Bayesian Models in Machine Learning

Approximate inference in Bayesian models

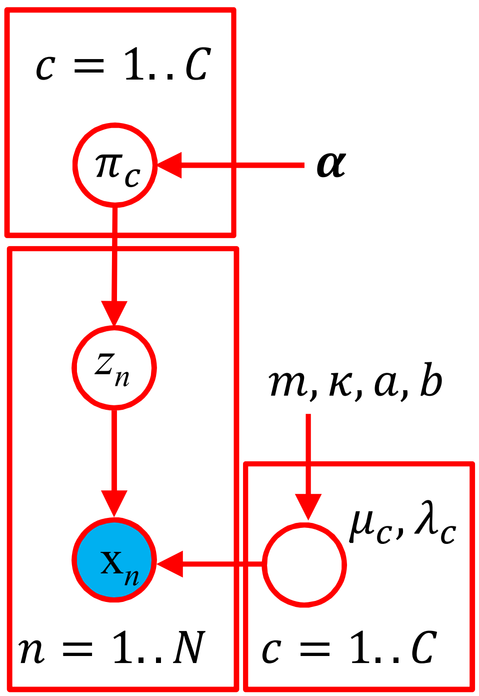
Lukáš Burget

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Escuela de Ciencias Informáticas 2017

Buenos Aires, July 24-29 2017

Bayesian Gaussian Mixture Model



- We assume that the observed data were generated as follows:
  - $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
  - For Gaussian component $c = 1 \ldots C$
    - $\mu_c, \lambda_c \sim \mathrm{NormalGamma}(\mu_c, \lambda_c | m, \kappa, a, \mathrm{b})$
  - For each observation $n = 1 \ldots N$
    - $z_n \sim \mathrm{P}(z_n | \boldsymbol{\pi}) = \mathrm{Cat}(z_n | \boldsymbol{\pi})$
    - $x_n \sim p(x_n | z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathcal{N}\left(x_n | \mu_{z_n}, \ \lambda_{z_n}^{-1}\right)$

- The task is to infer the posterior distribution of parameters $\mathrm{p}(\boldsymbol{\pi}, \mu_1, \lambda_1, \ldots \mu_C, \lambda_C | \mathbf{x})$ given some observed data $\mathbf{x} = [x_1, x_2, \ldots, x_N]$

- Intractable: need for approximations

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) = \prod_n p(x_n | z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) \prod_n P(z_n | \boldsymbol{\pi}) \prod_c p(\mu_c, \lambda_c) p(\boldsymbol{\pi})$$

Approximate inference (for Bayesian GMM)

- Variational Bayes
  - Approximate intractable $p(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}, \mathbf{z}|\mathbf{X})$ with tractable $q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}, \mathbf{z})$
  - Iteratively tune parameters of $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \mathbf{z})$ minimize $D_{KL}\ (q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \mathbf{z})||p(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}, \boldsymbol{z}|\mathbf{X}))$

- Gibbs sampling
  - Instead of obtaining $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{z}|\mathbf{X})$, we only generate samples from this distribution
  - Integrating over $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{z}|\mathbf{X})$ (e.g. for predictive distribution) can be approximated with *empirical expectations*

