Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With an Application to Speech Recognition

Fengchun Peng
Department of Mathematics and Statistics
University of Nebraska, Lincoln

Robert A. Jacobs
Department of Brain and Cognitive Sciences
University of Rochester

Martin A. Tanner
Department of Statistics
Northwestern University

September 1995

This paper has been accepted for publication in the *Journal of the American Statistical Association*

We wish to thank J. Kolassa, R. Neal, B. Yakir, two anonymous referees, and the Associate Editor for their insightful comments, and V. de Sa for providing us with the speech dataset. M. Tanner was supported by NIH research grant RO1-CA35464. R. Jacobs was supported by NIH research grant R29-MH54770. F. Peng was supported by NIH research grant T32-CA09667.

Abstract

Machine classification of acoustic waveforms as speech events is often difficult due to context-dependencies. A vowel recognition task with multiple speakers is studied in this paper via the use of a class of modular and hierarchical systems referred to as mixtures-of-experts and hierarchical mixtures-of-experts models. The statistical model underlying the systems is a mixture model in which both the mixture coefficients and the mixture components are generalized linear models. A full Bayesian approach is used as a basis of inference and prediction. Computations are performed using Markov chain Monte Carlo methods. A key benefit of this approach is the ability to obtain a sample from the posterior distribution of any functional of the parameters of the given model. In this way, more information is obtained than provided by a point estimate. Also avoided is the need to rely on a normal approximation to the posterior as the basis of inference. This is particularly important in cases where the posterior is skewed or multimodal. Comparisons between a hierarchical mixtures-of-experts model and other pattern classification systems on the vowel recognition task are reported. The results indicate that this model showed good classification performance, and also gave the additional benefit of providing for the opportunity to assess the degree of certainty of the model in its classification predictions.

1 Introduction

Psychological studies of human speech perception have documented people's extraordinary abilities to categorize acoustic waveforms as speech events. These abilities seem all the more impressive when we consider that people can accurately perceive speech while listening to unfamiliar voices in noisy surroundings with widely varying rates of speech production. How we do it, and how we can program computers to do it, are topics of much recent research.

Advances in our understanding of speech perception have been achieved through the use of a variety of complementary research programs. Psychoacousticians, perceptual and cognitive psychologists, and psycholinguists study the perceptual, cognitive, and linguistic mechanisms that allow people to process speech in an efficient and seemingly effortless manner (O'Shaughnessy, 1987). Related research is pursued by statisticians and computer scientists who attempt to build machines that can recognize speech in a wide range of environments (Rabiner and Juang, 1993). These research paths have converged in pinpointing the factors that make speech perception a challenging task. Several of these factors stem from the fact that speech production is highly context dependent (O'Shaughnessy, 1987). For example, it is often difficult to partition a speech waveform into the acoustic segments that correspond to distinct phonetic segments. This is primarily due to coarticulation; the vocal articulatory movements for successive sounds overlap in time such that the motions used to produce a particular phonetic segment are dependent on the preceding and succeeding phonetic segments. As a second example, identical phonemes are spoken differently by different speakers, and by the same speaker at different points in time, yet must be classified as equivalent events. Differences in speakers' speech are often correlated with speakers' gender, age, and dialect.

There are at least three common approaches to overcoming the problems associated with context-dependency (Rabiner and Juang, 1993). One approach is to detect invariant features. For example, while some features of the acoustic waveform corresponding to a phonetic segment may vary with a speaker's speech characteristics, there may be other acoustic features that are common to all speakers. A related approach is to normalize the acoustic waveform for the context. There may exist, for example, a transformation that removes the features that arise due to the speaker's characteristics. A third approach is to use knowledge of context to aid in the classification of speech events. For example, different classifiers may be used depending on the gender of the speaker. In addition to classifying the speech events, this approach must also classify the context.

This paper considers the application of a class of statistical models to a speech recognition task. The models classify speech events by detecting features that are invariant to speakers' speech characteristics. The data in Figure 1 represent instances of vowels spoken by a large number of speakers. The instances were taken from utterances of ten words, each of which began with an "h", contained a vowel in the middle, and ended with a "d". These data are from seventy-five speakers who uttered each word twice, with the words in different random orders for each presentation. Thirty-two of the speakers were male adults, twenty-eight were female adults, and fifteen were children. A spectral analysis was performed on each utterance; the portion of the spectrogram corresponding to the vowel was hand segmented, and the first two formants were extracted from the middle portion of the segmented region. Formants are the vocal tract's resonant frequencies. Different placements of the speech articulators, corresponding to different vowels, alter the vocal tract's shape and, thus, its frequency response. The horizontal and vertical axes of Figure 1 give the first and second formant values (properly normalized) for each vowel instance. Ten classes of vowels are represented in the figure; the individual data points in the figure correspond to the instances, and are labeled with the digits zero through nine so as to indicate the vowel class to which they belong. For each vowel class, Table 1 gives the word from which the vowel was segmented, and the digit used to label the class in the figure. This data was originally collected by Peterson and Barney (1952) and is a benchmark database in the speech recognition literature.

We consider the task of classifying the vowel instances on the basis of the values of each instance's first and second formants. We address this problem using a class of modular and hierarchical systems known as mixtures-of-experts (ME) and hierarchical mixtures-of-experts (HME) models. These models attempt to solve problems using a "divide-and-conquer"

