement measurements collected by a network of GPS sensors in Southern California, (2) a 40-year record of seismic activity in Southern California, and (3) high-frequency surface velocity measurements collected by a network of broadband seismic stations in Southern California. In each case, we were able to show the success of the method in classifying observations and identifying previously unknown physical phenomena related to earthquake processes.

References

[ACK93] O. Arslan, P.D.L. Constable, and J.T. Kent. "Domains of convergence for the EM algorithm - a cautionary tale in a location estimation problem." *Statistics and Computing*, **3**(3):103–108, 1993.

[Bau72] L. E. Baum. "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes." *Inequalities*, **3**:1–8, 1972.

[BBS86] R. Bahl, P.F. Brown, P.V. De Souza, and R.L. Mercer. "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, 1986.

[BE67] 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. Math. Soc.*, **73**:360–363, 1967.

[BK93] J. Buhmann and H. Kuhnel. "Complexity optimized data clustering by competitive neural networks." *Neural Computation*, **5**:75–88, 1993.

[BM01] E. Bocchieri and B.K.W. Mak. "Subspace distribution clustering hidden Markov model." *IEEE Trans. on Speech and Audio Proc.*, **9**(3):264–275, 2001.

[BN90] J.R. Bellegarda and L.R. Nahamoo. "Tied mixture continuous parameter modeling for speech recognition." *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, **38**(12):2033–2045, 1990.

[BP66] L. E. Baum and T. Petrie. "Statistical inference for probabilistic functions of finite state Markov Chains." *Ann. Math. Stat.*, **37**:1554–1563, 1966.

[BPS70] 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. Stat.*, **41**(1):164–171, 1970.

[Bra99] M. Brand. "Structure learning in conditional probability models via an entropic prior and parameter extinction." *Neural Computation*, **11**:1155–1182, 1999.

[BS68] L. E. Baum and G. R. Sell. "Growth functions for transformations on manifolds." *Pac J Math*, **27**(2):211–227, 1968.

[Bun94] W. Buntine. "Operations for learning with graphical models." *Journal of Artificial Intelligence Research*, **2**:159–225, 1994.

[CKM94] I.B. Collings, V. Krishnamurthy, and J.B. Moore. "Online identification of hidden Markov models via recursive prediction error techniques." *IEEE Transactions On Signal Processing*, **42**(12):3535–3539, 1994.

[CL99] G.A. Churchill and B. Lazareva. "Bayesian restoration of a hidden Markov chain with applications to DNA sequencing." *Journal Of Computational Biology*, **6**(2):261–277, 1999.

[CLJ94] W. Chou, C.H. Lee, B.H. Juang, and F.K. Soong. "A minimum error rate pattern recognition approach to speech recognition." *Int. J. Pattern Recogn. Artificial Intelligence*, **8**(1):5–31, 1994.

[CRI03] G. Ciuperca, A. Ridolfi, and J. Idier. "Penalized maximum likelihood estimator for normal mixtures." *Scandinavian Journal Of Statistics*, **30**(1):45–59, 2003.

[DLR77] A. D. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, **B-39**:1–37, 1977.

[EDR89] Y. Ephraim, A. Dembo, and L.R. Rabiner. "A minimum discrimination information approach for hidden Markov modeling." *IEEE Transactions On Information Theory*, **35**(5):1001–1013, 1989.

[ENF02] G. Elidan, M. Ninio, and N. Friedman. "Data perturbation for escaping local maxima in learning." In *Proc. AAAI-2002*, Edmonton, July 2002.

[FF56] L. R. Ford and D. R. Fulkerson. "Maximal flow through a network." *Canadian Journal of Mathematics*, **8**:399–404, 1956.

[FH94] J.A. Fessler and A.O. Hero. "Space-alternating generalized expectation- maximization algorithm." *IEEE Transactions on Signal Processing*, **42**(10):2664–2677, 1994.

[FL89] A. Farago and G. Lugosi. "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, 1989.

[GD02] R. Granat and A. Donnellan. "A hidden Markov model based tool for geophysical data exploration." *Pure and Applied Geophysics*, **159**(10):2271–2283, 2002.

