A General Evidence Framework for Regularized Probabilistic Models

Anonymous Author(s)
Affiliation
Address
email

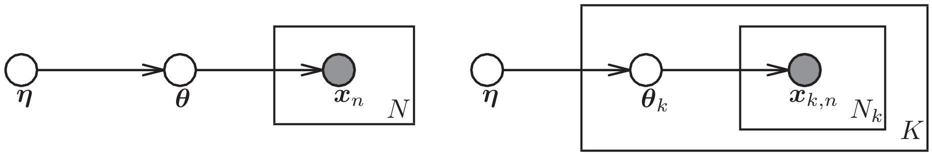
Abstract

To jointly learn both prior and posterior distribution from ill-posed data, we present a concrete solution to empirical Bayesian learning for regularized exponential family distributions and their mixture model. Under the evidence framework, we first develop an expectation-maximization (EM) algorithm to find optimal prior by maximizing the evidence function in presence of latent model parameters. Consequently, we extend the solution to deal with mixture of exponential family distributions by making use of the variational method, and come up with a double-looped EM procedure. In addition, we analyze the physical meaning and the computational behavior of the proposed method. Finally, we conceptually illustrated on a Gaussian case, and experimentally verified by a noisy speech recognition task using hidden Markov models. Results show that the proposed algorithm is effective in automatically learning the decent hyperparameters, and is adaptive to the variations in the training data.

1 Introduction

We address the problem of empirical Bayesian learning [1, 2], namely, to determine the prior distribution from data, for a wide range of statistical models including exponential family distributions and their mixture models. Our goal is to present the theoretical solutions to Bayesian learning in the absence of strict prior knowledge to guide the estimation of hyperparameters. In comparison with the maximum likelihood (ML) estimation, the Bayesian treatment has been shown promising in more reliably estimating statistical models with insufficient or noisy training data. This is achieved by regarding model parameters as uncertain, and introducing a prior distribution to regularize them. As we know, for real world problems, we can never find concise, analytically tractable models to perfectly fit our observation, especially when the observations are ill-posed. A flexible model, if can be reliably estimated, is usually endowed with better capability to describe a stochastic phenomenon. Hence, Bayesian learning, either in its precise or conceptually similar embodiments, has been shown successful in many real problems, ranging from the smoothed n-gram language model [3], robust speech recognition [4], to text corpora modeling [5].

An important issue in Bayesian learning is how to reasonably find prior distributions. This issue has twofold considerations: First, we prefer to give an appropriate analytic form to the prior, which is required more by mathematical convenience than by Bayesian learning itself. Second, given a family of candidate prior distributions, we should reasonably select one of them by specifying its hyperparameters. This can be done either subjectively or objectively [1]: In subjective Bayes, the prior distribution is selected based on some background knowledge; while in the latter approach, also referred to as *empirical Bayes* [1, 2], the prior distribution is automatically learned from the training data. In machine learning community, the paradigm is also known as the *evidence framework* [6, 7], which reveals that the criteria to determine the optimal prior is the evidence function, e.g., marginalized likelihood by integrating over the model parameters. The evidence framework



(a) without prior sharing

(b) with prior sharing

Figure 1: Graphical representation for Bayesian learning

has been shown successful for some specific probabilistic models, e.g., linear regression model [6], support vector regression model [8], neural network [7], and semantic analysis models [5].

In this study, we strive towards a general evidence framework to build regularized probabilistic models for exponential family distributions and their mixture models. The reason of why we focus on the exponential family is not only its intensive use in machine learning, but also insightful conclusions we can summarize from our investigations. By treating the model parameters as hidden variables, we come up with a general expectation-maximization (EM) [9] solution to the evidence framework for exponential family distributions. By further adopting the variational method [10, 11], we extend the solution to deal with the mixture of exponential family distributions, with a double-looped EM algorithm. By using the proposed method, the prior distribution, as well as the corresponding posterior distributions, can be simultaneously learned from data. Experiments show that the evidence framework solution is powerful in assign appropriate priors for a complex model set, and it leads to considerable improvement in classification accuracy, and significantly better convenience on a noisy speech recognition task, compared with the non-empirical Bayesian methods.

2 Backgrounds

2.1 Bayesian learning

The concept of Bayesian learning is depicted by the graphical model in Fig.1(a): Given a probabilistic model, we assume its parameter $\boldsymbol{\theta}$ as a random variable, which is generated by a *prior distribution* governed by the hyperparameter $\boldsymbol{\eta}$. Given a set of training samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ and a *prior distribution* $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, there are two basic Bayesian inference problems:
1. *Model estimation*, in which we evaluate the probability of a model parameter:

$$p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{\eta}) = p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta}) \tag{1}$$

The distribution $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{\eta})$ is named the *posterior distribution*. Accordingly, the concept of *conjugate prior*, which ensures a posterior having the same functional form as the prior, is important because it can dramatically simplify the mathematical derivation. In the context of conjugate prior, we can rewrite $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{\eta})$ as $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}})$, where $\tilde{\boldsymbol{\eta}}$ denotes the posterior parameters.
2. *Prediction*, in which we evaluate the probability of a newly observed $\boldsymbol{x}$:

$$p(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\eta}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{\eta})\mathrm{d}\boldsymbol{\theta} \tag{2}$$

In these two problems, how to determine $\boldsymbol{\eta}$ dramatically affects the behavior, but manual tuning for decent $\boldsymbol{\eta}$ on a large set of statistical models is not feasible. Hence, empirical learning of hyperparameters, which can be addressed by the evidence framework [6], has been explored and shown promising in many real-world applications [6, 7, 8, 5]. The evidence framework is given by:

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \int_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})\mathrm{d}\boldsymbol{\theta} \tag{3}$$

2.2 Exponential family distributions

The exponential family [12, 13] contains various statistical models, ranging from Gaussian distribution to maximum entropy model [14]. These distributions can be further used as assembling components for more elaborate models. The canonical form of the exponential family is:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x})\exp\{\boldsymbol{\theta}^\top \mathbf{u}(\boldsymbol{x}) - g(\boldsymbol{\theta})\} \tag{4}$$

where $\mathbf{u}(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ and $g(\boldsymbol{\theta})$ is a normalization term. In its ML estimation, the *sufficient statistics* [12] plays a significant role because it summarizes the data by a compact number of quantities. For notation convenience, we define a general statistics w.r.t. an arbitrary function $\boldsymbol{f}$:

$$\boldsymbol{\gamma}[\boldsymbol{f}(\boldsymbol{x})] \quad = \textstyle\sum_{n=1}^{N} \boldsymbol{f}(\boldsymbol{x}_n) \tag{5}$$

Accordingly, the sufficient statistics takes form $\boldsymbol{\gamma}[\mathbf{u}(\boldsymbol{x})]$, and the ML estimation is given by:

$$\nabla g(\boldsymbol{\theta}) = \boldsymbol{\gamma}[\mathbf{u}(\boldsymbol{x})]/\gamma(1) \tag{6}$$

For an exponential family distribution, there exists a generic conjugate prior distribution:

$$p(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) = \exp\left\{\boldsymbol{\chi}^{\top}\boldsymbol{\theta} - \nu g(\boldsymbol{\theta}) - b(\boldsymbol{\chi}, \nu)\right\} \tag{7}$$

where $\boldsymbol{\eta}^{\top} = (\boldsymbol{\chi}^{\top}, \nu)$ is the hyperparameter. Note that this conjugate prior also belongs to the exponential family, which can be more clearly shown if we denote $\mathbf{s}^{\top}(\boldsymbol{\theta}) = [\boldsymbol{\theta}^{\top}, -g(\boldsymbol{\theta})]$. Given sufficient statistics, the posterior parameters obtained in Eq.(1) are given by:

$$\tilde{\boldsymbol{\chi}} = \boldsymbol{\chi} + \boldsymbol{\gamma}[\mathbf{u}(\boldsymbol{x})], \quad \tilde{\nu} = \nu + \gamma(1) \tag{8}$$

It reveals that ν can be viewed as a virtual number of observations, or confidence of the model.

3 Regulzarized Exponential Family Distributions

3.1 A generic EM solution to evidence framework

As we shall discuss in section 3.3, without prior sharing across models, we can only obtain a trivial solution to the evidence framework. Hence, we instead study the case with prior sharing, as shown in Fig.1(b). In this case, K statistic models, governed by $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K$, share an identical prior governed by $\boldsymbol{\eta}$, and $\boldsymbol{X}_k = \{\boldsymbol{x}_{k,1}, \cdots, \boldsymbol{x}_{k,N_k}\}$ represents the data set associated with the k^{th} model. Based upon Eq.(3), now the evidence framework problem is generally written:

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \mathcal{F}(\boldsymbol{\eta}) = \arg\max_{\boldsymbol{\eta}} \prod_{k=1}^{K} \int_{\boldsymbol{\theta}_k} p(\boldsymbol{X}_k|\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta}_k \tag{9}$$

Obviously, Eq.(3) is a special case of it. Eq.(9) can be viewed as a ML estimation problem with respect to $\boldsymbol{\eta}$, by regarding all the model parameters $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K$ as hidden variables, and it is then natural to apply the EM algorithm [9] for parameter estimation. In the E-step, given the old hyperparameter $\boldsymbol{\eta}^{\text{old}}$, we evaluate the following auxiliary function:

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\eta}, \boldsymbol{\eta}^{\text{old}}) &= \sum_{k=1}^{K} \int_{\boldsymbol{\theta}_k} p(\boldsymbol{\theta}_k|\boldsymbol{X}_k, \boldsymbol{\eta}^{\text{old}}) \ln p(\boldsymbol{X}_k, \boldsymbol{\theta}_k|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta}_k \\ &= \sum_{k=1}^{K} \int_{\boldsymbol{\theta}_k} p(\boldsymbol{\theta}_k|\boldsymbol{X}_k, \boldsymbol{\eta}^{\text{old}}) \ln p(\boldsymbol{\theta}_k|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta}_k + \text{const} \end{aligned} \tag{10}$$

We focus on the cases in which all $p(\boldsymbol{x}|\boldsymbol{\theta}_k)$ belong to the exponential family. By adopting conjugate prior, we can rewrite $p(\boldsymbol{\theta}|\boldsymbol{X}_k, \boldsymbol{\eta}^{\text{old}})$ as $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k^{\text{old}})$, where the posterior parameters $(\tilde{\boldsymbol{\eta}}_k^{\text{old}})^{\top} = [(\tilde{\boldsymbol{\chi}}_k^{\text{old}})^{\top}, \tilde{\nu}_k^{\text{old}}]$ can be calculated by Eq.(8). Accordingly, we can rewrite Eq.(10) as:

$$\mathcal{Q}(\boldsymbol{\eta}, \boldsymbol{\eta}^{\text{old}}) = \sum_{k=1}^{K} \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k^{\text{old}}) \ln p(\boldsymbol{\theta}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta} + \text{const} \tag{11}$$