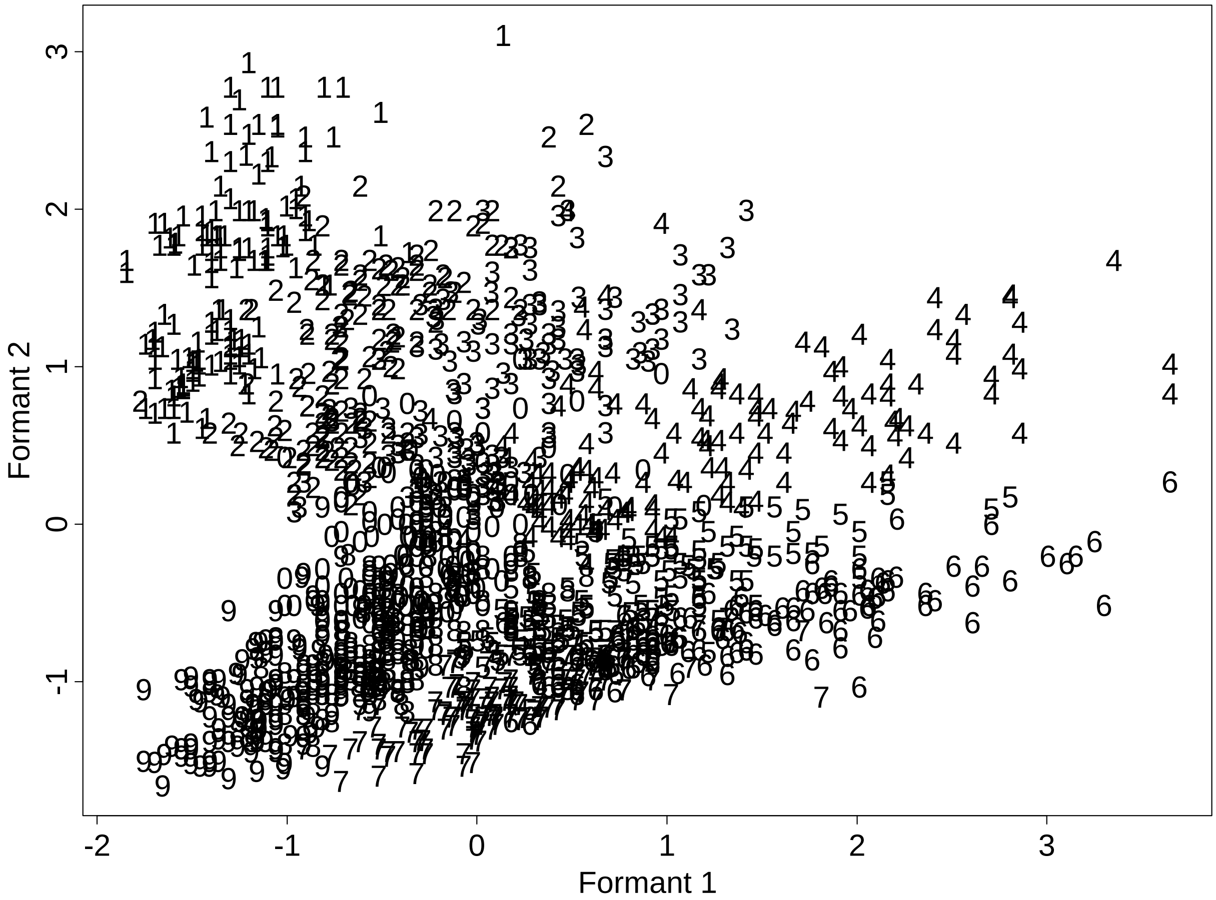

Figure 1: The horizontal and vertical axes give the values of the first and second formants respectively. Each data point is labeled with a digit (0-9) that indicates the vowel class to which the data point belongs.

Vowel Class	Word	Label
[ϖ]	heard	0
[i]	heed	1
[I]	hid	2
[ε]	head	3
[æ]	had	4
[Λ]	hud	5
[a]	hod	6
[ς]	hawed	7
[U]	hood	8
[u]	who'd	9

Table 1: For each vowel class, this table gives the word from which the vowel was segmented, and the digit used to label the class.

strategy; that is, complex problems are decomposed into a set of simpler sub-problems. We assume that the data can be adequately summarized by a collection of functions, each of which is defined over a local region of the domain. The approach adopted here attempts to allocate different modules to summarize the data located in different regions.

ME and HME models can be characterized as fitting a piecewise function to the data. The data are assumed to form a countable set of paired variables $\mathcal{X} = \{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t=1}^{T}$, where $\mathbf{x}$ is a vector of explanatory variables, also referred to as covariates, and $\mathbf{y}$ is a vector of responses. ME and HME models divide the covariate space, meaning the space of all possible values of the explanatory variables, into regions, and fit simple surfaces to the data that fall in each region. Unlike many other piecewise approximators, these models use regions that are not disjoint. The regions have "soft" boundaries meaning that data points may lie simultaneously in multiple regions. In addition, the boundaries between regions are themselves simple parameterized surfaces whose parameter values are estimated from the data.

ME and HME models combine properties of generalized linear models with those of mixture models. Like generalized linear models, they are used to model the relationship between a set of covariate and response variables. Typical applications include regression and binary or multiway classification. Unlike standard generalized linear models, however, they assume that the conditional distribution of the responses (given the covariates) is a finite mixture distribution. Because ME and HME models assume a finite mixture distribution, they provide a motivated alternative to non-parametric models, and provide a richer class of distributions than standard generalized linear models.

This paper proposes the use of Markov chain Monte Carlo methodology for inference in the context of the ME and HME models. Sections 2 and 3 present the mixtures-of-experts and hierarchical mixtures-of-experts models, respectively. Section 4 presents a Bayesian approach for training an HME model on a multiway classification problem. Section 5 reports on the application of the methodology to the speech recognition task.

2 Mixtures-of-Experts Model

To motivate the mixtures-of-experts model (Jacobs, Jordan, Nowlan, and Hinton, 1991), assume that the process generating the data is decomposable into a set of sub-processes defined on (possibly overlapping) regions of the covariate space. For each data item, a sub-process is selected, based on the covariate $\mathbf{x}^{(t)}$, and the selected sub-process maps $\mathbf{x}^{(t)}$ to the response $\mathbf{y}^{(t)}$. More precisely, data are generated as follows. For each covariate $\mathbf{x}^{(t)}$,

- a label i is chosen from a multinomial distribution with probability $P(i|\mathbf{x}^{(t)}, V)$, where $V = [\mathbf{v}_1, \ldots, \mathbf{v}_I]$ is the matrix of parameters underlying the multinomial distribution;
- a response $\mathbf{y}^{(t)}$ is generated with probability $P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_i, \Phi_i)$, where U_i is a parameter matrix and Φ_i represents other (possibly nuisance) parameters, for $i = 1, \ldots, I$.

To more explicitly relate this approach to the generalized linear models framework, we assume that the conditional probability distribution $P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_i, \Phi_i)$ is a member of the exponential family of distributions. The expected conditional value of the response $\mathbf{y}^{(t)}$, denoted $\boldsymbol{\mu}_i^{(t)}$, is defined to be a generalized linear function f of $\mathbf{x}^{(t)}$ and parameter matrix U_i. The quantities $\boldsymbol{\eta}_i = f^{-1}(\boldsymbol{\mu}_i)$ and Φ_i are, respectively, the natural parameter and dispersion parameter of the response's conditional distribution $P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_i, \Phi_i)$. The total probability of generating $\mathbf{y}^{(t)}$ from $\mathbf{x}^{(t)}$ is given by the mixture density

$$P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, \boldsymbol{\Theta}) = \sum_i P(i|\mathbf{x}^{(t)}, V) P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_i, \Phi_i), \tag{1}$$

where $\boldsymbol{\Theta} = [\mathbf{v}_1, \ldots, \mathbf{v}_I, U_1, \ldots, U_I, \Phi_1, \ldots, \Phi_I]^T$ is the matrix of all parameters. Assuming independently distributed data, the total probability of the dataset $\mathcal{X}$ is the product of T such densities, with the likelihood given by:

$$L(\boldsymbol{\Theta}|\mathcal{X}) = \prod_t \sum_i P(i|\mathbf{x}^{(t)}, V) P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_i, \Phi_i). \tag{2}$$

Thus the ME model can be viewed as a mixture model in which the mixing weights and the mixed distributions are dependent on the covariate $\mathbf{x}$.