[HC93] Q. Huo and C. Chan. "The gradient projection method for the training of hidden Markov models." *Speech Communication*, **13**:307–313, 1993.

[HHK99] H. Hirose, K. Hirahara, F. Kimata, N. Fujii, and S. Miyazaki. "A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan." *Geophysical Research Letters*, **26**(21):3237–3240, 1999.

[HKM99] Q.H. He, S. Kwong, K.F. Man, and K.S. Tang. "Improved maximum model distance for HMM training." *Electronics Letters*, **35**(10):783–785, 1999.

[HKM00] Q.H. He, S. Kwong, K.F. Man, and K.S. Tang. "An improved maximum model distance approach for HMM-based speech recognition systems." *Pattern Recognition*, **33**(10):1749–1758, 2000.

[HMT97] K. Heki, S. Miyazaki, and H. Tsuji. "Silent fault slip following an interplate thrust earthquake at the Japan Trench." *Nature*, **386**(6625):595–598, 1997.

[JJ93] M. Jamshidian and R. I. Jennrich. "Conjugate gradient acceleration of the EM algorithm." *Journal of the American Stastical Association*, **88**:221–228, 1993.

[JLS86] B.H. Juang, S.E. Levinson, and M.M. Sondhi. "Maximum likelihood estimation for multivariate mixture observations of Markov chains." *IEEE Transactions On Information Theory*, **32**(2):307–309, 1986.

[JR85] B.H. Juang and L.R. Rabiner. "Mixture autoregressive hidden Markov models for speech signals." *IEEE Transactions On Acoustics Speech And Signal Processing*, **33**(6):1404–1413, 1985.

[JR90] B.H. Juang and L.R. Rabiner. "The segmental k-means algorithm for estimatinG parameters of hidden Markov models." *IEEE Transactions On Acoustics Speech And Signal Processing*, **38**(9):1639–1641, 1990.

[JR91] B.H. Juang and L.R. Rabiner. "Hidden Markov models for speech recognition." *Technometrics*, **33**(3):251–272, 1991.

[KCM01] S. Kwong, C.W. Chau, K.F. Man, and K.S. Tang. "Optimisation of HMM topology and its model parameters by genetic algorithms." *Pattern Recognition*, **34**(2):509–522, 2001.

[KGV83] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. "Optimization by simulated annealing." *Science*, **220**(4598):671–680, 1983.

[KK96] S. Kedar and H. Kanamori. "Continuous monitoring of seismic energy release associated with the 1994 Northridge earthquake and the 1992 Landers earthquake." *Bulletin Of The Seismological Society Of America*, **86**(1):255–258, 1996.

[Lee90] K.F. Lee. "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition." *IEEE Transactions On Acoustics Speech And Signal Processing*, **38**(4):599–609, 1990.

[LH89] K.F. Lee and H.W. Hon. "Speaker-independent phone recognition using hidden Markov models." *IEEE Transactions On Acoustics Speech And Signal Processing*, **37**(11):1641–1648, 1989.

[LR94] C. Liu and D. B. Rubin. "The ECME algorithm: a simple extension of EM and ECM with faster than monotone convergence." *Biometrika*, **81**:633–648, 1994.

[MMJ02] M.M. Miller, T. Melbourne, D.J. Johnson, and W.Q. Sumner. "Periodic slow earthquakes from the Cascadia subduction zone." *Science*, **295**(5564):2423–2423, 2002.

[MW02] T.I. Melbourne and F.H. Webb. "Precursory transient slip during the 2001 M_w=8.4 Peru earthquake sequence from continuous GPS." *Geophysical Research Letters*, **29**(21):art. no.–2032, 2002.

[MW03] T.I. Melbourne and F.H. Webb. "Slow but not quite silent." *Science*, **300**(5627):1886–1887, 2003.

[MWP00] G. McGuire, F. Wright, and M.J. Prentice. "A Bayesian model for detecting past recombination events in DNA multiple alignments." *Journal Of Computational Biology*, **7**(1-2):159–170, 2000.

