Bayesian Model Selection for Markov, Hidden Markov, and Multinomial Models

Mathias Johansson and Tomas Olofsson

***Abstract*—Model selection based on observed data sequences is used to decide between different model structures within the class of multinomial, Markov, and hidden Markov models. In a unified Bayesian treatment, we derive posterior probabilities for different model structures without assuming prior knowledge of transition probabilities. We emphasize the following tests: 1) Given a particular data sequence of n outcomes, is each state equally likely? 2) Do the data support an independent model, or is a Markov model a more plausible description? 3) Are two data sequences generated from a) the same Markov model? b) the same hidden Markov model? For Markov models and independent multinomial models, all results are exact. For hidden Markov models, the exact solution is computationally prohibitive, and instead, an approximate solution is proposed.**

***Index Terms*—Bayes procedures, hidden Markov models (HMMs), Markov models.**

I. Introduction

SEQUENCE data may be modeled probabilistically using several different model families, where the (independent) multinomial, discrete Markovian, and hidden Markov models (HMMs) are some of the most commonly used. Two important questions that arise in sequence modeling are those of how to determine whether 1) one sequence belongs to a class of models with a fixed structure but with undetermined parameters, or how to determine whether 2) two or more sequences have been generated by the same or by different models. This letter gives a unified treatment of problems 1) and 2) using Bayesian inference for the multinomial, Markovian, and HMMs.

A. Contributions and Related Work

Multinomial and Markov models have been used in a vast variety of fields over a long time; hence, publications on model selection for these model families have appeared spread out over a large literature. Several results that we present here from a unified Bayesian derivation have been derived before in specific scenarios, but we include them as a service to the reader. As these results have appeared in very different application areas, it is difficult for a general reader to find the results and to see their relations and tacit assumptions. We believe that this motivates a unified derivation of these important results, which also highlights the basic unity of them all. Our treatment of HMMs has, to the best of our knowledge, not been presented before. Here lies the main novelty of our work, although our showing that all model selection problems involving multinomial, Markov, and HMMs are special cases of a particular integral (see (6) below) is perhaps more important.

Bayesian model selection is based on computing the evidence (or, the marginal likelihood; see below for explanations) for a given model. The Bayesian test of equiprobability of a multinomial model, a special case of (10) derived below, has been derived by others before (see Good's paper [1]). Using a similar Bayesian approach, Cooper and Herskovits [2] considered evidence computations for Bayesian networks for inferring the most probable network structures. As Markov models are a special case of a Bayesian network, it is possible to derive our (non-hidden) Markov model results from their expressions. Indeed, in [3], the results of Cooper and Herskovits were applied to clustering of different Markov processes.

Model selection for HMMs is typically done using asymptotically motivated criteria, such as variants of Akaike's or the Bayesian information criteria (see, e.g., [4]). However, in order to use them, a particular state sequence must be singled out, and the uncertainty concerning the actually traversed states is thus neglected.

II. Bayesian Model Comparison

Our basic problem is that of determining the probability for some clearly specified hypothesis M_i that describes the observed data, D, based on all available prior information I that is relevant to the problem. The posterior probability is given by

$$P(M_i|D,I) = \frac{P(D|M_i,I)P(M_i|I)}{P(D|I)}. \tag{1}$$

When the prior information I does not give any preference for any model over any other, the principle of indifference [5] yields the prior $P(M_i|I) = 1/\mathcal{M}$, where $\mathcal{M}$ is the overall number of considered models. The numerator is obtained from

$$P(D|I) = \sum_i P(D|M_i,I)P(M_i|I) \tag{2}$$

where the sum is over all considered models. The crucial factor, and our main concern, is thus the marginal likelihood, or the *evidence*, for the model $P(D|M_i,I)$. As we shall see, the computation of this factor follows a similar pattern for all models considered in this letter.

We note that model comparison between M_k and M_l can be performed by evaluating the posterior odds

$$O = \frac{P(D|M_k,I)}{P(D|M_l,I)}\frac{P(M_k|I)}{P(M_l|I)}. \tag{3}$$

With equal prior probabilities for the two alternatives, the posterior odds (3) reduces to the evidence ratio $(P(D|M_k,I))/(P(D|M_l,I))$. Henceforth, we will assume equal prior probabilities for the models under consideration in all tests. For more on Bayesian model selection, see [5, Ch. 20].