Figure 2 presents a graphical representation of the ME model. The model consists of I modules referred to as *expert networks*. These networks approximate the data within each region of the covariate space: expert network i maps its input, the covariate vector $\mathbf{x}$, to an output vector $\boldsymbol{\mu}_i$. It is assumed that different expert networks are appropriate in different regions of the covariate space. Consequently, the model requires a module, referred to as a *gating network*, that identifies for any covariate $\mathbf{x}$, the expert or blend of experts whose output is most likely to approximate the corresponding response vector $\mathbf{y}$. The gating network outputs are a set of scalar coefficients g_i that weight the contributions of the various experts. For each covariate $\mathbf{x}$, these coefficients are constrained to be nonnegative and to sum to one. The total output of the model, given by

$$\boldsymbol{\mu} = \sum_{i=1}^{n} g_i \boldsymbol{\mu}_i, \tag{3}$$

is a convex combination of the expert outputs for each $\mathbf{x}$.

From the perspective of statistical mixture modeling, we identify the gating network with the selection of a particular sub-process. That is, the gating outputs g_i are interpreted

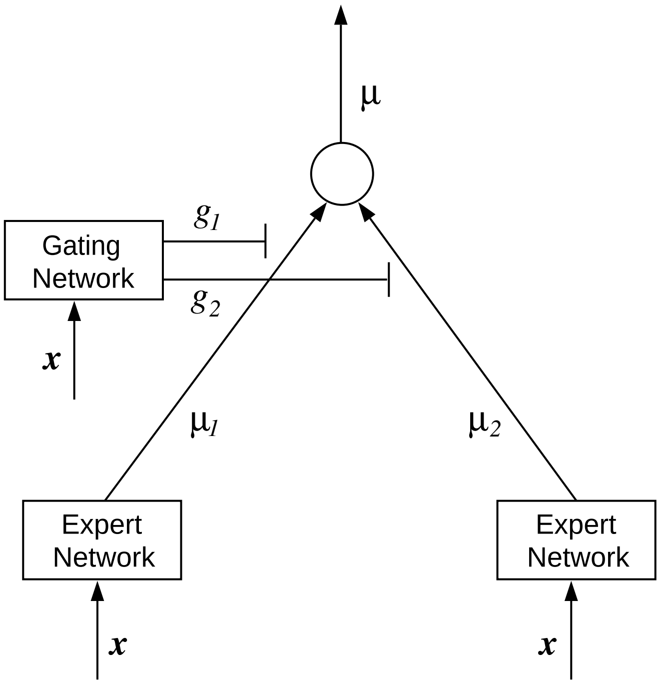

Figure 2: A mixtures-of-experts model.

as the covariate-dependent, multinomial probabilities of selecting sub-process i. Different expert networks are identified with different sub-processes; each expert models the covariate-dependent distributions associated with its corresponding sub-process.

The expert networks map their inputs to their outputs in a two-stage process. During the first stage, each expert multiplies the covariate vector $\mathbf{x}$ by a matrix of parameters. (The vector $\mathbf{x}$ is assumed to include a fixed component of one to allow for an intercept term.) For expert i, the matrix is denoted as U_i and the resulting vector is denoted as $\boldsymbol{\eta}_i$:

$$\boldsymbol{\eta}_i = U_i \mathbf{x}. \tag{4}$$

During the second stage, $\boldsymbol{\eta}_i$ is mapped to the expert output $\boldsymbol{\mu}_i$ by a monotonic, continuous nonlinear function f.

The selection of the nonlinear function f is based on the nature of the problem. For regression problems, f may be taken as the identity function (i.e. the experts are linear).

Moreover, the probabilistic component of the model may be Gaussian. In this case, the likelihood is a mixture of Gaussians:

$$L(\mathbf{\Theta}|\mathcal{X}) = \prod_t \sum_i g_i^{(t)} \, |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}[\mathbf{y}^{(t)} - \boldsymbol{\mu}_i^{(t)}]^T \Sigma_i^{-1} [\mathbf{y}^{(t)} - \boldsymbol{\mu}_i^{(t)}]}. \tag{5}$$

The expert output $\boldsymbol{\mu}_i^{(t)}$ and expert dispersion parameter Σ_i are interpreted as the mean and covariance matrix of expert i's Gaussian distribution for response $\mathbf{y}^{(t)}$ given covariate $\mathbf{x}^{(t)}$. The output of the entire model $\boldsymbol{\mu}^{(t)}$ is, therefore, interpreted as the expected value of $\mathbf{y}^{(t)}$ given $\mathbf{x}^{(t)}$.

For binary classification problems, f may be the logistic function:

$$\mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{6}$$

In this case, η_i may be interpreted as the log-odds of "success" under a Bernoulli probability model. The probabilistic component of the model is assumed to be the Bernoulli distribution. The likelihood is a mixture of Bernoulli densities:

$$L(\mathbf{\Theta}|\mathcal{X}) = \prod_t \sum_i g_i^{(t)} \, [\mu_i^{(t)}]^{y^{(t)}} \, [1 - \mu_i^{(t)}]^{1-y^{(t)}}. \tag{7}$$

The quantity $\mu_i^{(t)}$ is expert i's conditional probability of classifying the covariate $\mathbf{x}^{(t)}$ as success, and μ is the expected success of $\mathbf{x}^{(t)}$. Other problems (e.g., multiway classification, counting, rate estimation, survival estimation) may suggest other choices for f. In all cases, we take the inverse of f to be the canonical link function for the appropriate probability model (McCullagh and Nelder, 1989).

The gating network also forms its outputs in two stages. During the linear stage, it computes the intermediate variables ξ_i as the inner product of the covariate vector $\mathbf{x}$ and the vector of parameters $\mathbf{v}_i$:

$$\xi_i = \mathbf{v}_i^T \mathbf{x}. \tag{8}$$

The ξ_i are mapped to the gating outputs g_i during the nonlinear stage. This mapping is performed by using a generalization of the logistic function:

$$g_i = \frac{e^{\xi_i}}{\sum_j e^{\xi_j}}. \tag{9}$$

Note that the inverse of this function is the canonical link function for a multinomial response model (McCullagh and Nelder, 1989).

In this paper, Bayesian inference regarding the gating and expert networks' parameters is performed using Markov chain Monte Carlo methods. In particular, we make use of the fact that the problem is greatly simplified by augmenting the observed data with the unobserved indicator variables indicating the expert in the model. This idea of using data augmentation (Tanner and Wong, 1987) in mixture problems, where the mixing parameters are unknown scalars, is discussed in detail in Diebolt and Robert (1994). Jordan and Jacobs (1994) showed how the parameters of the mixtures-of-experts (as well as the hierarchical mixtures-of-experts) model can be estimated using this augmentation idea in the context of the EM algorithm. Jordan and Xu (1993) presented an analysis of the convergence properties of the EM algorithm as applied to ME and HME models. A discussion of the relationship between ME models and neural networks is presented in Peng, Jacobs, and Tanner (1995), which also presents an ME model for regression and illustrates Bayesian learning in a nonlinear regression context.