If we treat $\boldsymbol{\theta}$ as data, maximizing $\mathcal{Q}$ in the M-step is essentially finding the ML solution of $\boldsymbol{\eta}$, given an empirical distribution proportional to $\sum_{k=1}^{K} p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k^{\text{old}})$. As we mentioned in section 2.2, $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ also belongs to the exponential family, so the ML estimation can be obtaind by solving an equation corresponding to Eq.(6). We come up with the following M-step solution:

$$\langle \mathbf{s}(\boldsymbol{\theta}) \rangle_{p(\boldsymbol{\theta}|\boldsymbol{\eta}^{\text{new}})} \quad = \tfrac{1}{K} \textstyle\sum_{k=1}^{K} \langle \mathbf{s}(\boldsymbol{\theta}) \rangle_{p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k^{\text{old}})} \tag{12}$$

where $\langle \cdot \rangle$ denotes an expectation operation. The physical meaning of Eq.(12) can be clarified as: we should find the new hyperparameter $\boldsymbol{\eta}^{\text{new}}$ by matching the expectation of $\mathbf{s}(\boldsymbol{\theta})$ with the ensemble average of K expectations given their corresponding old posterior parameters $\tilde{\boldsymbol{\eta}}_k^{\text{old}}$.

The EM process is shown in Table 1. Meanwhile, we only need visit the training data once to collect statistics, so the computational cost is $O(\sum_{1}^{K} N_k)$, i.e, the same as the ML training. It is notable that although we aim at find optimal prior in this procedure, the posterior parameters are actually by-products of it. Namely, the algorithm can learn prior and posteriors from data simultaneously.

Table 1: An EM solution to the evidence framework for the exponential family (**Algorithm I**)

```
Collect sufficient statistics
    for each $k$
        given $\boldsymbol{X}_k$, accumulate $\boldsymbol{\gamma}_k[\mathbf{u}(\boldsymbol{x})], \gamma_k(1) = N_k$
Set $\boldsymbol{\eta}^{\text{old}}$ to its seed value
Do
    E-step: given $\boldsymbol{\eta}^{\text{old}}, \gamma_k(1), \boldsymbol{\gamma}_k[\mathbf{u}(\boldsymbol{x})]$, calculate $\tilde{\boldsymbol{\eta}}_k^{\text{old}}$ by Eq.(8)
    M-step: given $\tilde{\boldsymbol{\eta}}_k^{\text{old}}$, solve Eq.(12) for $\boldsymbol{\eta}^{\text{new}}$, $\boldsymbol{\eta}^{\text{old}} \leftarrow \boldsymbol{\eta}^{\text{new}}$
Until $\|\boldsymbol{\eta}^{\text{old}} - \boldsymbol{\eta}^{\text{new}}\|$ is small enough
```

3.2 Concavity analysis

Now let us check what kind of optimum we achieve in Algorithm I. It can be derived that the Hessian matrix of $\mathcal{Q}(\boldsymbol{\eta}, \boldsymbol{\eta}^{\text{old}})$, with respect to $\boldsymbol{\eta}$, takes the following form:

$$\nabla^2 \mathcal{Q} \quad = \quad -K^2 \text{cov}_{p(\boldsymbol{\theta}|\boldsymbol{\eta})}\left[\mathbf{s}(\boldsymbol{\theta})\right] \tag{13}$$

Obviously, $\nabla^2 \mathcal{Q}$ is semi-negative definite because a covariance matrix is guaranteed to be semi-positive definite [15]. Hence, we can assure a global optimum of $\mathcal{Q}$ in each iteration.

The Hessian matrix of the original evidence function can be derived as:

$$\nabla^2 \ln \mathcal{F}(\boldsymbol{\eta}) \quad = \sum_{k=1}^{K} \left\{ \text{cov}_{p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k)}[\mathbf{s}(\boldsymbol{\theta})] - \text{cov}_{p(\boldsymbol{\theta}|\boldsymbol{\eta})}[\mathbf{s}(\boldsymbol{\theta})] \right\} \tag{14}$$

Although both $\text{cov}_{p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}}_k)}[\mathbf{s}(\boldsymbol{\theta})]$ and $\text{cov}_{p(\boldsymbol{\theta}|\boldsymbol{\eta})}[\mathbf{s}(\boldsymbol{\theta})]$ are semi-positive definite, there is no general conclusion can be drawn on their difference, which implies in general we may not achieve the global evidence maximum by the EM algorithm. However, let us check this problem from the engineer perspective: In principle, the posterior becomes sharper along with size of observed data increasing. In the extreme case, if we have infinite data, $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ converges to $\delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{MAP}})$, where $\boldsymbol{\theta}^{\text{MAP}}$) denotes the *maximum a posteriori (MAP)* value [12], and $\text{cov}_{p(\boldsymbol{\theta}|\tilde{\boldsymbol{\eta}})}[\mathbf{s}(\boldsymbol{\theta})]$ converges to $\mathbf{0}$. Hence, we can approximately treat $\ln \mathcal{F}$ as concave in quite a few practical cases.