We now turn our attention to the specific tests, starting with tests on the multinomial model.

Manuscript received May 2, 2006; revised June 30, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Deniz Erdogmus.

The authors are with the Signals and Systems Group, Uppsala University, SE-75120 Uppsala, Sweden (e-mail: mj@signal.uu.se).

Digital Object Identifier 10.1109/LSP.2006.882094

A. Independent Multinomial Models

Independent multinomial models are completely parameterized by K probabilities, one for each possible state. Let f_k denote the probability for observing state k at any time. We have K such probabilities that we gather in the parameter vector $\mathbf{f} = (f_1, \dots, f_K)^T$. They sum to unity and are assumed constant over time and independent of previous states. The probability for observing a specific sequence S of states is thus simply

$$P(S|\mathbf{f}, I) = f_1^{n_1} \cdots f_K^{n_K} \tag{4}$$

where n_k denotes the number of observations of state k. Let $n = \sum_{k=1}^{K} n_k$ denote the total number of observations, i.e., the length of the sequence S.

Let H_N denote the hypothesis stating that a sequence is independent multinomial. We now wish to compute the evidence $P(S|H_N, I) = \int P(S|\mathbf{f} H_N I) P(\mathbf{f}|H_N I) d\mathbf{f}$, where $d\mathbf{f}$ is shorthand for the volume element $df_1 \cdots df_K$.

If our prior information I suggests no preference for any vector $\mathbf{f}$ over any other, we express this state of indifference by assigning a uniform prior over the values that sum to unity

$$P(\mathbf{f}|H_N I) = C\delta\left(\sum_{k=1}^{K} f_k - 1\right), \quad f_k \geq 0 \tag{5}$$

where $\delta(x) = 1$ if $x = 0$, $\delta(x) = 0$ elsewhere, and C is a normalization constant.

Below, we will use the result (see the Appendix)

$$\int f_1^{n_1} \cdots f_K^{n_K} \delta\left(\sum_{k=1}^{K} f_k - 1\right) d\mathbf{f} = \frac{\prod_{k=1}^{K} \Gamma(n_k + 1)}{\Gamma(n + K)}. \tag{6}$$

By noting that the prior (5) must integrate to unity, we find by inserting $n_1 = \ldots = n_K = 0$ that the normalization constant becomes $C = (K-1)!$. The evidence for a multinomial model now follows directly from the integral result (6) as

$$P(S|H_N I) = \frac{(K-1)! \prod_{k=1}^{K} \Gamma(n_k + 1)}{\Gamma(n + K)}. \tag{7}$$

This result can be extended to the case when we have observed two sequences, S^1 and S^2. Let n_k^1 and n_k^2 denote the number of observations of state k in S^1 and S^2, respectively, and let $\tilde{n}_k = n_k^1 + n_k^2$, i.e., the accumulated number of observations of state k in S^1 and S^2. Finally, let $\tilde{n} = \sum_{k=1}^{K} \tilde{n}_k$.

Consider now the case that these sequences both originate from the same multinomial model, and denote that hypothesis H_{sN}. This is equivalent to the statement that the same $\mathbf{f}$ has generated both sequences. Consequently, the evidence for H_{sN} is obtained directly from (7) with $\tilde{n}_k$ in place of n_k

$$P(S^1, S^2|H_{sN} I) = \frac{(K-1)! \prod_{k=1}^{K} \Gamma(\tilde{n}_k + 1)}{\Gamma(\tilde{n} + K)}. \tag{8}$$

We now have the necessary means to set up the first test.

Test 1 (Same Versus Different Multinomial Models): Let H_{sN} and H_{dN} denote the hypotheses stating that two sequences, S^1 and S^2, originate from the same multinomial model and different multinomial models, respectively. The evidence of the first is given by (8), and the evidence of the latter is simply the product of the evidences for the respective sequences given by (7). We have

$$\frac{P(D|H_{sN}, I)}{P(D|H_{dN}, I)} = \frac{1}{(K-1)!} \frac{\Gamma(n^1 + K)\Gamma(n^2 + K)}{\Gamma(\tilde{n} + K)} \times \frac{\prod_{k=1}^{K} \Gamma(\tilde{n}_k + 1)}{\prod_{k=1}^{K} \Gamma\left(n_k^2 + 1\right) \prod_{k=1}^{K} \Gamma\left(n_k^1 + 1\right)}. \tag{9}$$

