Variational Bayesian Dirichlet-Multinomial Allocation for Exponential Family Mixtures

Shipeng Yu* ‡, Kai Yu‡, Volker Tresp‡, Hans-Peter Kriegel*

* Institute for Computer Science, University of Munich, and ‡ Siemens AG, Corporate Technology

{spyu, kriegel}@dbs.ifi.lmu.de, {kai.yu, volker.tresp}@siemens.com

Ludwig-Maximilians-Universität München

SIEMENS

Introduction

We study a Bayesian framework for density modeling with mixture of exponential family distributions. Our contributions:

- *A variational Bayesian solution for finite mixture models*
- *Show that finite mixture models (with a Bayesian setting) can determine the mixture number automatically*
- *Justify this result with connections to Dirichlet Process mixture models*
- *A fast variational Bayesian solution for Dirichlet Process mixture models*

Exponential Family

The probability distribution of $\mathbf{x} \in \mathcal{X}$ given parameters $\boldsymbol{\theta}$ takes the form

$$P(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\left\{\boldsymbol{\theta}^\top \phi(\mathbf{x}) - A(\boldsymbol{\theta})\right\}, \qquad (1)$$

where

- $\phi(\mathbf{x})$ is the sufficient statistics; $\boldsymbol{\theta}$ is the natural parameter.
- $A(\boldsymbol{\theta})$ is the log-partition function: $A(\boldsymbol{\theta}) = \log \int_{\mathcal{X}} h(\mathbf{x}) \exp\left\{\boldsymbol{\theta}^\top \phi(\mathbf{x})\right\} d\mathbf{x}$.
- Example distributions: Gaussian, Multinomial, Poisson, Beta, Dirichlet, ...

Conjugate Family

A prior family for exponential family distributions:

$$P(\boldsymbol{\theta}|\boldsymbol{\gamma}, \eta) = g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{\gamma} - \eta A(\boldsymbol{\theta}) - B(\boldsymbol{\gamma}, \eta)\right\}. \qquad (2)$$

This family also belongs to exponential family with sufficient statistics $\binom{\boldsymbol{\theta}}{-A(\boldsymbol{\theta})}$ and natural parameter $\binom{\boldsymbol{\gamma}}{\eta}$.

Exponential Family Mixtures

Suppose we have (a fixed number of) K component distributions, and each of them takes the same exponential family distribution.

Generative process:

- Pick one of the K components with weights $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ $(\sum_{k=1}^K \pi_k = 1)$;
- Generate one data point from the cluster-specific probability distribution.

Likelihood (for N data points IID, with $\boldsymbol{\Theta} := \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$):

$$P(\mathcal{D}|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \prod_{i=1}^N \sum_{k=1}^K P(c_i = k|\boldsymbol{\pi}) P(\mathbf{x}_i|\boldsymbol{\theta}_k) = \prod_{i=1}^N \sum_{k=1}^K \pi_k P(\mathbf{x}_i|\boldsymbol{\theta}_k). \qquad (3)$$

Priors (see plate model in Figure 1 left):

- The mixing weights $\boldsymbol{\pi}$ follow a Dirichlet distribution: $\boldsymbol{\pi} \sim \mathrm{Dir}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$;
- Each $\boldsymbol{\theta}_k$ follows (2): $P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}_0, \eta_0) = g(\boldsymbol{\theta}_k) \exp\left\{\boldsymbol{\theta}_k^\top \boldsymbol{\gamma}_0 - \eta_0 A(\boldsymbol{\theta}_k) - B(\boldsymbol{\gamma}_0, \eta_0)\right\}$. This distribution is denoted as G_0 in Figure 1.