3.3 Importance of prior sharing

Although the evidence framework provides us a decent approach to conduct Bayesian learning without explicitly specifying the hyperparameters, a reasonable prior sharing structure is intrinsically important to achieve a proper solution. Considering the case without prior sharing, namely, $K = 1$ in Eq.(12), we can obtain the solution of M-step by inspection: $\boldsymbol{\eta}^{\text{new}} = \boldsymbol{\eta}^{\text{old}}$, which implies the hyperparameter will keep unchanged in the M-step. By validating it with E-step of Eq.(8), where ν is always increasing, we can easily obtain: $\hat{\nu} = \infty$. It implies that without prior sharing, we can only achieve a trivial prior model with infinite confidence. Obviously, this case takes no advantage of Bayesian learning and is practically useless. This analysis reveals an inspiring insight to the physical meaning of the evidence framework: Beyond the mathematical attractiveness, what is even more important is to design a reasonable prior evolution space based upon our unique domain knowledge of the problem.

3.4 Case studies

3.4.1 Multinomial distribution

Multinomial distribution is widely used in characterizing discrete random variables. Multinomial variables can be used to describe the values being one of M possible mutually exclusive values. By applying the 1-of-M scheme, we can describe the multinomial distribution over $\boldsymbol{x}$ by:

$$p(\boldsymbol{x}|\boldsymbol{\mu}) \quad = \text{Mult}(\boldsymbol{x}; \boldsymbol{\mu}) \triangleq \prod_{m=1}^{M} \mu_m^{x_m} \tag{15}$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_m)^\top$, and the parameters μ_m are constrained to satisfy $\mu_m \geq 0$ and $\sum_m \mu_m = 1$. Apparently, multinomial distribution belongs to the exponential family. According to Eq.(7), its conjugate prior can be given by:

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \quad = \text{Dir}(\boldsymbol{\mu}; \boldsymbol{\alpha}) = \Gamma(\alpha_0) \left[\Gamma(\alpha_1) \cdots \Gamma(\alpha_M)\right]^{-1} \prod_{m=1}^{M} \mu_m^{\alpha_m - 1} \tag{16}$$

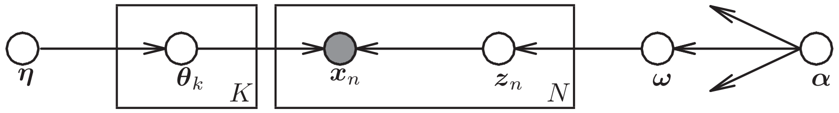


Figure 2: Graphical representation for Bayesian learning for probabilistic mixture model

which is called the Dirichlet distribution. Here $\boldsymbol{\alpha}$ denotes $(\alpha_1, \cdots, \alpha_M)^\top$, and $\alpha_0 = \sum_{m=1}^M \alpha_m$. By applying Eq.(8) and Eq.(12), we obtain the concrete EM solution for multinomial distributions:

1. E-step: $\quad \tilde{\boldsymbol{\alpha}}_k^{\text{old}} = \boldsymbol{\alpha}^{\text{old}} + \boldsymbol{\gamma}_k(\boldsymbol{x})$
2. M-step: $\quad \psi(\alpha_m^{\text{new}}) - \psi\left(\alpha_0^{\text{new}}\right) = \frac{1}{K}\sum_{k=1}^K \left[\psi(\alpha_{km}^{\text{old}}) - \psi\left(\alpha_{k0}^{\text{old}}\right)\right] \quad (1 \leq m \leq M)$

where $\psi(\alpha) \equiv \partial \ln \Gamma(\alpha)/\partial\alpha$ denotes the digamma function. The M-step can be solved by the Newton method.

3.4.2 Gaussian distribution

A D-dimensional Gaussian distribution takes the following form:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) = |\boldsymbol{\Lambda}|^{\frac{1}{2}}(2\pi)^{-\frac{D}{2}} \exp\left\{-(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})/2\right\} \tag{17}$$

A most frequently used conjugate prior of Gaussian parameters is the Gaussian-Wishart distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{v}, \boldsymbol{W}, \beta, \tau) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{\Lambda}^{-1}/\beta)\mathcal{W}(\boldsymbol{\Lambda}; \boldsymbol{W}, \tau) \tag{18}$$

where $\mathcal{W}(\boldsymbol{\Lambda}; \boldsymbol{W}, \nu) = B(\boldsymbol{W}, \tau)|\boldsymbol{\Lambda}|^{\frac{\tau-D+1}{2}} \exp\left\{-\text{tr}(\boldsymbol{W}^{-1}\boldsymbol{\Lambda})/2\right\}$ is called Wishart distribution. Note that the Gaussian-Wishart distribution can not be represented by the generic conjugate prior form of Eq.(7) because now there are two uncertainty variables β and τ. Conceptually, it is designed as follows: 1. Treat $\boldsymbol{\mu}$ as a free parameter, which leads to a Gaussian conjugate prior $\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{\Lambda}^{-1}/\beta)$ by applying Eq.(7); 2. Treat $\boldsymbol{\Lambda}$ as another free parameter, which leads to a Wishart conjugate prior $\mathcal{W}(\boldsymbol{\Lambda}; \boldsymbol{W}, \tau)$ by applying Eq.(7). It is easy to verify that such a decomposable prior distribution is also conjugate to the original distribution, and the solution given in Algorithm I is still valid. By substituting Eq.(17) and Eq.(18) into it, the EM steps are derived by:

1. E-step: $\quad \tilde{\beta}_k^{\text{old}} = \beta^{\text{old}} + \gamma_k(1), \quad \tilde{\tau}_k^{\text{old}} = \tau^{\text{old}} + \gamma_k(1)$

$$\begin{aligned}
\tilde{\boldsymbol{v}}_k^{\text{old}} &= [\beta^{\text{old}}\boldsymbol{v}^{\text{old}} + \boldsymbol{\gamma}_k(\boldsymbol{x})]/[\beta^{\text{old}} + \gamma_k(1)] \\
(\tilde{\boldsymbol{W}}_k^{\text{old}})^{-1} &= (\boldsymbol{W}^{\text{old}})^{-1} + \boldsymbol{\gamma}_k(\boldsymbol{x}\boldsymbol{x}^\top) - \boldsymbol{\gamma}_k(\boldsymbol{x})\boldsymbol{\gamma}_k^\top(\boldsymbol{x}) + \\
&\quad \gamma_k(1)\beta^{\text{old}}(\tilde{\beta}_k^{\text{old}})^{-1}\left[\boldsymbol{\gamma}_k(\boldsymbol{x})/\gamma_k(1) - \boldsymbol{v}^{\text{old}}\right]\left[\boldsymbol{\gamma}_k(\boldsymbol{x})/\gamma_k(1) - \boldsymbol{v}^{\text{old}}\right]^\top
\end{aligned} \tag{19}$$

2. M-step: $\quad \boldsymbol{W}^{\text{new}} = \frac{1}{K}\sum_{k=1}^K \tilde{\boldsymbol{W}}_k^{\text{old}}, \quad \boldsymbol{v}^{\text{new}} = \frac{1}{K}(\boldsymbol{W}^{\text{new}})^{-1}\sum_{k=1}^K \tilde{\boldsymbol{W}}_k^{\text{old}}\tilde{\boldsymbol{v}}_k^{\text{old}}$

$$\begin{aligned}
(\beta^{\text{new}})^{-1} &= \frac{1}{K}\sum_{k=1}^K \left\{(\tilde{\beta}_k^{\text{old}})^{-1} + (\boldsymbol{v}^{\text{new}} - \tilde{\boldsymbol{v}}_k^{\text{old}})^\top \tilde{\boldsymbol{W}}_k^{\text{new}}(\boldsymbol{v}^{\text{new}} - \tilde{\boldsymbol{v}}_k^{\text{old}})\right\} \\
\phi(\tau^{\text{new}}) &= \frac{1}{K}\sum_{k=1}^K \left\{\phi(\tilde{\tau}_k^{\text{old}}) + \ln|\tilde{\boldsymbol{W}}_k^{\text{old}}| - \ln|\boldsymbol{W}^{\text{new}}|\right\}
\end{aligned} \tag{20}$$

where $\phi(\tau) \triangleq \psi(\tau/2) - \ln(\tau/2)$. The solution of $\boldsymbol{v}_{\text{new}}$ and $\boldsymbol{W}_{\text{new}}$ is intuitive, and are coincident to other prior estimation methods inspired by statistcs matching. In addition, an interesting property of the solution can be found if we sum up the equations related with β and τ in Eq.(20):

$$\phi(\tau^{\text{new}}) + \frac{1}{\beta^{\text{new}}} = \frac{1}{K}\left\{\sum_{k=1}^K \phi(\tilde{\tau}_k^{\text{old}}) + \frac{1}{\tilde{\beta}_k^{\text{old}}} + \text{KL}\left[\mathcal{N}(;\tilde{\boldsymbol{v}}_k^{\text{old}}, (\tilde{\boldsymbol{W}}_k^{\text{old}})^{-1})||\mathcal{N}(;\boldsymbol{v}^{\text{new}}, (\boldsymbol{W}^{\text{new}})^{-1})\right]\right\} \tag{21}$$

The result clearly shows that the two uncertainty variables β and τ jointlys characterizes the statistical difference, or KL divergence [16], between the MAP values of the newly estimated prior and all the seed posteriors, which reveals the physical meaning of the evidence framework.

4 Regularized Mixture of Exponential Family Distributions

Although the exponential family covers a widely range of statistical models, we often need further flexibility to describe more complicated stochastic observations. A well established scheme is to use mixture density models with relatively simple component distributions. In this section, we apply the evidence framework to the mixture model of the exponential family distributions, as depicted in Fig.1(b). By comparing Fig.1(b) with Fig.2, we find now the observation $\boldsymbol{x}$ depends on both a set of parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K\}$, and a component selecting variable $\boldsymbol{z}$, through a conditional distribution takes the form:

$$p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\Theta}) \quad = \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\theta}_k)^{z_k} \tag{22}$$

In this study, we assume all $p(\boldsymbol{x}|\boldsymbol{\theta}_k)$ belong to the exponential family with the same functional form, and $\boldsymbol{z}$ can be characterized by a Multinomial distribution $p(\boldsymbol{z}|\boldsymbol{\omega}) = \text{Mult}(\boldsymbol{z};\boldsymbol{\omega})$.

In Bayesian learning, we consider that the prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ is a conjugate prior, and the prior distribution of $\boldsymbol{\omega}$ is a Dirichlet distribution, namely, $p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\omega};\boldsymbol{\alpha})$. Given the training data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, now the joint distribution is given by:

$$p(\boldsymbol{X},\boldsymbol{Z},\Theta,\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{\alpha}) \quad = \quad \prod_{n=1}^{N} \text{Mult}(\boldsymbol{z}_n;\boldsymbol{\omega}) \prod_{k=1}^{K} p(\boldsymbol{x}_n|\boldsymbol{\theta}_k)^{z_{n,k}} p(\boldsymbol{\theta}_k|\boldsymbol{\eta}) \text{Dir}(\boldsymbol{\omega};\boldsymbol{\alpha}) \tag{23}$$

And the corresponding evidence framework problem is yielded by:

$$\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\eta},\boldsymbol{\alpha}} \sum_{\boldsymbol{Z}} \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\omega}} p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{\alpha}) \text{d}\boldsymbol{\Theta}\text{d}\boldsymbol{\omega} \tag{24}$$

Note that in Fig.2, we explicitly indicate that $\boldsymbol{\alpha}$ must be a shared prior for more than one multinomial distributions to obtain a non-trivial solution, as discussed in section 3.3.