3 Hierarchical Mixtures-of-Experts Model

The philosophy underlying the ME model is to solve complex modeling problems by dividing the covariate space into restricted regions, and to use different generalized linear models to fit the data in each region. If this "divide-and-conquer" approach is useful, then it seems desirable to pursue this strategy to its logical extreme. In particular, it seems sensible to divide the covariate space into restricted regions, and then to recursively divide each region into sub-regions. The hierarchical extension of the mixtures-of-experts model, referred to as the HME model, is a tree-structured model that implements this strategy (Jordan and Jacobs, 1994). In contrast to the single-level ME model, the HME model can summarize the data at multiple scales of resolution due to its use of nested covariate regions. Jordan and Jacobs (1994) have empirically found that models with a nested structure often outperform single-level models with an equivalent number of free parameters.

The HME model, unlike the ME model, contains multiple gating networks. The gating networks, located at the non-terminals of the tree, implement the recursive splitting of the covariate space. The expert networks are at the terminals of the tree. Because they are defined over relatively small regions of the covariate space, the experts can fit simple functions to the data (e.g., generalized linear models).

Figure 3 illustrates an HME model. For explanatory reasons, we limit our presentation to a two-level tree; the extension to trees of arbitrary depth and width is straightforward. The bottom-level of the tree contains a number of *clusters*, each itself an ME model (the tree in the figure contains two clusters, each enclosed in a box). The top-level of the tree contains an additional gating network that combines the outputs of each cluster. As a matter of notation, we use the letter i to index branches at the top-level, and the letter j to index branches at the bottom-level. The expert networks map the covariate $\mathbf{x}$ to the output vectors $\boldsymbol{\mu}_{ij}$. The output of the i^{th} cluster is given by

$$\boldsymbol{\mu}_i = \sum_j g_{j|i} \boldsymbol{\mu}_{ij} \tag{10}$$

where the scalar $g_{j|i}$ is the output of the gating network in the i^{th} cluster corresponding to the j^{th} expert. The output of the model as a whole is given by

$$\boldsymbol{\mu} = \sum_i g_i \boldsymbol{\mu}_i \tag{11}$$

where the scalar g_i is the output of the top-level gating network corresponding to the i^{th} cluster.

The probabilistic interpretation of the HME model is a simple extension of that given to the ME model. In short, we assume that the process that generates the data involves a nested sequence of decisions, based on the covariate $\mathbf{x}^{(t)}$. The outcome of this sequence is the selection of a sub-process that maps $\mathbf{x}^{(t)}$ to the response $\mathbf{y}^{(t)}$. In a two-level tree, for example, we assume that a label i is chosen from a multinomial distribution with probability $P(i|\mathbf{x}^{(t)}, V)$, where V is the matrix of parameters underlying this distribution. Next, a label j is chosen from another multinomial distribution with probability $P(j|\mathbf{x}^{(t)}, V_i)$, where the parameters V_i underlying this distribution are dependent on the value of label i. Sub-process

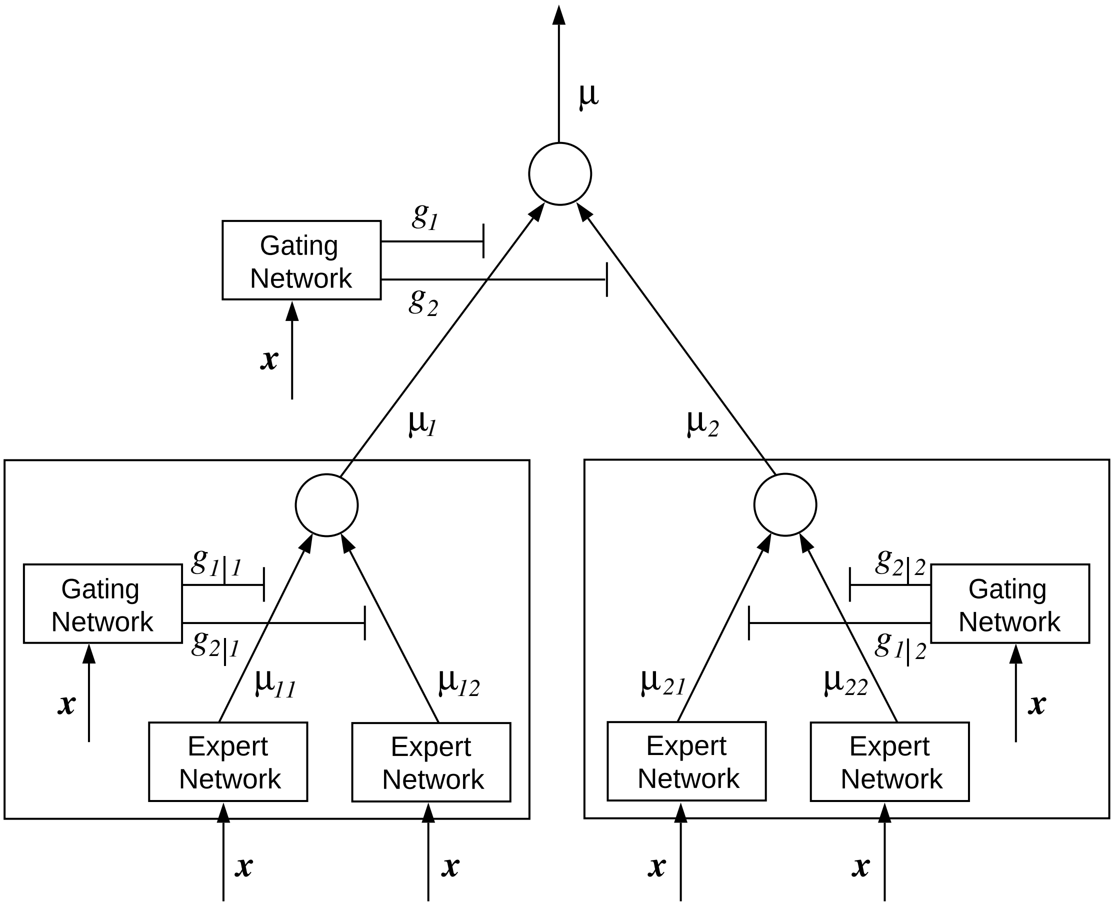

Figure 3: A hierarchical mixtures-of-experts model.