- …

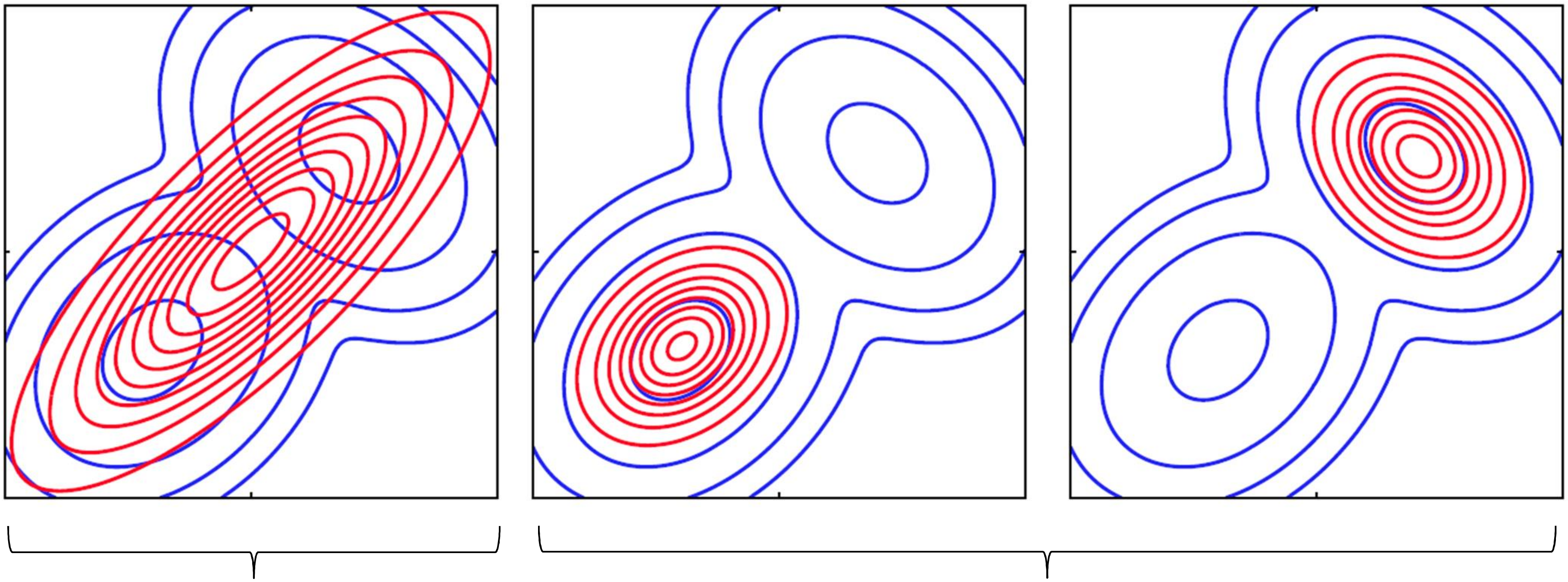
Variational Bayes

$$\ln p(\mathbf{X}) = \underbrace{\int q(\mathbf{Y}) \ln p(\mathbf{X}, \mathbf{Y}) \, \mathrm{d}\mathbf{Y} - \int q(\mathbf{Y}) \ln q(\mathbf{Y}) \, \mathrm{d}\mathbf{Y}}_{\mathcal{L}(q(\mathbf{Y}))} \underbrace{- \int q(\mathbf{Y}) \ln \frac{p(\mathbf{Y}|\mathbf{X})}{q(\mathbf{Y})} \, \mathrm{d}\mathbf{Y}}_{D_{KL}(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X}))}$$

- Find $q(\mathbf{Y})$, which is good approximation for the true posterior $p(\mathbf{Y}|\mathbf{X})$
- Maximize $\mathcal{L}(q(\mathbf{Y}))$ w.r.t. $q(\mathbf{Y})$, which in turn minimizes $D_{KL}(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X}))$
  - "Handcraft" a reasonable parametric distribution $q(\mathbf{Y}|\boldsymbol{\eta})$ and optimize $\mathcal{L}(q(\mathbf{Y}|\boldsymbol{\eta}))$ w.r.t. its parameters $\boldsymbol{\eta}$.
  - Mean field approximation assuming factorized form $q(\mathbf{Y})=q(\mathbf{Y}_1)q(\mathbf{Y}_2)q(\mathbf{Y}_3)\ldots$

Minimizing Kullback-Leibler divergence

- We optimize parameters of (simpler) distribution $q(\mathbf{Y})$ to minimize Kullback-Leibler divergence between $q(\mathbf{Y})$ and $p(\mathbf{Y}|\mathbf{X})$.



