STRUCTURE AND PARAMETER LEARNING VIA ENTROPY MINIMIZATION, WITH APPLICATIONS TO MIXTURE AND HIDDEN MARKOV MODELS

Matthew Brand

MERL — A Mitsubishi Electric Research Lab
Cambridge, MA 02139
brand@merl.com

ABSTRACT

We develop a computationally efficient framework for finding compact and highly accurate hidden-variable models via entropy minimization. The main results are: 1) An entropic prior that favors small, unambiguous, maximally structured models. 2) A prior-balancing manipulation of Bayes' rule that allows one to gradually introduce or remove constraints in the course of iterative re-estimation. #1 and #2 combined give the information-theoretic free energy of the model and the means to manipulate it. 3) Maximum *a posteriori* (MAP) estimators such that entropy optimization and deterministic annealing can be performed wholly within expectation-maximization (EM). 4) Trimming tests that identify excess parameters whose removal will *increase* the posterior, thereby simplifying the model and preventing over-fitting. The end result is a fast and exact hill-climbing algorithm that mixes continuous and combinatoric optimization and evades sub-optimal equilibria.

1. MOTIVATION

In pattern discovery we seek a model that reflects the structure of the data, which we hope in turn reflects the structure of the generating process. We presume that the universe is constructed out of mostly small processes; in order to produce an object (dataset) larger than itself, a small process must repeat or loop some of its steps, albeit with variation. Therefore a dataset is an "unrolled" record of the process' internal structure. If we can find a compact model of the data, we feel increasingly confident that 1) we can interpret the model and in doing so learn about the process, and 2) predictions made from the model will be consistent with new samples taken from the process. We now know that #2 is well-founded: A recent result assures us (in the Bayesian setting) that the most compact of all possible models is virtually always the best hypothesis and strategy for prediction [11].

Sadly, the function which yields the most compact model is not computable. Furthermore, we only know how to do efficient searches for accurate models (disregarding compactness) in quite restricted spaces, e.g., for locally optimal parameterizations of probabilistic models whose kind and structure we fix in advance. In this paper we broaden that scope by estimating both model structure and parameters. Fortuitously, our methods will also yield a fast hill-climbing procedure for finding nearly global optima.

Our strategy is to minimize all entropies associated with modeling. Because of the duality between entropy and expected code length, this can also be understood as finding a model such that a message consisting of the model and the data can be maximally compressed. To maximally compress the data, we need to remove all of its redundancy, or, equivalently, discover all of its structure. This structure will be encoded by the model. To compress the model, we must maximally sparsify it, e.g., we seek to pack as much information into as few parameters as possible.

Our main results are a Bayesian framework for entropy minimization, solutions for the maximum *a posteriori* estimators, and a generalization of the expectation-maximization (EM) algorithm that gives it all the favorable properties of deterministic annealing. We will demonstrate with novel applications to speech and language problems using mixture and hidden Markov models (HMMs).

It is useful to contrast our framework with two other broad approaches, model selection and model construction. Model selection, the comparison of two rival models, has an enormous literature going back two centuries. With the advent of computers, model selection has become the core of myriad generate-and-test search procedures, necessitating exponential amounts of speculative and wasted computation. For example Stolcke and Omohundro [10] fall back from EM to generate-and-test with gradient descent, in order to attempt model selection using minimum description-length (MDL) terms (à la Rissanen [6]). In model construction, one adds to an existing model by searching the training data for yet-unnoticed structures; typically the algorithms are heuristic, though in some cases here are proofs of approximate correctness (e.g., [8]). Except in special cases, the time complexity of search is at best high-order polynomial. We refer the reader to [2] for a survey of the literature on finding the structure of HMMs—of 33 recent papers detailing 12 different approaches, 31 propose generate-and-test searches and the remaining 2 propose heuristic constructive methods with clear vulnerabilities to degenerate cases. Our method is *eliminative*—it sculpts well-fit models out of monolithic blocks of random parameters. It has the same linear/quadratic complexity as conventional EM, but finds far superior optima.

