BAYESIAN LATENT VARIABLE MODELS FOR SPEECH RECOGNITION

Jen-Tzung Chien[†] *and Peng Liu*[‡]

[†]Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan
[‡]Sohu.com Inc., Beijing, China

jtchien@nctu.edu.tw & pengliubj8588@sohu-inc.com

ABSTRACT

We present a Bayesian framework to learn prior and posterior distributions for latent variable models. Our goal is to deal with model regularization and achieve desirable prediction using heterogeneous speech data. A variational Bayesian expectation-maximization algorithm is developed to establish a latent variable model based on the *exponential family distributions*. This algorithm does not only estimate model parameters but also their *hyperparameters* which reflect the *model uncertainties*. The uncertainty is compensated to construct a variety of regularized models. We realize this *full Bayesian* framework for uncertainty decoding of speech signals. Compared to maximum likelihood method and Bayesian approach with heuristically-selected hyperparameters, the proposed method achieves higher speech recognition accuracy especially in case of sparse and noisy training data.

Index Terms— Bayesian Learning, Exponential Family, Latent Variable Model, Speech Recognition

1. INTRODUCTION

In general pattern recognition, we aim to establish a concise and analytically tractable model from the collected data. However, in real world, training data may be abundant, sparse, noisy, mislabeled, misaligned, mismatched or ill-posed. The probabilistic models may be improperly-assumed, underestimated or overestimated. It is crucial to build a scalable and regularized model from heterogeneous training data. Considering the uncertainty in model construction is essential to achieve model robustness in adverse environments. Bayesian learning provides a powerful mechanism to fulfil model regularization [1] for pattern recognition including speech recognition [5][13][14][16][17], document categorization [2] and many others [12]. An important issue in Bayesian learning is to determine uncertainty [7][11] or prior distribution for a specific task at hand. This issue can be tackled as follows. First, we may adopt a prior model which is mathematically attractive for model inference. Then, we select the prior distribution or estimate its hyperparameters. Prior selection can be done either subjectively based on some background knowledge or objectively via empirical Bayes where prior is learnt from data. More attractively, the full Bayesian framework [1][12] could be applied to estimate hyperparameters by maximizing an *evidence function* which is calculated from training data. *No validation data* is needed. This framework has been successfully developed for linear regression/classification models [1], neural network model [12], support vector machine [10] and topic model [2].

In this study, we conduct Bayesian learning for a wide range of latent variable models based on the exponential family distribution. Our idea is to find the *distribution estimates* for a given model where the hyperparameters of prior distributions are estimated by maximizing the evidence function or the likelihood function which is marginalized over the latent mixture variables as well as the model parameters. A variational Bayesian expectation-maximization (VB-EM) algorithm [6][9] is developed for implementation of Bayesian latent variable models. Compared to the point estimates based on maximum likelihood (ML) or maximum *a posteriori* (MAP) criteria, this method is promising to realize the regularized models for robust speech recognition. The experiments on noisy speech data confirm the benefits of Bayesian approaches to model regularization and speech recognition.

2. BAYESIAN FRAMEWORK

2.1. Bayesian Learning

Figure 1(a) depicts graphical representation for Bayesian learning where the parameter $\boldsymbol{\theta}$ of a single probabilistic model is assumed to be random and is governed by a prior distribution with hyperparameter $\boldsymbol{\eta}$. Given a set of training vectors $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ and a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, there are two Bayesian inference problems: 1) *Model estimation* in which we evaluate the posterior distribution $p(\boldsymbol{\theta}|X, \boldsymbol{\eta}) = p(X|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})$. The *conjugate prior*, which ensures *a posteriori* distribution having the same functional form as the prior, is usually selected. Accordingly, we can rewrite $p(\boldsymbol{\theta}|X, \boldsymbol{\eta})$ as $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}})$ with the updated hyperparameter $\tilde{\boldsymbol{\eta}}$. 2) *Prediction* in which we evaluate the probability of a newly observed

Thanks to F. Soong and Y. Zhang for helpful discussion.