4.1 Variational inference for evidence framework of mixture models

Table 2: A variational solution to the evidence framework of mixture models (**Algorithm II**)

```
for $i = 1 \cdots \#$iterations
    variational E-step:
        for each $\boldsymbol{x}_n \quad (1 \leq n \leq N)$
            given $\boldsymbol{\eta}^{\text{old}}, \boldsymbol{\alpha}^{\text{old}}$, calculate the responsibility $r_{nk}$ by Eq.(26)
            accumulate the statistics $\{\gamma_k[\mathbf{u}(\boldsymbol{x})], \gamma_k(1)\}$ by Eq.(27) w.r.t. $r_{nk}$
    variational M-step:
        given $\boldsymbol{\gamma}_k[\mathbf{u}(\boldsymbol{x})], \gamma_k(1)$ for all $1 \leq k \leq K$, solve $\boldsymbol{\eta}^{\text{new}}$ by Algorithm I, $\boldsymbol{\eta}^{\text{old}} \leftarrow \boldsymbol{\eta}^{\text{new}}$
        given $\gamma_k(1)$ for all $1 \leq k \leq K$, solve $\boldsymbol{\alpha}^{\text{new}}$ by Algorithm I, $\boldsymbol{\alpha}^{\text{old}} \leftarrow \boldsymbol{\alpha}^{\text{new}}$
```

To solve Eq.(24), we also conduct EM, by treating $\boldsymbol{Z}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\omega}$ as hidden variables. A hurdle now arises in calculating the joint posterior $p(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\omega}|\boldsymbol{X},\boldsymbol{\eta}^{\text{old}},\boldsymbol{\mu}^{\text{old}})$, which is analytically intractable. Therefore, we resort to the variational method [11, 10] to conduct approximated inference. We adopt the following factorization of the posterior distribution: $p(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\omega}) \approx q(\boldsymbol{Z})q(\boldsymbol{\Theta},\boldsymbol{\omega})$. As discussed in [12], the optimal solution for variational distributions $q(\boldsymbol{Z})$ and $q(\boldsymbol{\Theta},\boldsymbol{\omega})$ can be derived as:

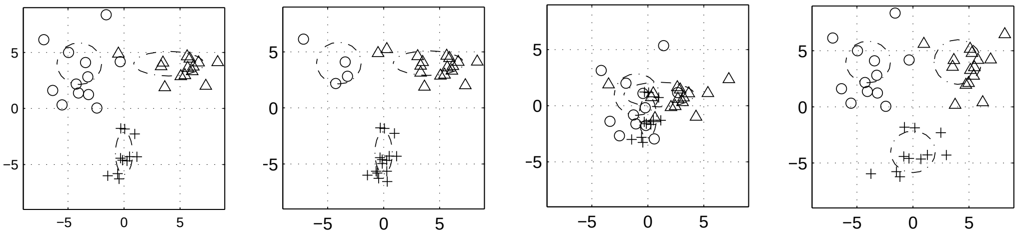
$$\begin{aligned} \ln q^*(\boldsymbol{Z}) = \quad & \langle \ln p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{\alpha}) \rangle_{p(\boldsymbol{\Theta},\boldsymbol{\omega})} + \text{const} \\ = \quad & \textstyle\sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const}' \\ \ln q^*(\boldsymbol{\Theta},\boldsymbol{\omega}) = \quad & \langle \ln p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{\alpha}) \rangle_{p(\boldsymbol{Z})} + \text{const} = \textstyle\sum_{k=1}^{K} \big[\ln p(\boldsymbol{\theta}_k|\boldsymbol{\eta}) + \\ & \textstyle\sum_{n=1}^{N} r_{nk} \ln p(\boldsymbol{x}_n|\boldsymbol{\theta}_k) + (\alpha_k - 1)\ln \omega_k + \sum_{n=1}^{N} r_{nk} \ln \omega_k \big] + \text{const}' \end{aligned} \tag{25}$$

where we denote:

$$\ln \rho_{nk} = \langle \ln \omega_k \rangle_{p(\boldsymbol{\omega}|\boldsymbol{\alpha})} + \langle \ln p(\boldsymbol{x}_n|\boldsymbol{\theta}_k) \rangle_{p(\boldsymbol{\theta}_k|\boldsymbol{\eta})}, \qquad r_{nk} = \rho_{nk}\alpha_k \big/ \textstyle\sum_{k=1}^{K} \rho_{nk}\alpha_k \tag{26}$$

Note that $q^*(\boldsymbol{Z})q^*(\boldsymbol{\Theta},\boldsymbol{\omega})$ further decomposes into $\prod_{k=1}^{K} q^*(\theta_k)q^*(\omega_k) \prod_{n=1}^{N} q^*(z_{nk})$. Based on it, we can redefine the generalized statistics of Eq.(5) with respect to $r_{nk} = q^*(z_{nk})$ as follows:

$$\boldsymbol{\gamma}_k[\boldsymbol{f}(\boldsymbol{x})] \quad = \textstyle\sum_{n=1}^{N} r_{nk} \boldsymbol{f}(\boldsymbol{x}_n) \tag{27}$$