- Minimizing $D_{KL}(p(\mathbf{Y}|\mathbf{X})||q(\mathbf{Y}))$.
- Not VB objective
- Expectation propagation

- Two local optima when (numerically) minimizing $D_{KL}(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X}))$.
- VB performs this optimization

VB – Mean field approximation

- Popular Variational Bayes optimization method
- Variant of Variational Bayes, where the set of model variables $\mathbf{Y}$, can be split into subsets $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$, ..., with conditionally conjugate priors
  - $p(\mathbf{Y}_\mathrm{i}|\mathbf{X}, \mathbf{Y}_{\forall \mathrm{j} \neq \mathrm{i}})$ is tractable with conjugate prior
  - E.g. for Bayesian GMM $p(\mu_c, \lambda_c|\mathbf{X}, \mathbf{z})$ has $\mathrm{NormalGamma}$ prior
- We assume factorized approximate posterior

$$q(\mathbf{Y}) = q(\mathbf{Y}_1)q(\mathbf{Y}_2)q(\mathbf{Y}_3) \ldots = \prod_{\boldsymbol{i}} q(\mathbf{Y}_\mathrm{i})$$

- This factorization dictates the optimal (conjugate) distributions for the factors $q(\mathbf{Y}_\mathrm{i})$ and brings well defined iterative update formulas:

$$q(\mathbf{Y}_i)^* \propto \exp\left(\int q(\mathbf{Y}_{\forall \mathrm{j} \neq \mathrm{i}}) \ln p(\mathbf{X}, \mathbf{Y}) \, \mathrm{d}\mathbf{Y}_{\forall \mathrm{j} \neq \mathrm{i}}\right)$$

Mean field - update

$$\mathcal{L}(q(\mathbf{Y})) = \int q(\mathbf{Y}) \ln p(\mathbf{X}, \mathbf{Y})\, \mathrm{d}\mathbf{Y} - \int q(\mathbf{Y}) \ln q(\mathbf{Y})\, \mathrm{d}\mathbf{Y} = \int \prod_{i=1}^{M} q(\mathbf{Y}_i) \left[\ln\, p(\mathbf{X}, \mathbf{Y}) - \ln \prod_{i} q(\mathbf{Y}_i) \right] d\mathbf{Y}$$

$$= \int \prod_{i=1}^{M} q(\mathbf{Y}_i) \left[\ln\, p(\mathbf{X}, \mathbf{Y}) - \sum_{i} \ln\, q(\mathbf{Y}_i) \right] \mathrm{d}\mathbf{Y}$$

- For example, let $M = 3$
- Now, lets optimize the lower bound $\mathcal{L}(q(\mathbf{Y}_1))$ w.r.t only one distribution $q(\mathbf{Y}_1)$

$$\mathcal{L}(q(\mathbf{Y}_1)) = \iiint q(\mathbf{Y}_1) q(\mathbf{Y}_2) q(\mathbf{Y}_3) \left[\ln p(\mathbf{X}, \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3) - \ln q(\mathbf{Y}_1) - \ln q(\mathbf{Y}_2) - \ln q(\mathbf{Y}_3)\right] \mathrm{d}\mathbf{Y}_1\, \mathrm{d}\mathbf{Y}_2\, \mathrm{d}\mathbf{Y}_3$$

$$= \int q(\mathbf{Y}_1) \underbrace{\iint q(\mathbf{Y}_2) q(\mathbf{Y}_3)\ \ln p(\mathbf{X}, \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3)\, \mathrm{d}\mathbf{Y}_2\, \mathrm{d}\mathbf{Y}_3}_{\ln \tilde{p}(\mathbf{Y}_1) + const} \mathrm{d}\mathbf{Y}_1 - \int q(\mathbf{Y}_1) \ln q(\mathbf{Y}_1)\, \mathrm{d}\mathbf{Y}_1 + const$$

$$= \int q(\mathbf{Y}_1) \ln \tilde{p}(\mathbf{Y}_1)\, \mathrm{d}\mathbf{Y}_1 - \int q(\mathbf{Y}_1) \ln q(\mathbf{Y}_1)\, \mathrm{d}\mathbf{Y}_1 + const = -D_{KL}(\tilde{p}(\mathbf{Y}_1) || q(\mathbf{Y}_1)) + const$$

where $\tilde{p}(\mathbf{Y}_1)$ is normalized to be a valid distribution (therefore $+const$)