2. MAXIMUM-STRUCTURE PRIORS

We begin with a set $\boldsymbol{X}$ of observations and a hypothesis class of hidden-variable models. We use hidden variables because we assume that the data has latent structure and/or is incomplete. The vector $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_N\}$ parameterizes all models having N or fewer parameters and specifies their structure via a sparse encoding in which $\theta_{f(i,j)} = 0$ means that variables i and j are independent (f and f^{-1} generically map between model and vector indices). Starting from a random $\boldsymbol{\theta}$ we seek an optimal model $\boldsymbol{\theta}^*$ that maximizes the posterior $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\boldsymbol{X})$ given by Bayes' rule: $P(\boldsymbol{\theta}|\boldsymbol{X}) \propto P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})$, where the likelihood $P(\boldsymbol{X}|\boldsymbol{\theta})$ measures accuracy in modeling the data and the prior $P(\boldsymbol{\theta})$ measures consis-

tency with our background knowledge.

What is our background knowledge? Let us assert that the vector of model parameters $\boldsymbol{\theta}$ is itself taken from a random variable $\boldsymbol{\Theta}$ which generates all possible processes. We know that on average, these processes are not infinitely unpredictable—otherwise learning would be impossible! Therefore our background knowledge is that, on average, learning will succeed. More formally, we say that the expected entropy of processes from $\boldsymbol{\Theta}$ is finite. We write this prior knowledge as $\xi\colon\ E_{\boldsymbol{\Theta}}[H(\boldsymbol{\theta})] = h$, where $H(\boldsymbol{\theta})$ is an entropy measure assessed on the model specified by $\boldsymbol{\theta}$, $E_{\boldsymbol{\Theta}}$ is the expectation with regard to $\boldsymbol{\Theta}$, and h is some finite value. The classic maximum-entropy method [4, §2] allows us to derive the mathematical form of a distribution from knowledge ξ about its expectations via Euler-Lagrange equations, yielding

$$P(\boldsymbol{\theta}|\xi) \propto \exp[-\lambda H(\boldsymbol{\theta})] \tag{1}$$

where the Lagrange multiplier λ depends on h and is unknown. We shall find meaningful interpretations for several values of λ, but we shall concentrate on the assumption that $\lambda = 1$. This we shall call the entropic prior

$$P_e(\boldsymbol{\theta}) \stackrel{\text{def}}{\propto} e^{-H(\boldsymbol{\theta})} \tag{2}$$

We assume ξ henceforth and drop it from notation. We note in passing that by making a similar assertion about the expected perplexity ($e^{H(\boldsymbol{\theta})}$) and assuming a measure on λ, it is also possible to integrate out the Lagrange multiplier and arrive directly at eqn. 2.

We call the reader's attention to two properties that derive from the definition of entropy: 1) $P_e(\cdot)$ is a bias for compact models having less ambiguity, more determinism, and therefore more structure. 2) $P_e(\cdot)$ is invariant to reparameterizations of the model, because the entropy is a property of the model, or because it can be specified in terms of a mean-value parameterization.

Eqn. 2 is a remarkable form which we will exploit to simultaneously estimate the structure and parameters of complex probability models—a mixed combinatorial (structure) and continuous (parameter) optimization problem. The results can be quite good; however, both hidden-variable and combinatorial optimization problems have notoriously rough energy surfaces. Before presenting estimators in §4 we will develop a technique that uses the entropic prior to improve the quality of local optima found by hill-climbing algorithms such as EM, and which finds the global optimum in the limiting case.

3. PRIOR BALANCING

We now introduce a generalization of the posterior, by rewriting Bayes' rule with a manipulation of the prior:

$$\tilde{P}_e(\boldsymbol{\theta}|\boldsymbol{X}, T, T_0) \stackrel{\text{def}}{\propto} P(\boldsymbol{X}|\boldsymbol{\theta})\ P(\boldsymbol{\theta})^{T_0 - T}\ \delta(T) \tag{3}$$

where T is normally positive and $\delta(T) \in (0, 1]$ is a driving term that monotonically increases as T approaches zero (e.g., the Gaussian $\delta(T) \propto \exp -T^2/2$). Varying T balances the prior against the likelihood, which is useful in iterative parameter estimation because it allows $\boldsymbol{\theta}$ to get into the right neighborhood with respect to one constraint before attempting to satisfy the other. Of course, to obtain a proper probability we are obliged to make $Z = T_0 - T$ converge to a meaningful value (e.g., $Z \to 1$) which can be done by following the gradient $\Delta T \propto \frac{\partial}{\partial T} \log \tilde{P}(\boldsymbol{\theta}|\boldsymbol{X}, T, T_0)$ or by iteratively tracking the maximum *a posteriori* (MAP) estimate of $\hat{T} = \operatorname{argmax}_T \tilde{P}(\boldsymbol{\theta}|\boldsymbol{X}, T, T_0)$.