Test 2 (Is a Particular Model Supported by Data?): In practice, we often have a suggestion for a particular multinomial model. We would then like to test whether this model is supported also by the observed data. Let the particular model, H_{pN}, be specified by a known probability vector $\mathbf{f}$. Given a sequence S, we then wish to test this model versus the class of all other multinomial models. The evidence for H_{pN} is simply given by (4). The posterior odds are thus obtained by dividing this by the evidence (7) for an unspecified multinomial model

$$O = \frac{P(S|H_{pN} I)}{P(S|H_N I)} = \frac{\Gamma(n + K) f_1^{n_1} \cdots f_K^{n_K}}{(K-1)! \prod_{k=1}^{K} \Gamma(n_k + 1)}. \tag{10}$$

If this ratio is larger than unity, the data support the particular model H_{pN}.

B. Markov Models

Let H_M denote the hypothesis stating that the sequence originates from a Markov process. The Markov model can be regarded as a generalization of the independent multinomial model, in which probabilities f_{jk} for the next state j depend on the current state k but no earlier states. A K-state Markov model is uniquely described by a transition matrix $\mathbf{F}$ defined by

$$\mathbf{F} = \begin{pmatrix} f_{11} & \cdots & f_{1K} \\ \vdots & \ddots & \vdots \\ f_{K1} & \cdots & f_{KK} \end{pmatrix} \tag{11}$$

and a vector $\mathbf{f}_0 = (f_{10}, \dots, f_{K0})^T$, where f_{k0} is the probability of starting in state k. For notational convenience, we also introduce the parameter vectors $\mathbf{f}_1, \dots, \mathbf{f}_K$, where $\mathbf{f}_k = (f_{1k}, \dots, f_{Kk})^T$, i.e., the transition probabilities associated with leaving state k.

For a known set of parameters, the probability for obtaining a specific sequence S of observations is

$$P(S|\mathbf{f}_0, \mathbf{F}, I) = f_{i0} \prod_{k=1}^{K} \prod_{l=1}^{K} f_{kl}^{m_{kl}} \tag{12}$$

where i is the first state, and m_{kl} is the number of observations of transitions from state l to state k. Note that the expression for the Markov likelihood is entirely analogous to the independent multinomial likelihood (4).

If we, again completely similar to the multinomial case, assume complete ignorance of the parameter values in $\mathbf{F}$ and $\mathbf{f}_0$, and assume no dependencies between the vectors $\mathbf{f}_0 \ldots, \mathbf{f}_K$, we obtain the parameter prior

$$\begin{aligned} P(\mathbf{f}_0, \mathbf{F}|I) &= P(\mathbf{f}_0|I) \prod_{l=1}^{K} P(\mathbf{f}_l|I) \\ &= C_0 \delta\left(\sum_{k=1}^{K} f_{k0} - 1\right) \prod_{l} C_l \delta\left(\sum_{k=1}^{K} f_{kl} - 1\right) \end{aligned} \tag{13}$$

where the normalization constants $C_l = (K-1)!$. It is now clear that the evidence for a Markov model is of the same form as in the independent multinomial case, and the final result follows from application of the basic integral result (6). By integrating over the parameter space and using (6), (12), and (13), we obtain the evidence $P(S|H_M, I)$

$$\begin{aligned} P(S|H_M,I) &= \int P(S|\mathbf{F},\mathbf{f}_0,I)P(\mathbf{F},\mathbf{f}_0,I)d\mathbf{f}_0 d\mathbf{F} \\ &= \int f_{i0}P(\mathbf{f}_0|H_M,I)d\mathbf{f}_0 \\ &\quad \times \prod_{l=1}^{K} \int f_{1l}^{m_{1l}} \dots f_{Kl}^{m_{Kl}} P(\mathbf{f}_l|I)d\mathbf{f}_l \\ &= \frac{1}{K}\prod_{l}(K-1)!\frac{\prod_{k=1}^{K}\Gamma(m_{kl}+1)}{\Gamma(m_l+K)} \end{aligned} \tag{14}$$

where $m_l = \sum_{k=1}^{K} m_{kl}$, and f_{i0} is the probability for the observed initial state. Note that (14) is just a product over evidence factors of the same form as in the multinomial case (7) multiplied by $1/K$, accounting for the probability for the initial state.