$\mathbf{x}$, namely $p(\mathbf{x}|X,\boldsymbol{\eta}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|X,\boldsymbol{\eta})d\boldsymbol{\theta}$, which is a marginal likelihood over model parameter $\boldsymbol{\theta}$. However, how to determine the hyperparameter $\boldsymbol{\eta}$ is a critical issue. Heuristically selecting $\boldsymbol{\eta}$ from validation data is impractical. We can apply the ML type II estimation and learn the hyperparameter by maximizing the marginal likelihood of training data X given by a general latent variable model

$$\boldsymbol{\eta}_{\text{ML2}} = \arg\max_{\boldsymbol{\eta}} \int_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta}. \tag{1}$$

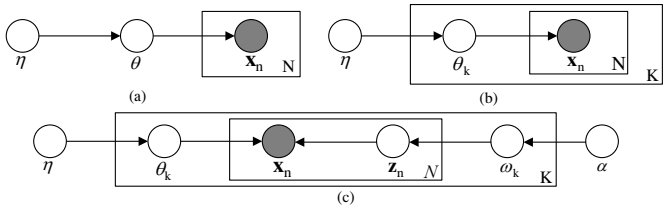


Fig. 1. Graphical model for Bayesian learning of (a)(b) a single model and (c) a latent variable model

2.2. Probabilistic Distribution

The exponential family distribution [1] provides a general representation of probabilistic models, ranging from multinomial distribution to Gaussian distribution, by using

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})\exp\{\boldsymbol{\theta}^T\mathbf{u}(\mathbf{x}) - g(\boldsymbol{\theta})\} \tag{2}$$

where $\mathbf{u}(\mathbf{x})$ is a function of $\mathbf{x}$ and $g(\boldsymbol{\theta})$ is a normalization term. Using ML estimation, the sufficient statistics plays an important role because it summarizes training data X using a compact value in a form of $\boldsymbol{\gamma}[\mathbf{u}(\mathbf{x})]$ where $\boldsymbol{\gamma}[\mathbf{u}(\mathbf{x})] = \sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)$ and $\gamma[1] = N$. The ML estimate $\boldsymbol{\theta}_{\text{ML}}$ satisfies $\nabla g(\boldsymbol{\theta}_{\text{ML}}) = \boldsymbol{\gamma}[\mathbf{u}(\mathbf{x})]/\gamma[1]$. In addition, the conjugate prior for the exponential family distribution is given by

$$p(\boldsymbol{\theta}|\boldsymbol{\chi},v) = \exp\{\boldsymbol{\chi}^T\boldsymbol{\theta} - vg(\boldsymbol{\theta}) - b(\boldsymbol{\chi},v)\} \tag{3}$$

where the hyperparameter is defined by $\boldsymbol{\eta} = [\boldsymbol{\chi}^T, v]^T$. This prior also belongs to the exponential family with the extended parameter vector $\mathbf{s}(\boldsymbol{\theta}) = [\boldsymbol{\theta}^T, -g(\boldsymbol{\theta})]^T$, i.e. $p(\boldsymbol{\theta}|\boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^T\mathbf{s}(\boldsymbol{\theta}) - b(\boldsymbol{\eta})\}$. By combining likelihood function of training data X and conjugate prior, the posterior distribution is calculated as a new exponential family distribution with the updated hyperparameters $\tilde{\boldsymbol{\chi}} = \boldsymbol{\chi} + \boldsymbol{\gamma}[\mathbf{u}(\mathbf{x})]$ and $\tilde{v} = v + \gamma[1]$. Here, v reveals an effective number of pseudo-observations in the prior.