Using the shorthand $H = -\log P(\boldsymbol{\theta})$, the gradient and MAP estimate for the Gaussian driving function are

$$\Delta T \propto H - T \quad ; \quad \hat{T} = H \tag{4}$$

The rate at which T decays can be controlled by adjusting the variance of δ, or by choosing a different δ function entirely.

This technique works for any prior, but it becomes especially meaningful with the entropic prior (eqn. 2). Consider the posterior transformed into a penalty function $\tilde{F} \stackrel{\text{def}}{=} -\log \tilde{P}(\boldsymbol{\theta}|\boldsymbol{X}, T, T_0)$:

$$\tilde{F} = E - (T - T_0) H(\boldsymbol{\theta}) - \log \delta(T) \quad \geq \quad F \stackrel{\text{def}}{=} E - TH \tag{5}$$

where $E = -\log P(\boldsymbol{X}|\boldsymbol{\theta})$ is the error or energy cost of a given parameterization, and T is understood as temperature. $\tilde{F}$ is an upper bound on the Helmholtz free energy F of thermodynamics, with equality at $\delta(T) = 1$, $T_0 = 0$. Maximizing the modified posterior thus minimizes the free energy, which is analogous to finding an equilibrium configuration in a complex model whose different parts compete to explain the data.

Four interesting cases immediately fall out of $\tilde{F}$: **1)** Iteratively re-estimating $\boldsymbol{\theta}$ while $T \to 0$ gives deterministic annealing (DA) [9, 5], a pseudo-global optimizer for pock-marked energy surfaces. DA belongs to a family of continuation techniques that convolve (smooth) a energy surface to make it globally convex, then track the optimum as gradual deconvolution re-introduces the hills and valleys of the surface. (In statistics, robust M-estimation [3] with decaying scale has analogous behavior.) During DA, entropy at high T keeps the system from prematurely committing to nearby local optima and forces it to explore the energy surface's large-scale structure. In a hidden-variable model, this is equivalent to maximizing almost equally w.r.t. all possible hypotheses within the model (e.g., all possible paths through an HMM), then concentrating on the most promising hypotheses as the temperature declines. Note that our modified posterior gives a useful amendment to DA—an automatic annealing schedule that tracks the quality (inversely, the entropy) of the model via the gradients or MAP estimates w.r.t. T.

The remaining cases of interest arise out of different terminal temperatures: **2)** At $T - T_0 = 1$ we obtain a maximum-entropy solution. **3)** At $T - T_0 = 0$ we obtain the maximum-likelihood (ML) solution. **4)** At $-Z = T - T_0 = -1$ we obtain the maximum-structure solution. Conveniently, we have derived a single MAP estimator for all three cases.

4. MAP ESTIMATORS

We obtain MAP parameter estimators by solving for maxima of $\boldsymbol{\theta}$-terms in the log of the entropic posterior,

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} [\log P(\boldsymbol{X}|\boldsymbol{\theta}) - ZH(\boldsymbol{\theta}) + c] \tag{6}$$

Although this can lead to systems of transcendental equations, we have obtained solutions for most simple distributions and, by subadditivity principles, for all models composed thereof. We will state basic results here; derivations can be found in [2, 1]. For the multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{K}$ the entropy is $\frac{1}{2}\log((2\pi)^d e|\boldsymbol{K}|)$. The entropic prior is thus $P_e(\boldsymbol{\mu}, \boldsymbol{K}) \propto |\boldsymbol{K}|^{-1/2}$, which is uniform in $\boldsymbol{\mu}$ and inversely proportional to the volume of $\boldsymbol{K}$. The MAP estimator is essentially an $N + Z$ normalization of the scatter of N samples:

$$\hat{\boldsymbol{K}} = \frac{\sum_n^N \boldsymbol{x}_n \boldsymbol{x}_n^\top}{N + Z} \tag{7}$$

Note that the maximum-structure (Z=1) MAP estimator yields the best mean squared-error estimate, while the maximum-entropy (Z=−1) MAP estimator yields the best unbiased estimate. Similarly normalized estimators give the scale parameters of related distributions, e.g., exponential, Laplace, and gamma.