[MWS02] T.I. Melbourne, F.H. Webb, J.M. Stock, and C. Reigber. "Rapid postseismic transients in subduction zones from continuous GPS." *Journal Of Geophysical Research – Solid Earth*, **107**(B10):art. no.–2241, 2002.

[NML95] R. Noumeir, G.E. Mailloux, and R. Lemieux. "An expectation maximization reconstruction algorithm for emission tomography with nonuniform entropy prior." *International Journal Of Bio-Medical Computing*, **39**(3):299–310, 1995.

[OT96] D. Ormoneit and V. Tresp. "Improved Gaussian Mixture Density Estimates Using Baysian Penalty Terms and Network Averaging." *Advances in Neural Information Processing Systems 8*, pp. 542–548, 1996.

[Rab89] L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition." *Proc. IEEE*, **77**(2):257–286, 1989.

[RD03] G. Rogers and H. Dragert. "Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip." *Science*, **300**(5627):1942–1943, 2003.

[RGF92] K. Rose, E. Gurewitz, and G. C. Fox. "Vector quantization by deterministic daing." *IEEE Trans. Inf. Theory*, **38**(4):1249–1257, 1992.

[RJL85] L.R. Rabiner, B.H. Juang, S.E. Levinson, and M.M. Sondhi. "Some properties of continuous hidden Markov model representations." *AT&T Technical Journal*, **64**(6):1251–1270, 1985.

[Ros98] K. Rose. "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems." *Proc. IEEE*, **86**(11):2210–2239, 1998.

[RR01] K. Rose and A. V. Rao. "Deterministically annealed design of hidden Markov model speech recognizers." *IEEE Trans. on Speech and Audio Processing*, **9**(2):111–126, 2001.

[RW84] R. A. Redner and H. F. Walker. "Mixture densities, maximum likelihood, and the EM algorithm." *SIAM Review*, **26**:195–239, 1984.

[SC98] P. Stolorz and P. Cheeseman. "Onboard science data analysis: Applying data mining to science-directed autonomy." *IEEE Intell Syst App*, **13**(5):62–68, 1998.

[SIG99] P. Smyth, K. Ide, and M. Ghil. "Multiple regimes in Northern hemisphere height fields via mixture model clustering." *Jour. Atmos. Sci.*, **56**(21):3704–3723, 1999.

[TKA04] V.C. Tsai, H. Kanamori, and J. Artru. "The morning glory wave of southern California." *Journal Of Geophysical Research – Solid Earth*, **109**(B2):art. no.–B02307, 2004.

[TPM02] M. Turmon, J. Pap, and S. Mukhtar. "Statistical Pattern Recognition for Labeling Solar Active Regions: Application to SoHO/MDI Imagery." *Astrophysical Journal*, **568**(1):396–407, March 2002.

[TSM85] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, Chichester, UK, 1985.

[UN94] N. Ueda and R. Nakano. "Mixture density estimate via EM algorithm with deterministic daing." *Proceedings of the IEEE Neural Networks for Signal Processing*, pp. 69–77, 1994.

[UN98] N. Ueda and R. Nakano. "Deterministic annealing EM algorithm." *Neural Networks*, **11**(2):271–282, 1998.

[UNG00] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. "SMEM Algorithm for Mixture Models." *Neural Computation*, **12**(9):2109–2128, 2000.

[Won93] Y. Wong. "Clustering data by melting." *Neural Computation*, **5**:89–104, 1993.

[WT02] M. Whiley and D. M. Titterington. "Applying the deterministic annealing expectation maximisation algorithm to naive bayesian networks." *Univ. Glasgow Tech. Report*, 2002.

[YSU94] A. L. Yuille, P. Stolorz, and J. Utans. "Statistical physics, mixtures of distributions and the EM algorithm." *Neural Computation*, **6**:334–340, 1994.

[YW94] S.J. Young and P.C. Woodland. "State clustering in hidden Markov model-based continuous speech recognition." *Computer Speech And Language*, **8**(4):369–383, 1994.