2.3. A General Bayesian Model

As shown in Figure 1(b), we empirically estimate the hyperparameter from a set of K models with parameters $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K\}$ which are calculated from the associated training datasets $\{X_1, \cdots, X_K\}$ where $X_k = \{\mathbf{x}_{nk}\}$. The objective of marginal likelihood in (1) turns out to be $\mathcal{F}(\boldsymbol{\eta}) = \prod_{k=1}^{K}\int_{\boldsymbol{\theta}_k} p(X_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\boldsymbol{\eta})d\boldsymbol{\theta}_k$ where model parameters $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K\}$ are treated as hidden variables. EM algorithm [6] can be applied to find optimal hyperparameter $\hat{\boldsymbol{\eta}}$. In E-step, we evaluate an auxiliary function $\mathcal{Q}(\boldsymbol{\eta}|\boldsymbol{\eta}^{(t)})$ of current hyperparameter $\boldsymbol{\eta}$ given the old hyperparameter $\boldsymbol{\eta}^{(t)}$ at the t^{th} iteration. This function is calculated as an expectation of the logarithm of marginal likelihood with respect to latent variables, which is proportional to $\sum_{k=1}^{K}\int_{\boldsymbol{\theta}_k} p(\boldsymbol{\theta}_k|X_k,\boldsymbol{\eta}^{(t)})\ln p(\boldsymbol{\theta}_k|\boldsymbol{\eta})d\boldsymbol{\theta}_k$. Considering the probabilistic model in (2) and its conjugate prior in (3), the resulting posterior distribution is yielded as a new exponential family distribution $p(\boldsymbol{\theta}|X_k,\boldsymbol{\eta}^{(t)}) \triangleq p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k^{(t)})$ with $\tilde{\boldsymbol{\eta}}_k^{(t)} = [(\tilde{\boldsymbol{\chi}}_k^{(t)})^T, \tilde{v}_k^{(t)}]^T$. As a result, we may treat $\boldsymbol{\theta}$ as data point. Maximizing $\mathcal{Q}(\boldsymbol{\eta}|\boldsymbol{\eta}^{(t)})$ is equivalent to finding ML type II solution $\boldsymbol{\eta}_{\text{ML2}}$ given the training data points drawn from empirical distribution $\sum_{k=1}^{K} p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k^{(t)})$. In M-step, we maximize $\mathcal{Q}(\boldsymbol{\eta}|\boldsymbol{\eta}^{(t)})$ with respect to $\boldsymbol{\eta}$ and find new estimate $\boldsymbol{\eta}^{(t+1)}$ at the $(t+1)^{\text{th}}$ iteration which satisfies

$$\langle s(\boldsymbol{\theta})\rangle_{p(\boldsymbol{\theta}|\boldsymbol{\eta}^{(t+1)})} = \frac{1}{K}\sum_{k=1}^{K}\langle s(\boldsymbol{\theta})\rangle_{p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k^{(t)})} \tag{4}$$

where $\langle\cdot\rangle$ denotes an expectation operation. New hyperparameter $\boldsymbol{\eta}^{(t+1)}$ is estimated by *matching the expectation of* $s(\boldsymbol{\theta})$ *with the ensemble average of* K *expectations given the old posterior parameters* $\tilde{\boldsymbol{\eta}}_k^{(t)}$. Right-hand-side of (4) is seen as the sufficient statistics for finding $\boldsymbol{\eta}_{\text{ML2}}$ from K samples $\{\boldsymbol{\theta}_k\}$. This solution is coincident with the prior estimation through the *matching of statistics*. Notably, we only visit the training data once to collect statistics $\{\boldsymbol{\gamma}_k[\mathbf{u}(\mathbf{x})], \boldsymbol{\gamma}_k[1]\}$. The computational complexity is $O(\sum_{k=1}^{K} N_k)$ which is the same as that of ML training.