(i, j) generates a response $\mathbf{y}^{(t)}$ by sampling from a distribution with probability $P(\mathbf{y}^{(t)}| \boldsymbol{\mu}_{ij}, \Phi_{ij})$, where $\boldsymbol{\eta}_{ij} = f^{-1}(\boldsymbol{\mu}_{ij})$ is the natural parameter of this distribution and Φ_{ij} is its dispersion parameter. This latter probability may also be written as $P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})$. The total probability of generating $\mathbf{y}^{(t)}$ from $\mathbf{x}^{(t)}$ is given by the hierarchical mixture density

$$P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, \boldsymbol{\Theta}) = \sum_i P(i|\mathbf{x}^{(t)}, V) \sum_j P(j|\mathbf{x}^{(t)}, V_i) P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij}), \tag{12}$$

where $\boldsymbol{\Theta}$ is the matrix of all parameters.

To model this process with the HME model, we identify the top-level gating network with the selection of the label i, and the bottom-level gating networks with the selection of the label j. Different expert networks are identified with different sub-processes. To give just one example, the resulting likelihood function for a binary classification problem is a hierarchical mixture of Bernoulli densities:

$$L(\boldsymbol{\Theta}|\mathcal{X}) = \prod_t \sum_i g_i^{(t)} \sum_j g_{j|i}^{(t)} [\mu_{ij}^{(t)}]^{y^{(t)}} [1 - \mu_{ij}^{(t)}]^{1-y^{(t)}}. \tag{13}$$

The quantity μ_{ij} is expert (i, j)'s conditional probability of classifying the covariate $\mathbf{x}^{(t)}$ as success, and μ is the expected success of $\mathbf{x}^{(t)}$.

The HME model has some similarities to classification and regression tree models that have previously been proposed in the statistics literature. Like CART (Breiman, Friedman, Olshen, and Stone, 1984) and MARS (Friedman, 1991), the HME model is a tree-structured approach to piecewise function approximation. A major difference between the HME model and CART or MARS is in the way the covariate space is segmented into regions. Both CART and MARS use "hard" boundaries between regions; each data item lies in exactly one region. The HME model, in contrast, allows regions to overlap so that data items may reside in multiple regions. The boundaries between the regions are, therefore, said to be "soft." MARS is restricted to forming region boundaries that are perpendicular to one of the covariate space axes. Thus MARS is coordinate-dependent, i.e. it is sensitive to the particular choice of covariate variables used to encode the data. The HME model is not coordinate-dependent; boundaries between regions are formed along hyperplanes at arbitrary orientations in the covariate space. In addition, CART and MARS are nonparametric techniques; the HME model is a parametric model. Jordan and Jacobs (1994) compared the performance of the HME model with that of CART and MARS on a robot dynamics task, and found that the HME model (using the EM algorithm to estimate the parameters) yielded a modest improvement.

4 HME Model for Multiway Classification

A basic simplification is obtained by augmenting the observed data with the unobserved indicator variables indicating the expert in the model. If we define $z_i^{(t)} = 1$ with probability

$$h_i^{(t)} = \frac{P(i|\mathbf{x}^{(t)}, V)\sum_j P(j|\mathbf{x}^{(t)}, V_i)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})}{\sum_i P(i|\mathbf{x}^{(t)}, V)\sum_j P(j|\mathbf{x}^{(t)}, V_i)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})}, \tag{14}$$

and $z_{j|i}^{(t)} = 1$ with probability

$$h_{j|i}^{(t)} = \frac{P(j|\mathbf{x}^{(t)}, V_i)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})}{\sum_j P(j|\mathbf{x}^{(t)}, V_i)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})}, \tag{15}$$

then $z_{ij}^{(t)} = z_i^{(t)} \times z_{j|i}^{(t)}$ will take value 1 with probability

$$h_{ij}^{(t)} = \frac{P(i|\mathbf{x}^{(t)}, V)P(j|\mathbf{x}^{(t)}, V_i)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})}{\sum_i P(i|\mathbf{x}^{(t)}, V)\sum_j P(j|\mathbf{x}^{(t)}, V_i)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})}, \tag{16}$$

where U_{ij} is the matrix of parameters associated with the $(i,j)^{\text{th}}$ expert network, V is the matrix of parameters associated with the gating network at the top-level of the model, and V_i is the matrix of parameters associated with the gating network in the i^{th} cluster at the bottom-level. Let $\mathcal{X}' = \{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \mathbf{z}^{(t)})\}_{t=1}^T$ where $\mathbf{z}^{(t)}$ is the vector of indicator variables. Then the augmented likelihood for the HME model is

$$\begin{aligned} L(\mathbf{\Theta}|\mathcal{X}') &= \prod_t \prod_i \prod_j \left\{P(i|\mathbf{x}^{(t)}, V)P(j|\mathbf{x}^{(t)}, V_i)P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})\right\}^{z_{ij}^{(t)}} & (17) \\ &= \prod_t \prod_i \prod_j \left\{g_i^{(t)} g_{j|i}^{(t)} P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, U_{ij}, \Phi_{ij})\right\}^{z_{ij}^{(t)}} & (18) \end{aligned}$$

where g_i is the same as defined in (9) and

$$g_{j|i} = \frac{e^{\mathbf{v}_{ij}^T \mathbf{x}^{(t)}}}{\sum_k e^{\mathbf{v}_{ik}^T \mathbf{x}^{(t)}}}. \tag{19}$$

Restricting our attention to multiway classification problems, suppose that $\mathbf{y} = (y_1, \cdots, y_n)$ is the outcome from the n classes, taking the values either 0 or 1; $U_{ij} = (U_{ij1}, \ldots, U_{ijn})^T$ is the $n \times p$ matrix of parameters associated with the $(i,j)^{\text{th}}$ expert network; and $\Phi_{ij} = 1$ for all i and j. In this case the conditional probability model for $\mathbf{y}$ can be written as

$$P(\mathbf{y}|\mathbf{x}, U_{ij}, \Phi_{ij}) = \prod_{k=1}^{n} \mu_{ijk}^{y_k}. \tag{20}$$

If we express this probability model in the density form of the exponential family, the natural parameters are defined as

$$\begin{aligned} \eta_{ijk} &= \log \frac{\mu_{ijk}}{\mu_{ijn}} && (21) \\ &= U_{ijk}\mathbf{x}, && (22) \end{aligned}$$

and hence

$$\mu_{ijk} = \frac{e^{\eta_{ijk}}}{\sum_{k=1}^{n} e^{\eta_{ijk}}}. \tag{23}$$

As noted above, the inverse of (23) is the canonical link function for a multinomial response model (McCullagh and Nelder, 1989). It follows from (9), (19), and (23) that the augmented likelihood (18) can be re-written as

$$\begin{aligned} L(\mathbf{\Theta}|\mathcal{X}') &= \prod_t \prod_i \prod_j \prod_k \left\{ g_i^{(t)} g_{j|i}^{(t)} \mu_{ijk}^{(t)\, y_k^{(t)}} \right\}^{z_{ij}^{(t)}} && (24) \\ &= \prod_t \prod_i \prod_j \prod_k \left\{ \frac{e^{\mathbf{v}_i^T \mathbf{x}^{(t)}}}{\sum_{k=1}^{I} e^{\mathbf{v}_i^T \mathbf{x}^{(t)}}} \frac{e^{\mathbf{v}_{ij}^T \mathbf{x}^{(t)}}}{\sum_{k=1}^{J} e^{\mathbf{v}_{ik}^T \mathbf{x}^{(t)}}} \left(\frac{e^{U_{ijk}\mathbf{x}^{(t)}}}{\sum_{k=1}^{n} e^{U_{ijk}\mathbf{x}^{(t)}}} \right)^{y_k^{(t)}} \right\}^{z_{ij}^{(t)}}. && (25) \end{aligned}$$