Similar to the multinomial model, the evidence for a hypothesis, H_{sM}, stating that two sequences, S^1 and S^2, belong to the same Markov model can also be easily evaluated, cf. (8)

$$P(S^1,S^2|H_{sM},I) = \frac{c_1}{K(K+1)}\prod_{l}(K-1)!\frac{\prod_{k=1}^{K}\Gamma(\tilde{m}_{kl}+1)}{\Gamma(\tilde{m}_l+K)} \tag{15}$$

where $c_1 = 1$ if the initial states are different and $c_1 = 2$ if the states are the same.

Now we are prepared to present two tests, one for determining whether two sequences originate from the same process and one for determining whether a sequence is independent or Markovian.

Test 3 (Same Versus Different Markov): Let H_{sM} and H_{dM} denote the hypotheses stating that two sequences, S^1 and S^2, originate from the same Markov model or different Markov models. The evidence for the first is given by (15), and the evidence for the latter is simply the product of the evidences for the respective sequences. This yields the odds

$$\begin{aligned} O &= \frac{P(D|H_{sM},I)}{P(D|H_{dM},I)} \\ &= \frac{Kc_1}{(K+1)(K-1)!^K}\prod_{l=1}^{K}\frac{\Gamma\left(m_l^1+K\right)\Gamma\left(m_l^2+K\right)}{\Gamma(\tilde{m}_l+K)} \\ &\quad \times \prod_{k=1}^{K}\frac{\Gamma(\tilde{m}_{lk}+1)}{\Gamma\left(m_{lk}^1+1\right)\Gamma\left(m_{lk}^2+1\right)}. \end{aligned} \tag{16}$$

Test 4 (Test of Independence): We here compute the odds for the hypothesis of an independent multinomial model versus that of the Markov class of models. Before we calculate the odds, we draw attention to a slight difference in how the counters m_k and n_k are treated. For the multinomial model, n_k is the total number of occurrences of state k in the sequence, whereas m_k is the number of occurrences of state k, excluding the nth position. Thus, n_k and m_k are related through $n_k = m_k + \delta_{s_n=k}$, where $\delta_{s_n=k}$ is one if the last state in the sequence is k, otherwise zero. We obtain the odds

$$\begin{aligned} O &= \frac{P(S|H_N,I)}{P(S|H_M,I)} \\ &= \frac{K}{(K-1)!^{K-1}}\frac{\prod_l\Gamma(m_l+K)\prod_l\Gamma(m_l+\delta_{s_n=l}+1)}{\Gamma(n+K)\prod_l\prod_k\Gamma(m_{kl}+1)}. \end{aligned} \tag{17}$$

C. Hidden Markov Models

Let H_H denote the hypothesis that a HMM generates the observed data. A HMM is a generalization of a Markov model. The sequence generated from an HMM is no longer observations of the states s_t at times t but rather observations of data x_t that are thought of as being emitted by the states. The probability for observing the data sequence $X = \{x_1,\dots,x_n\}$ given the corresponding state sequence $S = \{s_1,\dots,s_n\}$, the emission probability, is assumed to be of the form

$$P(X|S,H_H,I) = \prod_{t=1}^{n}P(x_t|s_t,H_H,I) = \prod_{t=1}^{n}e_{x_t s_t} \tag{18}$$

where we introduce the shorthand notation $e_{x_t s_t} = P(x_t|s_t,H_H,I)$ for the emission probabilities. The $e_{x_t s_t}$ are assumed known and independent of previous observations. The evidence for an HMM is obtained by marginalizing the joint probability distribution for X, the unknown transition probabilities $\mathbf{F}$, and the state sequence S according to

$$\begin{aligned} P(X|H_H,I) &= \int\sum_{S}P(X|S,\mathbf{F},I)P(S|\mathbf{F},I)P(\mathbf{F}|I)d\mathbf{F} \\ &= \int\sum_{S}P(X|S,I)P(S|\mathbf{F},I)P(\mathbf{F}|I)d\mathbf{F} \\ &= \sum_{S}P(X|S,I)\int P(S|\mathbf{F},I)P(\mathbf{F}|I)d\mathbf{F}. \end{aligned} \tag{19}$$