2.4. Concavity Analysis

For concavity analysis, we calculate the Hessian matrix of auxiliary function $\mathcal{Q}(\boldsymbol{\eta}|\boldsymbol{\eta}^{(t)})$ with respect to $\boldsymbol{\eta}$ and obtain $\nabla^2\mathcal{Q} = -K^2\text{cov}_{p(\boldsymbol{\theta}|\boldsymbol{\eta})}[s(\boldsymbol{\theta})]$ which is semi-negative definite because the covariance matrix is semi-positive definite. We assure a global optimum of $\mathcal{Q}$ in each iteration. In addition, the Hessian matrix of the original objective $\nabla^2\ln\mathcal{F}(\boldsymbol{\eta})$ can be derived as $\sum_{k=1}^{K}(\text{cov}_{p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k)}[s(\boldsymbol{\theta})] - \text{cov}_{p(\boldsymbol{\theta}|\boldsymbol{\eta})}[s(\boldsymbol{\theta})])$ where $\text{cov}_{p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k)}[s(\boldsymbol{\theta})]$ and $\text{cov}_{p(\boldsymbol{\theta}|\boldsymbol{\eta})}[s(\boldsymbol{\theta})]$ are both semi-positive definite. There is no conclusion drawn on their difference. However, the posterior distribution becomes sharper as more data are observed. In an extreme case, if we have infinite data, $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}})$ converges to $\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})$ where $\boldsymbol{\theta}_{\text{MAP}}$ is the MAP estimate, and $\text{cov}_{p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}})}[s(\boldsymbol{\theta})]$ converges to 0. Thus, we can treat $\ln\mathcal{F}$ as concave in most practical cases.

3. BAYESIAN LATENT VARIABLE MODEL

We extend this full Bayesian framework [1] to establish a mixture model of exponential family distributions. A graphical representation for Bayesian latent variable model is depicted in Figure 1(c). Different from Figure 1(b), the observation $\mathbf{x}$ is not only generated by a set of mixture parameters $\Theta = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K\}$ but also by the latent variable label $\mathbf{z}$ through $p(\mathbf{x}|\mathbf{z}, \Theta) = \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\theta}_k)^{z_k}$ where $p(\mathbf{x}|\boldsymbol{\theta}_k)$ belong to the exponential family and $\mathbf{z}$ comes from a multinomial distribution $p(\mathbf{z}|\boldsymbol{\omega} = \{\omega_k\}) = \text{Mult}(\mathbf{z}; \boldsymbol{\omega})$. For Bayesian learning, we adopt conjugate prior $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ in (3) for exponential family parameter $\boldsymbol{\theta}$ and use the conjugate prior $p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\omega}; \boldsymbol{\alpha})$ for multinomial parameter $\boldsymbol{\omega}$ with hyperparameter $\boldsymbol{\alpha}$. Given a set of training data $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, the joint distribution $p(X, Z, \Theta, \boldsymbol{\omega}|\boldsymbol{\eta}, \boldsymbol{\alpha})$ is calculated by

$$\prod_{n=1}^{N} \text{Mult}(\mathbf{z}_n; \boldsymbol{\omega}) \prod_{k=1}^{K} p(\mathbf{x}_n|\boldsymbol{\theta}_k)^{\mathbf{z}_{nk}} p(\boldsymbol{\theta}_k|\boldsymbol{\eta}) \text{Dir}(\boldsymbol{\omega}; \boldsymbol{\alpha}). \quad (5)$$

ML type II estimates of two hyperparameters $\{\boldsymbol{\eta}_{\text{ML2}}, \boldsymbol{\alpha}_{\text{ML2}}\}$ are calculated by optimizing the likelihood $\sum_Z \int_\Theta \int_{\boldsymbol{\omega}} p(X, Z, \Theta, \boldsymbol{\omega}|\boldsymbol{\eta}, \boldsymbol{\alpha}) d\Theta d\boldsymbol{\omega}$ which is marginalized over three latent variables $\{Z, \Theta, \boldsymbol{\omega}\}$.