A multinomial has entropy $-\sum_i \theta_i \log \theta_i$ and entropic prior $P_e(\boldsymbol{\theta}) \propto \boldsymbol{\theta}^{\boldsymbol{\theta}} = \prod_i \theta_i^{\theta_i}$. For the MAP estimator given a vector $\boldsymbol{\omega}$ of evidence for each alternative, we set the derivative of the log-posterior to zero, using a Lagrange multiplier to ensure $\sum_i \theta_i = 1$,

$$0 = \frac{\partial}{\partial \theta_i}\left(\log \prod_i \theta_i^{\omega_i + Z\theta_i} + \lambda \sum_i \theta_i\right) = \frac{\omega_i}{\theta_i} + Z\log\theta_i + Z + \lambda \quad (8)$$

The resulting system of simultaneous transcendental equations is obviously nonalgebraic, but we have been able to solve it using the Lambert W function, a multi-valued inverse function satisfying $W(x)e^{W(x)} = x$:

$$\hat{\theta}_i = \frac{-\omega_i/Z}{W(-\omega_i e^{1+\lambda}/Z)} \quad (9)$$

Eqn. 8 and eqn. 9 form a fix-point equation for λ. It typically converges to machine precision in 2-5 iterations. See [2] for details on using and computing W.

It can be shown that the MAP estimate minimizes the sum

$$H(\boldsymbol{\theta}) + D(\boldsymbol{\omega}\|\boldsymbol{\theta}) + H(\boldsymbol{\omega}) \quad (10)$$

thus reducing all entropies (in the model and in the data's sufficient statistics) and cross entropies (between the model and the data). In practice, it does so by extinguishing weakly supported parts of the model that are ill-matched to the structure of the data.

5. TRIMMING

The prior allows us to identify excess parameters that can be trimmed from a multivariate model without loss of posterior probability, such that $P_e(\boldsymbol{\theta}\backslash\theta_i|\boldsymbol{X}, Z) \geq P_e(\boldsymbol{\theta}|\boldsymbol{X}, Z)$. Expanding via Bayes' rule, taking logarithms, and rearranging, we obtain

$$Z(H(\boldsymbol{\theta}) - H(\boldsymbol{\theta}\backslash\theta_i)) \geq \log P(\boldsymbol{X}|\boldsymbol{\theta}) - \log P(\boldsymbol{X}|\boldsymbol{\theta}\backslash\theta_i) \quad (11)$$

Operationally, a parameter is trimmed by setting it to a default or minimal entropy value, e.g., for multinomials $\theta_i \leftarrow 0$, for variances $k_{ii} \leftarrow 1$ (equivalently $\log k_{ii} \leftarrow 0$). When zeroing, eqn. 11 can be approximated via differentials, yielding

$$Z\frac{\partial H(\boldsymbol{\theta})}{\partial \theta_i}|\theta_i| \quad > \quad |\theta_i|\frac{\partial \log P(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_i} \quad (12)$$

The sides of eqns. 11 and 12 can be mixed and matched to obtain mathematically convenient forms. For the multinomial trim test, we set l.h.s. eqn. 11 (=$-Z\theta_i \log \theta_i$) against r.h.s. eqn. 12 to obtain

$$\theta_i < \exp\left[-\frac{1}{Z}\frac{\partial \log P(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_i}\right] \quad (13)$$

Typically the gradient of the log-likelihood has already been computed for re-estimation, so these trimming tests are very cheap.

Solving eqn. 11 via W gives a variance trimming ($k_{ii} \leftarrow 1$) criterion for Gaussians:

$$k_{ii} \geq -\hat{k}_{ii}/W_0(-\hat{k}_{ii}e^{-\hat{k}_{ii}}) \quad \text{iff} \quad \hat{k}_{ii} > 1 \quad (14)$$

$$k_{ii} \leq -\hat{k}_{ii}/W_{-1}(-\hat{k}_{ii}e^{-\hat{k}_{ii}}) \quad \text{iff} \quad \hat{k}_{ii} < 1 \quad (15)$$

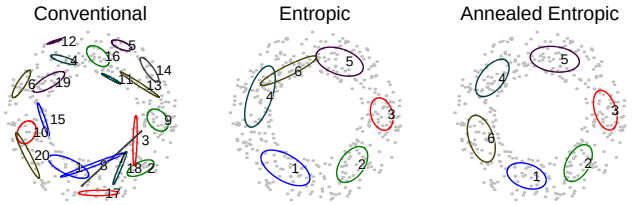

Figure 1: Mixture models estimated from identical initial conditions, superimposed on the data.