For notational compactness, the conditioning on H_H on the right-hand side is not written out, and $\mathbf{f}_0$ is included in $\mathbf{F}$. Recall that the joint probability $P(X|S,I)$ for the observed sequence is given by (18). We see that the final integral over $\mathbf{F}$ in (19) is the evidence (14) for a standard, non-hidden, Markov model. The total evidence (19) is thus

$$\begin{aligned} P(X|H_H,I) &= \frac{(K-1)!^K}{K}\sum_{S}P(X|S,I) \\ &\quad \times \prod_{l}\frac{\prod_{k=1}^{K}\Gamma\left(m_{kl}^S+1\right)}{\Gamma\left(m_l^S+K\right)}. \end{aligned} \tag{20}$$

Note here that the number of transitions m_{kl}^S and m_l^S depends on the state sequence S. Since an exact computation of (20) requires a summation over all possible sequences, this is generally not feasible. Instead, we propose an estimate that approximately evaluates all possible sequence paths.

We use the standard Markov model, but instead of increasing a single counter m_{kl} by one at a particular observation time, all counters are increased by an amount less than one, with the total at each time step summing to one. The amount that we add to m_{kl} at time t is the probability for the state transition l to k given the observations at t and $t-1$. An interesting interpretation of this approach is obtained by noting that this corresponds to a situation where all possible sequence paths are traversed simultaneously, with each state being occupied by a certain percentage $P(s_t = k, s_{t-1} = l|x_t, x_{t-1}, I)$. We then use the expression for the evidence (14) from the standard Markov model but instead of m_{kl} use

$$\begin{aligned} \hat{m}_{kl}(t) &= \hat{m}_{kl}(t-1) + P(s_t = k, s_{t-1} = l|x_t, x_{t-1}, I) \\ &= \hat{m}_{kl}(t-1) + \frac{p(x_t|s_t = k, I)p(s_t = k|I)}{P(x_t|I)} \\ &\quad \times \frac{p(x_{t-1}|s_{t-1} = l, I)p(s_{t-1} = l|I)}{P(x_{t-1}|I)} \\ &= \hat{m}_{kl}(t-1) + \frac{e_{x_t s_t = k}e_{x_{t-1}s_{t-1}=l}}{\sum_{s_t}e_{x_t s_t}\sum_{s_{t-1}}e_{x_{t-1}s_{t-1}}}. \end{aligned} \tag{21}$$

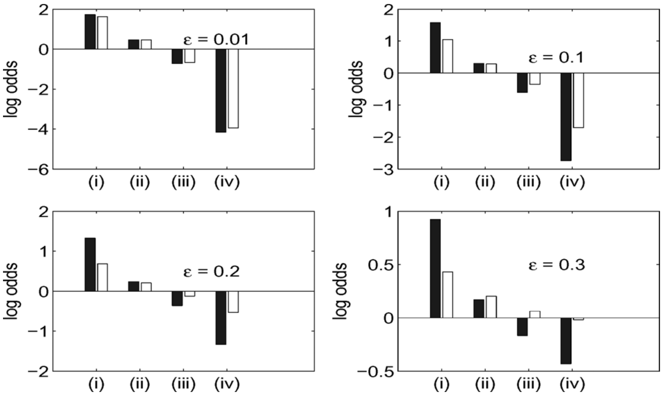

Fig. 1. Exact (dark) and approximative (light) log odds for tests between $X1 = \{1,1,1,1,1,1,1,1\}$ and the sequences (i) $X1$, (ii) {1,1,1,2,2,2,1,1}, (iii) {1,1,2,2,1,1,2,2}, and (iv) {1,2,1,2,1,2,1,2}.

In summary, we compute

$$P(X|H_H,I) = \int P(X|\mathbf{F},I)P(\mathbf{F}|I)d\mathbf{F} \qquad (22)$$

using the approximation

$$P(X|\mathbf{F},I) \approx f_{i0} \prod_{k=1}^{K}\prod_{l=1}^{K} f_{kl}^{\hat{m}_{kl}} \qquad (23)$$

with $\hat{m}_{kl}$ given by (21).