3.1. Variational Inference

EM algorithm should be applied to solve this incomplete data problem. However, due to the coupling of three latent variables, the joint posterior distribution $p(Z, \Theta, \boldsymbol{\omega}|X, \boldsymbol{\eta}^{(t)}, \boldsymbol{\alpha}^{(t)})$ is intractable. E-step can not be realized. Therefore, we resort to the approximate inference using variational Bayesian (VB) method [1][2][15][16] where the factorization of posterior distribution $p(Z, \Theta, \boldsymbol{\omega}|X, \boldsymbol{\eta}^{(t)}, \boldsymbol{\alpha}^{(t)}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \hat{q}(z_{nk})$ $\hat{q}(\boldsymbol{\theta}_k)\hat{q}(\omega_k)$ is assumed. Using VB-EM algorithm, the hyperparameters are estimated by maximizing the lower bound of log marginal likelihood or equivalently minimizing the Kullback-Leibler divergence between true posterior $p(Z, \Theta, \boldsymbol{\omega}|X, \boldsymbol{\eta}^{(t)}, \boldsymbol{\alpha}^{(t)})$ and approximate posterior $q(Z)q(\Theta)q(\boldsymbol{\omega})$. In VB E-step, the optimal solution to variational distribution $q(Z)$ is derived as

$$\begin{aligned} \ln \hat{q}(Z) &\propto \langle \ln p(X, Z, \Theta, \boldsymbol{\omega}|\boldsymbol{\eta}, \boldsymbol{\alpha}) \rangle_{p(\Theta, \boldsymbol{\omega})} \\ &\propto \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} \triangleq \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \hat{q}(z_{nk}) \end{aligned} \quad (6)$$

where $\ln \rho_{nk} = \langle \ln \omega_k \rangle_{p(\boldsymbol{\omega}|\boldsymbol{\alpha}^{(t)})} + \langle \ln p(\mathbf{x}_n|\boldsymbol{\theta}_k) \rangle_{p(\boldsymbol{\theta}|\boldsymbol{\eta}^{(t)})}$ and $r_{nk} \triangleq \hat{q}(z_{nk}) = \rho_{nk} / \sum_{k'=1}^{K} \rho_{nk'}$ is a responsibility of the k^{th} latent variable on the n^{th} sample. Similarly, we can derive the other two variational distributions $\hat{q}(\boldsymbol{\theta}_k)$ and $\hat{q}(\omega_k)$. Using this model, the sufficient statistics is refined as $\tilde{\boldsymbol{\gamma}}_k[\mathbf{u}(\mathbf{x})] = \sum_{n=1}^{N} r_{nk} \mathbf{u}(\mathbf{x}_n)$ with an additional weight r_{nk}. Notably, the optimal factorized posteriors are now analogous to the posterior distributions in auxiliary function $Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(t)})$ except that the sufficient statistics is updated as $\boldsymbol{\gamma}[\mathbf{u}(\mathbf{x})] \leftarrow \tilde{\boldsymbol{\gamma}}[\mathbf{u}(\mathbf{x})]$. Next, the VB-M step is implemented by maximizing again the lower bound given $\hat{q}(Z)\hat{q}(\Theta)\hat{q}(\boldsymbol{\omega})$ so as to find hyperparameters $\{\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\alpha}}\}$. This step is similarly performed by (4). As shown in Table 1, Bayesian latent variable model is implemented by a two-stage procedure. First, we calculate r_{nk} by using current distributions $p(\boldsymbol{\theta}|\boldsymbol{\eta}^{(t)})$ and $p(\boldsymbol{\omega}|\boldsymbol{\alpha}^{(t)})$. Next, we find new estimates $\boldsymbol{\eta}^{(t+1)}$ and $\boldsymbol{\alpha}^{(t+1)}$ by using the posterior statistics computed via $\tilde{\boldsymbol{\gamma}}[\mathbf{u}(\mathbf{x})]$.