Since this equation with independent normal priors (mean = 0; variance = σ_0^2) on $\mathbf{\Theta}$ does not yield standard densities for the full conditionals, we apply the approach of Müller (1995) to draw the posterior sample for the top-level gating network parameter matrix V, the bottom-level gating network parameter matrices V_i, and the bottom-level expert network parameter matrices U_{ij}. In particular, to draw a deviate from a full conditional we use the Metropolis algorithm. As an example, consider the top-level gating network parameters. A candidate value for the next point in the Metropolis chain ($V^{(k+1)}$) is drawn from the multivariate normal distribution with the current sample values as its mean and a diagonal variance-covariance matrix to allow for variation around the current sample values, i.e.

$V^{(k+1)} \longleftarrow N(V^{(k)}, \gamma^2 \mathbf{I})$. This candidate value is accepted or rejected according to the standard Metropolis scheme (Tanner, 1993). This Metropolis algorithm is iterated and the final value in this chain is treated as a deviate from the full conditional distribution.

5 The Speech Recognition Problem

The HME model was applied to the speech recognition task described above. The dataset consists of 1494 data items (75 speakers $\times$ 10 vowel classes $\times$ 2 presentations of each vowel; note that three speakers' data for the vowel [ς] are missing from the dataset). From this collection, 149 items were randomly selected and assigned to a training set; the remaining 1345 were assigned to a prediction set. The HME model consisted of three gating networks and ten expert networks arranged in a two-level tree structure. The bottom level had two clusters, each with one gating network and five expert networks. The outputs of these two clusters were combined by a gating network located at the top level.

The model was trained via the Gibbs sampler algorithm described above and also via the EM algorithm. For the Gibbs sampler, ten chains of length 7,500 were created. It is known that posterior distributions based on mixture models can be highly multimodal. We used two approaches in the context of the Gibbs sampler to address this aspect of the problem. In one approach we allowed for mode jumping as the chain moved through the parameter space—hopefully avoiding oversampling in the neighborhoods of isolated minor modes. To facilitate this, after every 100 iterations of the Gibbs sampler the parameter vector $\mathbf{\Theta}$ was selected at random from the modes defined by twenty independent and randomly initialized runs of the EM algorithm. This candidate value was compared to the current point in the Gibbs sampler chain using a Metropolis test based on the observed posterior. The data in Table 2 and Figures 4-6 are based on the final 500 iterations of each chain. The Metropolis algorithm used to sample from the full conditional distributions was run for 40 iterations, with the variance on the normal distribution equal to 1.0. The variance of each of the normal priors on the components of the expert and gating networks parameters was equal to 50,000.

It is noted that this mode jumping approach introduces a possible approximation due to the fact that the transition kernel of the Metropolis test is discrete, while the posterior distribution is continuous. The expectation is that this algorithm will quickly converge to a good approximation. As a check, we adopted a second approach based on Gelman and Rubin (1992). In particular, we formed a mixture of normals to approximate the observed posterior. We then drew a sample of size 1000 from this approximation. Importance sampling was then used to obtain ten starting points for ten Gibbs sampler chains which were run without mode jumping. This second approach confirmed the results of the first approach; that is, the two approaches yielded qualitatively similar outcomes. An alternative technique for sampling from multimodal posteriors is presented in Neal (1994).

The convergence of the ten chains was evaluated using the technique of Gelman and Rubin (1992). Intuitively, this Gelman and Rubin technique for assessing convergence compares the between-chain variation to the within-chain variation; an algorithm is said to converge when the between-chain and within-chain variations are comparable in size. To assess convergence of our Gibbs sampler algorithm, we used as input to the Gelman and Rubin technique the conditional predictive probability that a data item belongs to a particular vowel class [for vowel class j, this is $p(y_i = j|\mathbf{x}, \mathbf{\Theta})$]. The use of the conditional predictive probability when assessing convergence can also be found in Neal (1991). For the covariate values that we examined, the Gelman and Rubin R value did not exceed 1.1, meaning that there was no suggested evidence of lack of convergence for the present dataset and model.

Table 2 shows the classification performances of four systems on the data from the prediction set. The first two systems are the HME model trained with the Gibbs sampler algorithm and with the EM algorithm respectively. The number of correct classifications is an average over the ten chains for the first system, and over twenty instances of the second system where the instances differ due to the random initial settings of the system's parameter values. In both cases, the vowel class corresponding to the response variable with the largest expected value was used as the model's classification. The third and fourth systems are two versions of CART (Breiman, Friedman, Olshen, and Stone, 1984). In the first version, the splits of the covariate space were restricted to be perpendicular to one of the covariate axes; in the

Table 2: Classification performances of four systems on the data from the prediction set. The first two systems are the HME model trained with the Gibbs sampler algorithm and with the EM algorithm respectively; the last two systems are two versions of CART.

Category	Total in Category	Gibbs # Correct	EM # Correct	CART I # Correct	CART II # Correct
0	127	123.3	123.3	114	124
1	132	96.3	97.9	94	112
2	139	95.9	92.0	30	60
3	137	84.1	79.2	90	109
4	133	104.6	100.9	73	117
5	136	109.1	106.3	102	109
6	128	88.5	84.4	90	96
7	132	74.5	78.0	72	71
8	140	80.7	77.1	49	38
9	141	58.0	50.2	87	60
0-9	1345	915.0	889.3	801	896

second version, the splits were allowed to be linear combinations of the covariate variables. The HME model trained with the Gibbs sampler algorithm showed the best performance. Similar performance levels were achieved by the HME model trained with the EM algorithm and the second version of CART. The first version of CART showed the worst performance.