The Algorithm

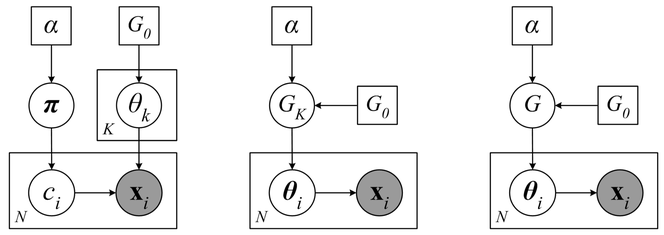


Figure 1. Plate models. Left: exponential family finite mixtures; Middle: equivalent model with G_K the finite discrete measure; Right: DP mixture model ($K \to \infty$).

Connections to Dirichlet Process Mixture Model

For finite mixture models, (3) indicates that $\boldsymbol{\theta}$ is sampled from distribution

$$G_K(\boldsymbol{\theta}) := P(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \textstyle\sum_{k=1}^K \pi_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$$

for each data point $\mathbf{x}$, which defines a discrete prior for $\boldsymbol{\theta}$ (Figure 1 middle). When $K \to \infty$ it is know in statistics that the finite mixture model approaches a DP mixture model, with α the concentration parameter, and G_0 the base distribution (see Figure 1 right). This approximation is also called Dirichlet-Multinomial Allocation (DMA).

Variational Bayesian DMA for DP Mixture Model

The VBDMA first approximates DP mixtures with finite mixtures using a large enough K, and then approximates the true posterior $P(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{c}|\mathcal{D}, \alpha, \boldsymbol{\gamma}_0, \eta_0)$ with

$$Q(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{c}|\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\varphi}) := Q(\boldsymbol{\pi}|\boldsymbol{\lambda}) \textstyle\prod_{k=1}^K Q(\boldsymbol{\theta}_k|\boldsymbol{\gamma}_k, \eta_k) \prod_{i=1}^N Q(c_i|\boldsymbol{\varphi}_i).$$

The variational Bayesian solution then maximizes a lower bound of data likelihood with respect to these variational parameters and model parameters iteratively.

- **E-step**: Update variational parameters $\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\varphi}$ analytically as

$$\varphi_{i,k} \propto \exp\left\{\mathbb{E}_{\boldsymbol{\gamma}_k, \eta_k}[\boldsymbol{\theta}_k^\top \phi(\mathbf{x}_i) - A(\boldsymbol{\theta}_k)] + \mathbb{E}_{\boldsymbol{\lambda}}[\log \pi_k]\right\}, \qquad (4)$$

$$\boldsymbol{\gamma}_k = \sum_{i=1}^N \varphi_{i,k} \phi(\mathbf{x}_i) + \boldsymbol{\gamma}_0, \quad \eta_k = \sum_{i=1}^N \varphi_{i,k} + \eta_0, \quad \lambda_k = \sum_{i=1}^N \varphi_{i,k} + \frac{\alpha}{K}. \qquad (5)$$

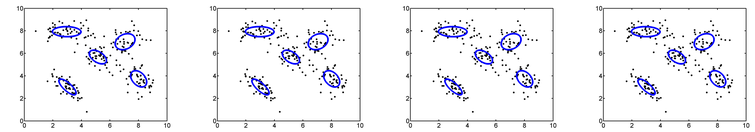
- **M-step**: Update model parameters $\alpha, \boldsymbol{\gamma}_0, \eta_0$ by matching expected sufficient statistics, e.g., $\sum_{k=1}^K \mathbb{E}_\alpha[\log \pi_k] = \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\lambda}}[\log \pi_k]$.

Key Observations:

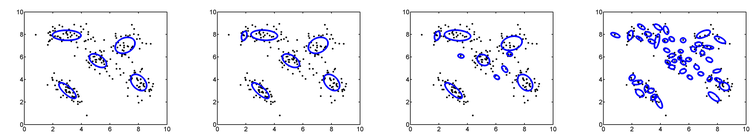
- Sparsity occurs in VBDMA with a large K; some components will get zero weights.
- α is the single parameter to control sparsity; small α leads to less components.
- Both of these two observations are due to the approximation to DP mixture models.
- VBDMA coincides with variational Bayesian solution for finite mixture models, which explains why Bayesian finite mixture models can also have sparse solutions.