Table 1. VB inference for Bayesian latent variable model

for each VB-EM iteration	
VB-E step:	
for each $\mathbf{x}_n$, $1 \le n \le N$	
given $\{\boldsymbol{\eta}^{(t)}, \boldsymbol{\alpha}^{(t)}\}$, calculate r_{nk}	
calculate $\{\tilde{\gamma}_k[\mathbf{u}(\mathbf{x})], \tilde{\gamma}_k[1]\}$ w.r.t r_{nk}	
VB-M step:	
given $\{\tilde{\gamma}_k[\mathbf{u}(\mathbf{x})], \tilde{\gamma}_k[1]\}$, solve $\boldsymbol{\eta}^{(t+1)}$, set $\boldsymbol{\eta}^{(t)} \leftarrow \boldsymbol{\eta}^{(t+1)}$	
given $\tilde{\gamma}_k[1]$, solve $\boldsymbol{\alpha}^{(t+1)}$, set $\boldsymbol{\alpha}^{(t)} \leftarrow \boldsymbol{\alpha}^{(t+1)}$	

3.2. Case Studies

We have presented a Bayesian latent variable model based on the exponential family which could be realized into many Bayesian models including latent Dirichlet allocation (mixture of multinomial distributions) [2], mixture of Gaussian distributions (MoG) and hidden Markov model (HMM) [16].

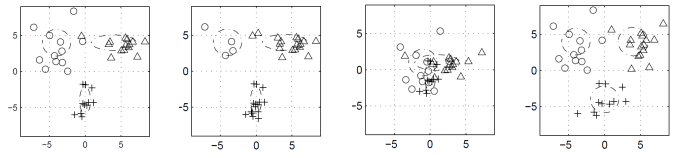
In MoG model, the likelihood function of D-dimensional observation is given by $p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \Lambda_k^{-1})$ with mixture weight ω_k, mean vector $\boldsymbol{\mu}_k$ and precision matrix Λ_k. The conjugate prior for multinomial parameter ω_k is known as a Dirichlet distribution and that for Gaussian parameters is given by a Gaussian-Wishart distribution $p(\boldsymbol{\mu}, \Lambda|\boldsymbol{\nu}, W, \beta, \tau) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\nu}, (\beta\Lambda)^{-1}) \mathcal{W}(\Lambda; W, \tau)$ where $\mathcal{W}$ is a Wishart distribution with a $D \times D$ symmetric, positive definite matrix W and a degree of freedom τ. The responsibility r_{nk} in VB E-step is determined based on log variational parameter $\ln \rho_{nk}$ which is calculated by adding the expected log probabilities $\langle \ln \omega_k \rangle_{p(\boldsymbol{\omega}|\boldsymbol{\alpha}^{(t)})}$ and $\langle \ln p(\mathbf{x}_n|\boldsymbol{\mu}_k, \Lambda_k) \rangle_{p(\boldsymbol{\mu}, \Lambda|\boldsymbol{\nu}^{(t)}, W^{(t)}, \beta^{(t)}, \tau^{(t)})}$ to obtain [1]

$$\begin{aligned} &\psi(\alpha_k^{(t)}) - \psi(\sum_{k'=1}^{K} \alpha_{k'}^{(t)}) + \frac{1}{2}[\sum_{i=1}^{D} \psi(\frac{\tau_k^{(t)} + 1 + i}{2}) + \ln |W_k^{(t)}| \\ &- D(\ln \pi + (\beta_k^{(t)})^{-1}) - \tau_k^{(t)} (\mathbf{x}_n - \boldsymbol{\nu}_k^{(t)})^T W_k^{(t)} (\mathbf{x}_n - \boldsymbol{\nu}_k)^{(t)}] \end{aligned} \quad (7)$$

where $\psi(\alpha) \triangleq \frac{d}{d\alpha} \ln \Gamma(\alpha)$ denotes the digamma function. As β_k and τ_k increases in (7), ρ_{nk} gets closer to a sharp Gaussian. In the other extreme case, if $\tau_k \to 0^+$ or $\beta_k \to 0^+$, which implies that prior density $p(\boldsymbol{\mu}, \Lambda|\boldsymbol{\nu}, W, \beta, \tau)$ is extremely uncertain, ρ_{nk} will yield the same value for all samples $\mathbf{x}_n$.