There are also trimming tests for covariances ($k_{ij} \leftarrow 0$) and for removing dimensions, but they are beyond the scope of this paper.

Trimming increases but does not maximize the posterior. The reason why it becomes desirable in iterative estimation of models containing hidden variables is that it can radically sparsify the model, which in turn reduces ambiguity in the data's expected sufficient statistics, and speeds learning. Not only does trimming protect against over-fitting, but by sculpting the model to fit the data, it often reveals a simple machine that provides insight into the causality of the process that produced the data (e.g., a finite-state machine from an HMM; a circuit from a NN).

6. EXAMPLES

Mixture model of a textbook problem: To illustrate, a mixture model was fitted in Cartesian space to an ring of samples taken from a uniform distribution over the polar coordinates $r \in [1, 2]$; $\alpha \in [0, 2\pi]$. The figures show the model estimated conventionally; entropically with trimming; and entropically with trimming and deterministic annealing. Initial conditions were identical for all three cases (figure 1). In the maximum likelihood case, over-fitting has resulted in a model of the accidental properties of the data (e.g., the clumpiness of the sample). In the second case, trimming has removed excess components and the resulting model looks much more like the essential structure of the data, but a large under-sampled region near the top still affects the model. In the third case, DA circumvents this local optimum and finds a model which generalizes even better.

Mixture model of of benchmark vowel acoustic data: We obtained the British English vowel recognition dataset from the CMU Neural-Bench Archive. Each example of a vowel is characterized by LPC coefficients from two time-frames. This dataset has been treated several times using perceptrons, neural networks, Kanerva models, radial basis function networks (RBFNs), and non-parametric methods (nearest neighbor). The last yielded the best classification of the test set (56%), while RBFNs peaked at 53% with 528 Gaussian basis functions [7]. We combined entropically estimated mixture models of each vowel's training data into an RBFN and obtained 58.7% correct classification on the test set. Entropic estimation found mixtures of 1-3 components per vowel, resulting in an RBFN of only 22 basis functions.

HMM of handwriting data: We obtained handwriting samples by 10 writers from the UNIPEN archive. Figure 2 shows two models of the pen-strokes for the digit "5," estimated from pen-position data taken at 5msec intervals from 10 different individuals via an electro-magnetic resonance sensing tablet. Ellipses show iso-probability contours for each state; ×s and arcs indicate state dwell and transition probabilities, respectively, by their thicknesses. En-

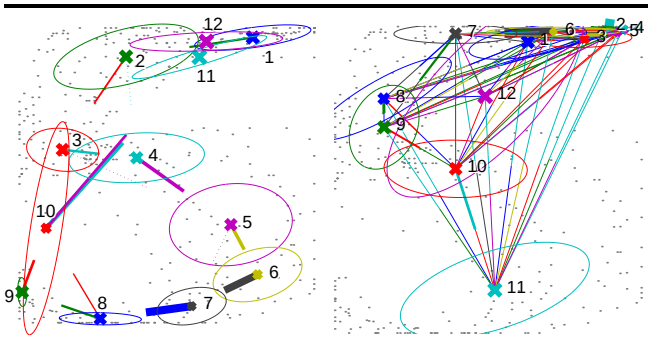

Figure 2: Hidden Markov models of writing "5," estimated entropically and conventionally (no annealing), and superimposed on the data. The entropic model clearly shows the structure of the pen-strokes, as well as variations in their ordering between writers.

tropic estimation induces an interpretable automaton that captures essential structure and timing of the pen-strokes. 50 of the original 80 dynamical parameters were trimmed. Estimation without the entropic prior results in a wholly opaque model, in which none of the original dynamical parameters were trimmed. Over 10 trials with different parts of the database held out, entropic models yielded 96% correct classification; ML models yielded 93%.