Empirical Studies

VBDMA for mixture of Gaussians on a 2D toy data (5 clusters, 50 points per cluster), with different initial values for α and K:



$\alpha = 1, K = 10$ $\alpha = 1, K = 20$ $\alpha = 1, K = 100$ $\alpha = 1, K = 250$



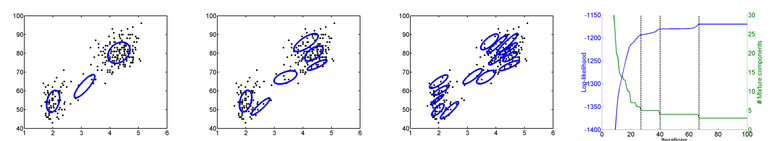
$\alpha = 1, K = 250$ $\alpha = 10, K = 250$ $\alpha = 100, K = 250$ $\alpha = 1000, K = 250$

The number of learned mixture components (mean ± std) in VBDMA (top) and VBTDP (bottom) for the same toy data. The experiments are repeated 20 times. VBTDP is the variational method for DP mixture models from Blei and Jordan, 2004. In VBDMA α better controls the sparsity of the mixture modeling.

	$K=5$	$K=10$	$K=20$	$K=50$	$K=100$	$K=250$
$\alpha=1$	4.45 ± 0.60	6.00 ± 1.03	6.70 ± 0.86	7.15 ± 1.27	6.85 ± 1.42	6.25 ± 1.16
$\alpha=10$	4.95 ± 0.22	7.80 ± 1.01	8.65 ± 1.14	7.35 ± 1.04	7.10 ± 1.37	6.45 ± 1.10
$\alpha=100$	5.00 ± 0.00	10.00 ± 0.00	19.90 ± 0.31	21.20 ± 1.58	11.40 ± 1.76	7.80 ± 1.40
$\alpha=1000$	5.00 ± 0.00	10.00 ± 0.00	20.00 ± 0.00	49.65 ± 0.49	69.05 ± 2.19	45.05 ± 2.06
$\alpha=10000$	5.00 ± 0.00	10.00 ± 0.00	20.00 ± 0.00	49.90 ± 0.31	85.10 ± 2.47	87.75 ± 2.07

	$K=5$	$K=10$	$K=20$	$K=50$	$K=100$	$K=250$
$\alpha=1$	4.50 ± 0.61	6.30 ± 1.03	7.35 ± 1.46	8.15 ± 1.39	8.55 ± 1.23	9.00 ± 1.62
$\alpha=10$	4.65 ± 0.49	6.75 ± 0.91	7.85 ± 1.14	8.50 ± 1.24	8.80 ± 1.32	9.15 ± 1.09
$\alpha=100$	4.60 ± 0.60	7.55 ± 1.15	8.95 ± 1.79	9.60 ± 1.70	9.90 ± 1.21	10.10 ± 1.33
$\alpha=1000$	4.65 ± 0.49	7.80 ± 1.01	10.45 ± 1.47	10.80 ± 2.07	11.15 ± 2.06	11.10 ± 2.31
$\alpha=10000$	4.60 ± 0.50	7.75 ± 1.02	10.20 ± 1.32	11.05 ± 2.01	11.50 ± 1.82	11.40 ± 2.19

VBDMA on the "Old Faithful" 2D data set ($N = 272$). We always initialize $K = N$. Different α values result in different mixture modeling. The right figure shows that each time the component number decreases, the log-likelihood increases. For more results on real world data sets see Shipeng Yu's Ph.D. thesis, 2006.



$\alpha = 100, K = 272$ $\alpha = 500, K = 272$ $\alpha = 1000, K = 272$ $\alpha = 100$