Considering the HMMs with output distribution as a latent variable model of the exponential family distributions, we can apply VB-EM algorithm in Table 1 and construct

the regularized HMMs where the optimal hyperparameters are estimated from training data X. Given a L-state HMM with initial state probabilities $\boldsymbol{\pi} = \{\pi_i\}_{1\leq i\leq L}$, state transition probabilities $A = \{a_{ij}\}_{L\times L}$ and output distributions $B = \{\sum_{k=1}^{K}\omega_{ik}p(\mathbf{x}|\boldsymbol{\theta}_{ik})\}_{1\leq i\leq L}$, the full Bayesian framework [1] for HMMs can be implemented. In VB E-step, we conduct the standard Baum-Welch algorithm and calculate ρ_{nik} for state i by (7) and determine the variational state occupation probability r_{nik}. Meanwhile, we collect the statistics $\{\tilde{\boldsymbol{\gamma}}_{ik}[\mathbf{u}(\mathbf{x})], \tilde{\gamma}_{ik}(1)\}$ for all mixture probabilities $p(\mathbf{x}|\boldsymbol{\theta}_{ik})$. In VB M-step, the optimal hyperparameters are estimated by using these statistics. A special case of HMMs with MoG output distributions was addressed in [16][17].



(a)$\hat{\beta} = 0.115, \hat{\tau} = 1.48$ (b)$\hat{\beta} = 0.0874, \hat{\tau} = 1.35$ (c)$\hat{\beta} = 0.354, \hat{\tau} = 3.62$ (d)$\hat{\beta} = 0.166, \hat{\tau} = 13.44$

Fig. 2. Effect of hyperparameters for Gaussian distribution.

Table 2. Word accuracy (%) versus no of training utterances

	no of utterances (CT)				no of utterances (MT)			
	100	500	2000	8440	100	500	2000	8440
ML	23.3	46.9	48.8	63.9	56.7	70.0	75.6	88.8
Bayesian I	38.2	53.0	53.5	66.1	64.2	72.1	76.8	89.2
Bayesian II	41.8	51.7	53.9	66.2	64.3	71.6	76.6	89.1
Bayesian III	45.9	55.4	54.7	66.4	64.1	68.9	74.2	88.8
Full Bayesian	52.1	56.8	55.5	67.1	65.9	72.6	77.2	89.2

4. EXPERIMENTS

4.1. Effect of Hyperparameters

We first illustrate the Bayesian framework for Gaussian distribution. In this experiment, three 2-dimensional, diagonal covariance Gaussians share an identical Gaussian-Wishart prior. The illustration for four training conditions are shown in Figure 2(a)-(d), in which the training samples belong to different Gaussians are marked by their point types. The optimal hyperparameters $\hat{\beta}$ and $\hat{\tau}$ are displayed. We take (a) as a baseline condition, which yields the optimal $\hat{\boldsymbol{\nu}} = (-0.217, 1.64)$ and $\hat{W} = \text{diag}(0.593, 0.523)$. We investigate three other conditions. Condition (b) is set up by taking away some samples of class '$\circ$' from condition (a). As a result, the optimal $\hat{\boldsymbol{\nu}} = (0.219, 1.65)$, $\hat{W} = \text{diag}(0.801, 0.524)$ are estimated so as to reflect the Gaussians with more training samples. In condition (c), the three Gaussians are moved closer to each other, which leads to a larger hyperparameter $\hat{\beta} = 0.354$. This is reasonable because we are more confident to locate the position of the Gaussian. In condition (d), three Gaussians are distorted to have more similar covariances. The estimated hyperparameter is increased as $\hat{\tau} = 13.44$ which results in a more confident estimation of covariance matrix. In summary, Bayesian framework could flexibly reflect the model uncertainty and estimate an appropriate prior density in various conditions. This framework is helpful for Bayesian learning with evolved hyperparameters [3].