Test 5 (Same Versus Different HMM): To test whether or not two observed sequences have been generated by the same HMM, the odds (16) are calculated using the estimates $\hat{m}_{kl}$ from (21). The parameter $c_1 = \prod_{k=1}^{K}\Gamma(\tilde{m}_{k0}+1)$, where $\tilde{m}_{k0} = \delta_{s_0^1=k} + \delta_{s_0^2=k}$, is approximated using

$$\delta_{s_t=k} \approx P(s_t=k|x_t,I) = \frac{P(x_t|s_t=k,I)}{\sum_{s_t=1}^{K} P(x_t|s_t,I)}. \qquad (24)$$

Test 6 (Test of Independence): To test whether a data sequence supports a hidden independent multinomial model in favor of a general HMM, the odds (17) are computed using $\hat{m}_{kl}$ from (21), and $\delta_{s_n=k}$ as above.

D. Freely Available Software Implementations of the Tests

We have implemented the tests described here in Matlab. The tests can be downloaded from http://www.signal.uu.se/Staff/mj/pub/MarkovTests.zip. In implementing the tests, we have used the log odds without explicit computation of the Gamma function. We then obtain exact results, even for large numbers of observations.

III. Examples

Test 5 was compared with an exact Bayesian test based on the evidences in (19). In the comparison, a sequence $X1 = \{1,1,1,1,1,1,1,1\}$ was tested against sequences of increasing dissimilarity (see Fig. 1). The underlying Markov chain had two possible states, and the emission probabilities were $e_{x_k=i,s_k=i} = 1-\epsilon$, for $i = 1, 2$, and $e_{x_k=i,s_k=j} = \epsilon$, for $i \neq j$. We examined the parameter values $\epsilon \in \{0.01, 0.1, 0.2, 0.3\}$.

We note in Fig. 1 that the approximation is accurate for small ϵ. For larger ϵ, i.e., for less informative sequences, it becomes more cautious and favours similar HMMs.

IV. Comments

The standard procedure for model selection today is chi-square tests using maximum likelihood estimates (MLEs) for the transition probabilities. Besides the *ad hoc* status of chi-square [5], the procedure to estimate transition probabilities based on few data and treating them as known can be dangerously misleading. The MLE for a transition probability f_{ji} is m_{ji}/m_i, which will yield zero probability for non-observed state transitions. If we have only observed a short data series, this is of course absurd. Using the approach we have taken here, the transition probabilities are not estimated at all; rather, we take all possible values into account, weighted by their respective probability. It can finally be noted that if we would wish to make an estimate of f_{ji}, the posterior expectation becomes $(m_{ji}+1)/(m_i+K)$.

Appendix

We wish to integrate the state probabilities out of the expression for the prior (5) for the sequence multiplied by the likelihood (4). We omit the normalization constant C in the prior and solve the more general integral expression

$$J(q) = \int_0^\infty \cdots \int_0^\infty f_1^{m_1}\cdots f_K^{m_K}\delta(f_1+\ldots+f_K-q)d\mathbf{f}. \quad (25)$$

We note that the Laplace transform of $J(q)$ is

$$\int_0^\infty \cdots \int_0^\infty e^{-s(f_1+\ldots+f_K)} f_1^{m_1}\cdots f_K^{m_K} d\mathbf{f} = \frac{m_1!\cdots m_K!}{s^{m+K}} \quad (26)$$

where $m = \sum_{i=1}^{K} m_i$. Taking the inverse transform yields

$$J(q) = \frac{m_1!\cdots m_K!}{(m+K-1)!}q^{m+K-1}. \qquad (27)$$

Using $q = 1$, we obtain

$$J(1) = \frac{m_1!\cdots m_K!}{(m+K-1)!} = \frac{\prod_{l=1}^{K}\Gamma(m_l+1)}{\Gamma(m+K)}. \qquad (28)$$

References

[1] I. J. Good, "A Bayesian significance test for multinomial distributions," *J. R. Statist. Soc. B*, vol. 29, no. 3, pp. 399–431, Mar. 1967.

[2] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Mach. Learn.*, vol. 9, pp. 309–347, 1992.

[3] M. Ramoni, P. Sebastiani, and P. Cohen, "Bayesian clustering by dynamics," *Mach. Learn.*, vol. 47, pp. 91–121, 2002.

[4] J.-T. Chien and S. Furui, "Predictive hidden Markov model selection for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 377–387, May 2005.

[5] E. T. Jaynes, *Probability Theory—The Logic of Science*. Cambridge, U.K.: Cambridge Univ. Press, Apr. 2003.