HMM of phonetic data: Using *only* phonetic transcriptions from the TIMIT databases of Western and Southern speakers, we sought to classify sentences by speaker group using HMMs. Note that since we are not using waveforms, we do not have acoustic information about people's accents. We trained 20-state discrete-output HMMs entropically and conventionally (maximum likelihood). Entropic estimation trimmed more than 1100 of the original 1620 parameters in each model. (Not all of this is due the the entropy terms; in the ML models, roughly 300 parameters were zeroed due to underflow.) Entropic estimation converged in roughly 2/3 the time of conventional training. The entropic models were able to correctly classify 57% of the sentences in the test set; the ML models, 54%.

7. DISCUSSION

We have used entropic estimation in many other domains to obtain highly highly sparsified mixture models, HMMs, RBFNs, and generalized recurrent neural networks, using benchmark datasets (e.g., Bach Chorales, Boston Home Values) as well as real-time data obtained directly from computer vision and speech analysis systems. The resulting models have been smaller, better generalizing, and more predictive than conventionally estimated models. Even though we are not using discriminant training, entropic models are tend to be more discriminative, simply because they're better models. Often they are interpretable, e.g., in the Bach HMMs we found many low-perplexity state subgraphs that capture standard melodic devices of Baroque composition [2]. In some cases, the models are so good they can be used to generate the signals they are trained to accept—for example, generating photo-realistic facial animation from entropically estimated HMMs.

We have also found a case where annealing to maximum-structure ($Z \to 1$) can be a liability: When the data is *atypical* of the generating process, the maximum structure model, being a stronger theory of the data, will probably not generalize as well as an annealed maximum entropy model. This situation will arise when the training sample is far too small to be a fair representation of the target concept. However, in a year of experiments, this case arose only once, when modeling the statistics of English from just a few pages of text.

In addition to the choice of terminal prior (temperature), our framework allows two other degrees of freedom: Firstly, $H(\boldsymbol{\theta})$ may be interpreted as the entropy of the model parameterized by $\boldsymbol{\theta}$ *or* as the entropy of the model's parameter vector $\boldsymbol{\theta}$. The former case can be understood as strict entropy minimization; the latter as MDL. Both are consistent with our framework; in some cases they are identical. When an analytic form of the model entropy is unavailable, the parameter entropy often can be used as an upper-bound. Secondly, our framework is agnostic about when one trims—although we have trimmed greedily, one might wait until near convergence or after DA. Various combinations of entropic training, trimming, and prior-balancing have intriguing physical analogues, including the metallurgical processes of casting, annealing, tempering, and bluing. Like special-purpose steels, variations may yield models that are highly malleable or selective etc., while simple defaults lead to general-purpose high-quality models.

8. REFERENCES

[1] Matthew Brand. Learning concise models of human activity from ambient video. Technical Report 97-25, Mitsubishi Electric Research Labs, November 1997.

[2] Matthew Brand. Structure discovery in conditional probability models via an entropic prior and parameter extinction. *Neural Computation (accepted 8/98)*, October 1997. See also www.merl.com/projects/structure.html

[3] P.J. Huber. *Robust Statistics*. Wiley and Sons, 1981.

[4] Edwin T. Jaynes. *Papers on probability, statistics, and statistical mechanics*, (Brandeis Lectures) (1963), pages 39–76. Kluwer Academic, 1982.

[5] D. Miller, A. Rao, K. Rose, and A. Gersho. A global optimization technique for statistical classifier design. *IEEE Transactions on Signal Processing*, 4:3108–3121, 1996.

[6] Jorma Rissanen.*Stochastic Complexity and Statistical Inquiry*. World Scientific, 1989.

[7] A. J. Robinson. *Dynamic Error Propagation Networks*. PhD thesis, Cambridge University Engineering Department, 1989. See www.boltz.cs.cmu.edu/benchmarks/vowel.html.

[8] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, December 1996.

[9] K. Rose, E. Gurewitz, and G.C. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594, 1990.

[10] Andreas Stolcke and Stephen Omohundro. Best-first model merging for hidden Markov model induction.ICSI TR-94-003, U.C. Berkeley. April 1994.

[11] P.M.B. Vitanyi and M. Li. Ideal MDL and its relation to Bayesianism. In *ISIS: Information, Statistics and Induction in Science*, pages 282–291. World Scientific, Singapore, 1996.