(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) $\hat{\beta}=0.166, \hat{\tau}=13.44$

Figure 3: An illustration of the evidence framework solution by Algorithm I for Gaussian cases

We find that the optimal factorized posteriors are analogous to the posterior distributions in conventional Bayesian learning of Eq.(8), except that the statistics is redefined in Eq.(27). Accordingly, the evidence framework solution for mixture models can be implemented by a double-looped procedure. In the outer loop, we calculate r_{nk} using the current distribution of $p(\boldsymbol{\theta}|\boldsymbol{\eta}^{\text{old}})$ and $p(\boldsymbol{\omega}|\boldsymbol{\alpha}^{\text{old}})$, and then call an inner-loop. In the inner loop, we conduct Algorithm I for $\boldsymbol{\eta}^{\text{new}}$ and $\boldsymbol{\alpha}^{\text{new}}$, by using the statistics derived by r_{nk}. The procedure is shown in Table 2.

4.2 Case studies

4.2.1 Mixture of Gaussians

A mixture of Gaussians (MoG) $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a mixture model with Gaussian components. By substituting it into Eq.(26), we obtain:

$$\begin{aligned}\rho_{nk} \propto \quad & \exp\big\{ - \big[2\psi(\alpha_k) - 2\psi(\alpha_0) + \textstyle\sum_{i=1}^{D} \psi\left[(\tau_k + 1 + i)/2\right] + \\ & \ln|\boldsymbol{W}_k| - D\left(\ln\pi + \beta_k^{-1}\right) - \nu_k(\boldsymbol{x}_n - \boldsymbol{v}_k)^\top \boldsymbol{W}_k(\boldsymbol{x}_n - \boldsymbol{v}_k)\big]/2\big\} \qquad (28)\end{aligned}$$

As β and ν increases, ρ gets closer to an accurate Gaussian probability. On the other extreme, if $\tau = 0$ or $\beta \to 0^+$, which implies that $p(\boldsymbol{\theta}_k|\boldsymbol{\eta})$ is extremely uncertain, the evaluation of ρ will yield a same value for all $\boldsymbol{x}$. More detailed discussions of this special case can be found in [12].

4.2.2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) [5] can be viewed as the Bayesian version of a mixture of multinomial distributions. It has been shown that in such a case, the evidence framework is promising in achieving a more reliable model estimation. Actually, it is easy to verify that LDA is a special case of Algorithm II, when the component distributions takes a multinomial form, and a minor difference is that in LDA, the uncertainty of $p(\boldsymbol{x}|\boldsymbol{\theta}_k)$ is ignored [5].

4.2.3 Hidden Markov model (HMM)

Considering HMMs [17] with mixtures of exponential family output distributions, we can also apply Algorithm II, with minor revision, to train regularized HMMs without explicitly specifying hyperparameters. Because of space limitation, we omit derivations and only present the solution here: Given a HMM ($\boldsymbol{\pi} = \{\pi_i\}_{1\le i\le L}, \boldsymbol{A} = \{a_{ij}\}_{L\times L}, \boldsymbol{B} = \{\sum_{k=1}^{K} \omega_{ik} p(\boldsymbol{x}|\boldsymbol{\theta}_{ik})\}_{1\le i\le L}$), and a T-frame observation sequence $\boldsymbol{x}_{1:T}$, based upon Algorithm II, we can solve the evidence framework as below: 1. In variational E-step, conduct the conventional Baum-Welch algorithm [17] to collect statistic $\boldsymbol{\gamma}_{ik}[\mathbf{u}(\boldsymbol{x})], \gamma_{ik}(1)$ for all the components $p(\boldsymbol{x}|\boldsymbol{\theta}_{ik})$. Meanwhile, calculate $\rho_{t,ik}$ by Eq.(26) instead of $\omega_{ik} p(\boldsymbol{x}_t|\boldsymbol{\theta}_{ik})$. 2. Given $\boldsymbol{\gamma}_{ik}[\mathbf{u}(\boldsymbol{x})], \gamma_{ik}(1)$, conduct the standard evidence framework solution as shown in Algorithm II. A special case of HMMs with MoG output distributions is discussed in [18].

5 Experiments

5.1 A Gaussian case

We first illustrate the capability of the evidence framework by a Gaussian case. In the experiment, three 2-dimensional, diagonal covariance Gaussians share an identical Gaussian-Wishart prior, and

Table 3: Word accuracy (%) by using the evidence framework and other methods

system	#utterances, clean train (CT)				#utterances, multi train(MT)			
	100	500	2000	all	100	500	2000	all
ML	23.28	46.89	48.76	63.88	56.66	70.03	75.61	88.78
$\beta, \tau = 0.1$ (best in MT)	38.19	52.98	53.48	66.11	64.22	72.05	76.75	89.16
$\beta, \tau = 0.5$	41.84	51.73	53.85	66.17	64.31	71.60	76.62	89.14
$\beta, \tau = 2.0$ (best in CT)	45.93	55.37	54.74	66.41	64.14	68.92	74.15	88.77
evidence framework	52.10	56.79	55.54	67.13	65.85	72.58	77.18	89.19