A major benefit of the sampling based approach to inference for HME models is the ability to assess the degree of certainty of the model in its classification. This is illustrated in Figures 4-6. The graphs in these figures show estimated posterior distributions of the probability that a data item belongs to a particular vowel class given the values of the covariates. The horizontal axis of each graph gives the probability and the vertical axis gives

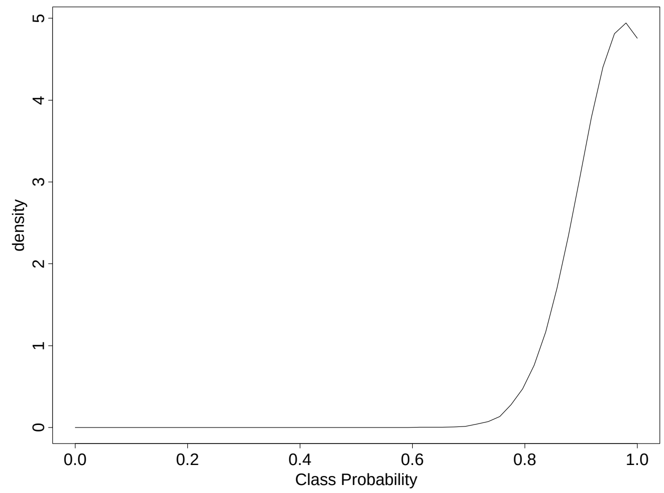

Figure 4: The estimated posterior distribution for the probability that the data item belongs to the correct vowel class one. Note that the mass is concentrated at unity, thus indicating that the model is confident in its prediction. This graph gives the result of one chain of the Gibbs sampler; the other nine chains produced nearly identical results. The Gelman and Rubin R value was equal to 1.0.

the density. The mean of each of the distributions presented in Figures 4-6 is the predictive probability of the corresponding class, i.e. $p(y_i = j|\mathbf{x}) = E[p(y_i = j|\mathbf{x}, \mathbf{\Theta})]$.

The first data item that we consider is a member of vowel class one. The graph in Figure 4 shows the estimated posterior distribution of the probability that this item belongs to the correct vowel class given the covariate values. This graph gives the result of one chain of the Gibbs sampler; the other nine chains produced nearly identical results. The distribution is highly skewed such that the majority of its mass is concentrated at a probability of unity. That is, the model is consistently and correctly confident that the data item belongs to vowel class one. Although not shown, the distributions corresponding to all other vowel classes have their mass concentrated at zero.

The second data item that we consider is a member of vowel class eight. Figure 5 shows the estimated posterior distributions of the probability that this item belongs to the correct vowel class, whereas Figure 6 shows the distributions for an incorrect vowel class (class nine).

The ten graphs in each of these figures correspond to the ten chains of the Gibbs sampler. Along the horizontal axis of each graph in these figures is a "v" indicating the expected value of the posterior distribution. The expected values are generally larger for the incorrect vowel class than they are for the correct vowel class meaning that the model has incorrectly classified this data item. Note, however, that the distributions for both vowel classes have their mass concentrated in the middle of the probability range suggesting that the model is not confident in its prediction for either of the vowel classes. The distribution for the remaining classes are not shown but have their mass concentrated near a probability of zero. Overall, the distributions for all the vowel classes indicate that the model is confident that the data item belongs to either vowel class eight or nine, though it is uncertain as to which of these two classes the item belongs.

6 Conclusions

This paper presented an approach for the full Bayesian analysis of ME and HME models. We presented the basic Gibbs sampler equations, as well as applied the methodology to a vowel recognition task. Comparisons of the classification performance of an HME model trained via the Gibbs sampler algorithm with that of the HME model trained via the EM algorithm and with that of two versions of CART were conducted. The results indicate that the HME model trained via the Gibbs sampler yielded good performance, and also gave the additional benefit of providing for the opportunity to assess the degree of certainty of the model in its classification predictions. This research has extended the work of Diebolt and Robert (1994) to the case where the mixing coefficients and mixture components are generalized linear models. We feel that the idea of "soft" boundaries between regions of the covariate space shows promise, and further study of this idea by the statistical community is warranted. Because the models can be used in any of the domains to which generalized linear models are typically applied, we see this work as having broad implications. Future work, for example, may apply this class of models to problems of censored regression data, Poisson regression, image analysis, and DNA structure prediction. Also of interest is the development

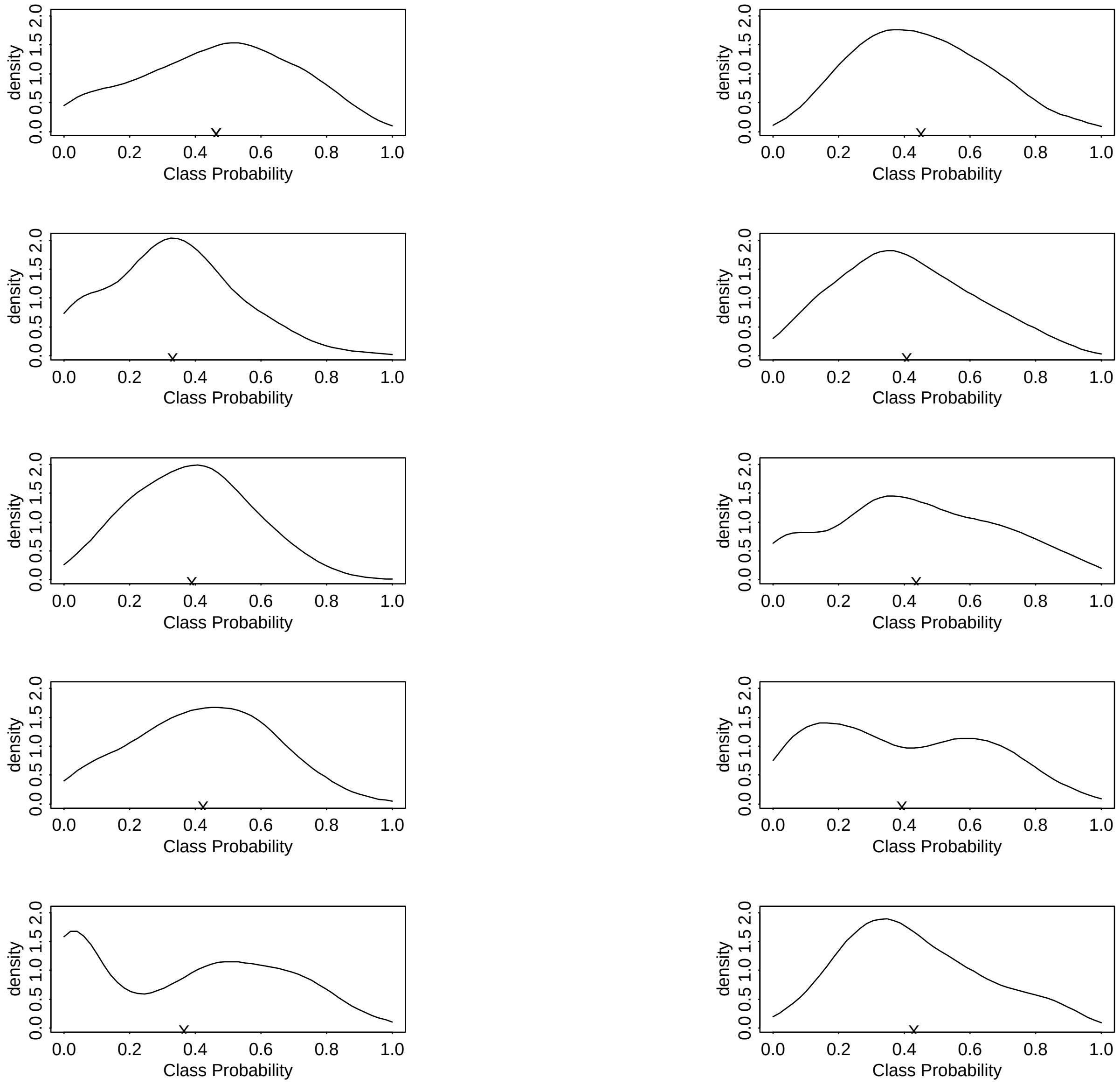

Figure 5: Ten estimated posterior distributions for the probability that the data item belongs to the correct vowel class eight. The "v" along the horizontal axis of each graph indicates the expected value of the distribution. Note that the mass is concentrated in the middle of the probability range, thus indicating that the model is not confident in its prediction. The Gelman and Rubin R value was equal to 1.024.