4.2. Noisy Speech Recognition

We further evaluate the proposed algorithm by using Aurora2, a connected-digit noisy speech recognition task. The HMMs were constructed for each of eleven digits ranging from 'zero' to 'nine', and 'oh'. The 3-component MoGs with diagonal covariance matrices were adopted as the output distributions for all HMM states. Each sample $\mathbf{x}_n$ consisted of 39-dimensional features based on MFCCs and their dynamic features. We trained the hyperparameters of HMMs, and then in test phase, the *uncertainty decoding* algorithm using marginal likelihood as shown in [3][4][5][8][16] was performed. The conditions of clean training (CT) and multi-conditional training (MT) were examined. There were 8440 utterances collected in different noise conditions. We set up the training conditions with 100, 500, 2000 and 8440 utterances. We compare the averaged word accuracies (%) over three systems: (a) ML training, (b) Bayesian training with heuristic hyperparameter [16], in which $\boldsymbol{\nu}$ and W are obtained by statistics matching while β and τ are selected between 0.001 and 10 [16], and (c) full Bayesian training with learning of hyperparameters. In case of Bayesian I, the best hyperparameters $\beta = \tau = 0.1$ for MT condition are selected. In Bayesian III, the best hyperparameters $\beta = \tau = 2.0$ for CT condition are selected. In Bayesian II, the intermediate hyperparameters $\beta = \tau = 0.5$ are specified. The results with CT and MT training modes are reported in Table 2. Note that the best heuristically-selected hyperparameters β and τ differ significantly for CT (2.0) and MT (0.1). Improvement is obvious in case of small amount of training data. However, an inappropriate hyperparameter may lead to even worse performance than ML training. Hyperparameters should be data-driven and adaptive for different conditions. The proposed full Bayesian framework is beneficial to achieve this goal. Our performance is better than conventional Bayesian learning via hyperparameter selection.

5. CONCLUSIONS

We concerned the model regularization and presented the general solution to Bayesian latent variable model where the exponential family distribution was adopted. We derived a VB-EM algorithm to estimate the hyperparameters where the computation cost was similar to that of ML training. Proposed algorithm has the potential for modeling the uncertainty in different latent variable models and has improved the robustness for noisy speech recognition.

6. REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science, 2006.

[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993-1022, 2003.

[3] J.-T. Chien, "A Bayesian prediction approach to robust speech recognition and online environmental learning", *Speech Communication*, vol. 37, nos. 3-4, pp. 321-334, 2002.

[4] J.-T. Chien, "Linear regression based Bayesian predictive classification for speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 1, pp. 70-79, 2003.

[5] J.-T. Chien and S. Furui, "Predictive hidden Markov model selection for speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 377-387, 2005.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of Royal Statistics Society (B)*, vol. 39, no. 1, pp. 1-38, 1977.

[7] J. Droppo, A. Acero and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition", in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 57-60, 2002.

[8] Q. Huo and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 200-204, 2000.

[9] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L. Saul, "An introduction to variational methods for graphical models", *Machine Learning*, vol. 37, pp. 183-233, 1999.

[10] J. T.-Y. Kwok, "The evidence framework applied to support vector machines", *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp. 1162-1173, 2000.

[11] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition", *Speech Communication*, vol. 50, pp. 265-277, 2008.

[12] D. J. C. MacKay, "The evidence framework applied to classification networks", *Neural Computation*, vol. 4, no. 5, pp. 720-736, 1992.

[13] G. Saon and J.-T. Chien, "Bayesian sensing hidden Markov models", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 43-54, 2012.

[14] G. Saon and J.-T. Chien, "Large-vocabulary continuous speech recognition systems - a look at some recent advances," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18-33, 2012.

[15] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference", *Foundations and Trends in Machine Learning*, vol. 1, nos. 1-2, pp. 1-305, 2008.

[16] S. Watanabe, Y. Minami, A. Nakamura and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 365-381, 2004.

[17] Y. Zhang, P. Liu, J.-T. Chien and F. Soong, "An evidence framework for Bayesian learning of continuous-density hidden Markov models", in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, pp. 3857-3860, 2009.