we adopted Algorithm I to learn the optimal hyperparameters. Four training cases as well as the corresponding results are shown in Figure 3, in which the training samples belong to different classes are distinguished by their point types, and the optimal uncertainty term $\hat{\beta}$ and $\hat{\tau}$ are labeled at each case. We take (a) as a baseline, which yields a optimal $\hat{\boldsymbol{v}} = (-0.217, 1.64), \hat{\boldsymbol{W}} = \mathrm{diag}(0.593, 0.523)$, and the other three as contrastive cases. In case (b), data in class '$\circ$' is relatively less. As a result, the optimal $\hat{\boldsymbol{v}} = (0.219, 1.65), \hat{\boldsymbol{W}} = \mathrm{diag}(0.801, 0.524)$, which biases the classes with more training samples. In case (c), the three classes are moved closer to each other, which leads to a prior with larger larger $\hat{\beta} = 0.354$. It is reasonable because now we are more confident to set the position of the Gaussians. In case (d), the classes are distorted to have more similar covariances, and the resultant prior has a larger $\hat{\tau} = 13.44$, which implies a more confident pre-estimation of covariance matrix.

This illustration shows us that the evidence framework, solved by Algorithm I, is flexible to balance the uncertainty and achieve an appropriate prior estimation in various cases, which is expected to be helpful in dealing with Bayesian learning problems with evolved prior setting.

5.2 Noisy speech recognition

Second, we tested the proposed algorithm on Aurora2, a connected digit noisy speech recognition task [19]. Meanwhile, separate HMMs were built for each of the eleven digits ranging form 'zero' to 'nine', and 'oh', and 3-component MoGs were adopted as the output distributions for all the states, with all the covariance matrices set diagonal. 39-dimensional Mel-frequency cepstral coefficients (MFCCs) [3] were adopted as the features. We applied the Algorithm II with the modification discussed in section 4.2.3 to train regularized HMMs, and in the testing phase, the decoding algorithm used in [4] was adopted.

In the experiment, we used a clean training (CT) set with 8440 utterances, as well as a multiple noisy level training (MT) set with 8440 utterances [19]. Both of them were sampled to different sizes to investigate the impact of data sufficiency. We compared the word accuracies on a multiple noise condition testing set [19] of three systems: (a) Evidence framework based training Beyesian training; (b) Bayesian training with heuristic prior setting by using the method in [4], in which $\boldsymbol{v}, \boldsymbol{W}$ are derived by statistics matching, while β, τ are heuristically set. We tried to set β and τ to a series of values ranging from 0.001 to 10, but only show several representative results including the best trials in CT and MT cases. The results are listed in Table 3. Note that in CT and MT, the best heuristic trials of β, τ differ significantly, and their inappropriate values can sometimes leads to even worse performance than the ML system. This means it is hard to suggest good hyperparameters for all cases. However, by using the evidence framework, we can universally obtain decent priors in terms of accuracy. Moreover, the empirical method make it possible to find different best prior for each output distribution, which accounts for why the evidence framework performs even better than the best heuristic trials.

6 Conclusions

We study on a generic solution of evidence framework for exponential family distributions and their mixture models. We derive an EM algorithm to deal with the exponential family with proper prior sharing, with similar computational cost with ML training. By adopting the variational method, we extend it to a double-looped iterative algorithm for mixture of exponential family distributions. The proposed algorithm is analyzed and verified both theoretically and experimentally, and it was shown that the proposed algorithm is promising in various real empirical Bayesian learning applications.

References

[1] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Wiley.

[2] Carlin, B.P. & Louis, T.A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis (2nd ed.)*. Chapman & Hall/CRC

[3] Huang, X.D., Axero, A. & Hon, H.-W. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR.

[4] Watanabe, S., Minami, Y., Nakamura, A. & Ueda, N. (2002) Application of variational Bayesian approach to speech recognition, *Proc. of NIPS15*:1261–1268. MIT Press.

[5] Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**(1):993–1022.

[6] MacKay, D.J.C. (1992) Bayesian interpolation, *Neural Computation*, **4**(3):415–447.

[7] MacKay, D.J.C. (1992) The evidence framework applied to classification networks. *Neural Computation*, **4**(5):720–736.

[8] Kwok, J.T.-Y. (2000) The evidence framework applied to support vector machines. *IEEE Trans. on Neural Networks*. **11**(5):1162–1173.

[9] Dempster, A.P., Laird, N.M. & Rubin, D.B., (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistics Society (B)*. **39**(1):1–38.

[10] Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. & Saul, L. (1999) An introduction to variational methods for graphical models. *Machine Learning*. **37**:183–233.

[11] Attias, H. (2000) A variational Bayesian framework for graphical models. *Proc. of NIPS12*:209–215. MIT Press.

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

[13] Duda, R.O. & Hart, P.E. (1973) *Pattern Classification and Scene Analysis*. Wiley.

[14] Berger, A.L., Della Pietra, S.A. & Della Pietra, V.J. (1996) A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1):39–71.

[15] Boyd, S. & Vandenberghe, L. (2004) *Convex Optimization*, Cambridge University Press.

[16] Kullback, S. & Leibler, R. A. (1951) On information and sufficiency. *Annals of Mathematical Statistics*, **22**(1):79–86.

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

[18] Zhang, Y., Liu, P., Chien, J.-T. & Soong, F. (2009) An evidence framework for Bayesian learning of continuous-density hidden Markov models. *Proc. of ICASSP2009*:3857–3860.

[19] Hirsch, H.G. & Pearce, D. (2000) The aurora experimental framework for the performance evaluation of speech recognition under niosy conditions, *Proc. of ISCA ITRW ASR2000*:181–188.