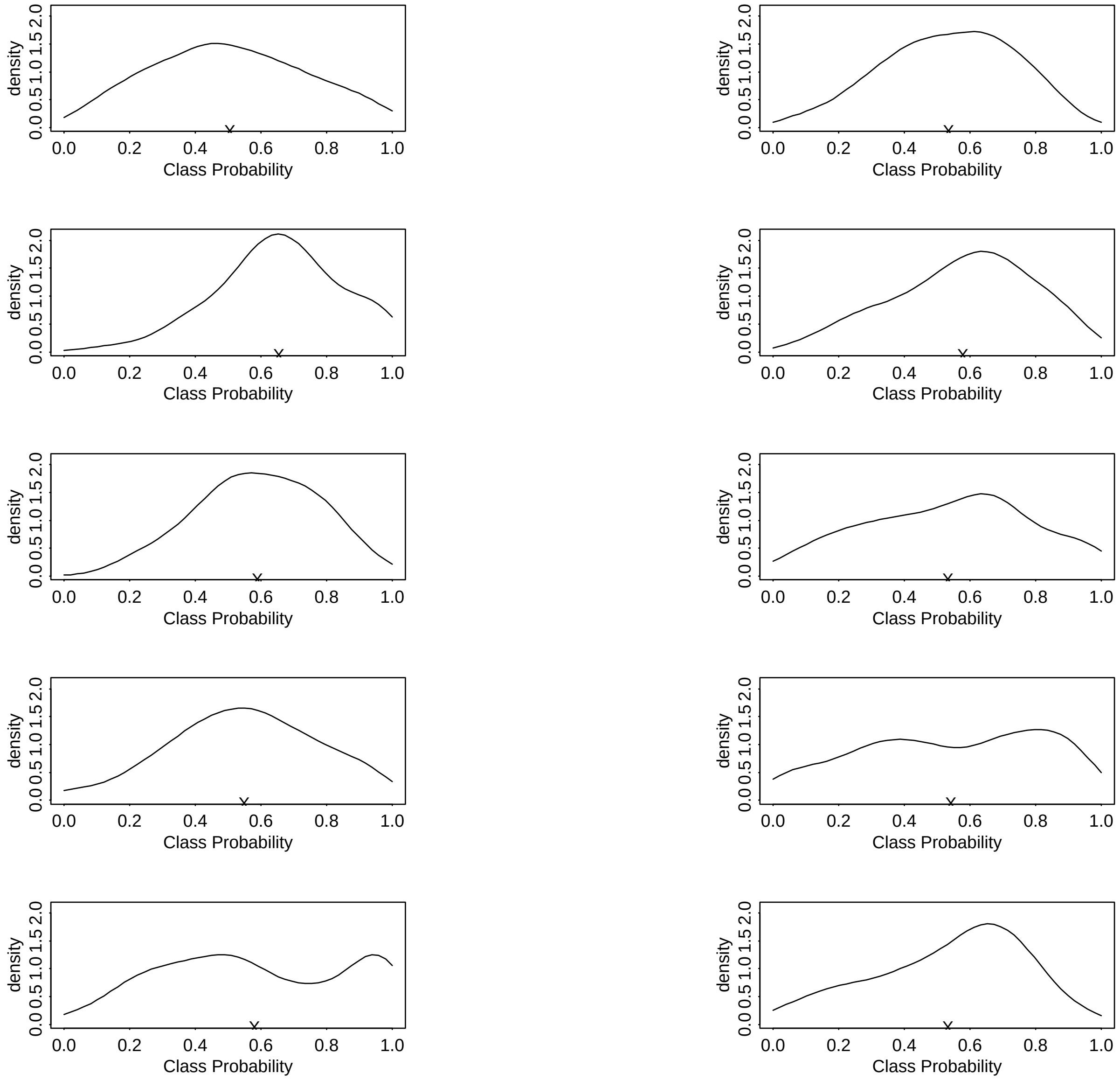

Figure 6: Ten estimated posterior distributions for the probability that the data item belongs to the incorrect vowel class nine. The "v" along the horizontal axis of each graph indicates the expected value of the distribution. Note that the mass is concentrated in the middle of the probability range, thus indicating that the model is not confident in its prediction. The Gelman and Rubin R value was equal to 1.021.

of algorithms that automatically specify the number of levels and the branching factors at each level of a model.

References

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984) *Classification and Regression Trees.* Belmont, CA: Wadsworth International Group.

Diebolt, J. & Robert, C.P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B*, 56, 363-375.

Friedman, J.H. (1991) Multivariate adaptive regression splines. *The Annals of Statistics*, 19, 1-141.

Gelman, A. & Rubin, D.B. (l992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., & Hinton, G.E. (1991) Adaptive mixtures of local experts. *Neural Computation*, 3, 79-87.

Jordan, M.I. & Jacobs, R.A. (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6, 181-214.

Jordan, M.I. & Xu, L. (1993) Convergence results for the EM approach to mixtures-of-experts architectures. Technical Report 9303, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models.* London: Chapman and Hall.

Müller P. (1995) Metropolis Posterior Integration Schemes. *Journal of the American Statistical Association*, in press.

Neal, R.M. (1991) Bayesian mixture modeling by Monte Carlo simulation. Technical Report CRG-TR-91-2, Department of Computer Science, University of Toronto.

Neal, R.M. (1994) Sampling from multimodal distributions using tempered transitions. Technical Report No. 9421, Department of Statistics, University of Toronto.

O'Shaughnessy, D. (1987) *Speech Communication: Human and Machine.* Reading, MA: Addison-Wesley Publishing Company.

Peng, F., Jacobs, R.A., & Tanner, M.A. (1995) Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts architectures. Technical Report, Department of Biostatistics, University of Rochester Medical Center.

Peterson, G.E. & Barney, H.L. (1952) Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, 24, 175-184.

Rabiner, L. & Juang, B.-H. (1993) *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice Hall.

Tanner, M.A. (1993) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* New York: Springer-Verlag.

Tanner, M.A. & Wong, W.H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528-550.