- $\mathcal{L}(q(\mathbf{Y}_1))$ is maximized by setting the D_{KL} term to zero, which implies

$$\ln q(\mathbf{Y}_1) = \ln \tilde{p}(\mathbf{Y}_1) = \iint q(\mathbf{Y}_2) q(\mathbf{Y}_3)\ \ln p(\mathbf{X}, \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3)\, \mathrm{d}\mathbf{Y}_2\, \mathrm{d}\mathbf{Y}_3 + const$$

- In general, we can iteratively update each $q(\mathbf{Y}_{\mathrm{i}})$ given the others $q(\mathbf{Y}_{\mathrm{i} \neq j})$ as:

$$q(\mathbf{Y}_j) \propto \exp \int q(\mathbf{Y}_{\forall j \neq i}) \ln p(\mathbf{X}, \mathbf{Y})\, \mathrm{d}\mathbf{Y}_{\forall j \neq i}$$

where each update guaranties to improve the lower bound $\mathcal{L}(q(\mathbf{Y}))$

Variational Bayes for GMM

- Joint likelihood for Bayesian GMM

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) = \prod_n p(x_n|z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) \prod_n P(z_n|\boldsymbol{\pi}) \prod_c p(\mu_c, \lambda_c) p(\boldsymbol{\pi})$$

$$\ln p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_n \ln p(x_n|z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \sum_n \ln P(z_n|\boldsymbol{\pi}) + \sum_c \ln p(\mu_c, \lambda_c) + \ln p(\boldsymbol{\pi})$$

where

$$\begin{aligned} p(x_n|z_n = c, \boldsymbol{\mu}, \boldsymbol{\lambda}) &= \mathcal{N}(x_n; \mu_c, \lambda_c^{-1}) \\ P(z_n = c|\boldsymbol{\pi}) &= \mathrm{Cat}(z_n = c|\boldsymbol{\pi}) = \pi_c \\ p(\mu_c, \lambda_c) &= \mathrm{NormalGamma}(\mu_c, \lambda_c|m, k, a, b) \\ p(\boldsymbol{\pi}) &= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \end{aligned}$$

- Mean field approximation $q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}, \mathbf{z}) = q(\mathbf{z}) q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi})$ dictates updates:

$$q(\mathbf{z})^* \propto \exp\left(\int q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \ln p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\lambda} \, \mathrm{d}\boldsymbol{\pi}\right)$$

$$q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi})^* \propto \exp\left(\sum_{\mathbf{z}} q(\mathbf{z}) \ln p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi})\right)$$

VBGMM – update for $q(\boldsymbol{z})$

$$
\begin{aligned}
q(\mathbf{z})^* &\propto \exp\left(\int q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \ln p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \ \mathrm{d}\boldsymbol{\mu} \ \mathrm{d}\boldsymbol{\lambda} \ \mathrm{d}\boldsymbol{\pi}\right) \\
&\propto \exp\left(\int q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \left(\sum_n \ln p(x_n | z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \sum_n \ln p(z_n | \boldsymbol{\pi})\right) \ \mathrm{d}\boldsymbol{\mu} \ \mathrm{d}\boldsymbol{\lambda} \ \mathrm{d}\boldsymbol{\pi}\right) \\
&= \exp\left(\sum_n \int q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \left(\ln p(x_n | z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \ln p(z_n | \boldsymbol{\pi})\right) \ \mathrm{d}\boldsymbol{\mu} \ \mathrm{d}\boldsymbol{\lambda} \ \mathrm{d}\boldsymbol{\pi}\right) \\
&\propto \prod_n q(z_n)^*
\end{aligned}
$$

- We see that $q(\mathbf{z})$ further factorizes - so called induced factorization

Similar to responsibilities from EM

$$
\begin{aligned}
q(z_n = c)^* &= \gamma_{nc} \\
&\propto \exp\left(\int q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \ln \mathcal{N}(x_n; \mu_c, \lambda_c^{-1}) \ \mathrm{d}\boldsymbol{\mu} \ \mathrm{d}\boldsymbol{\lambda} + \int q(\boldsymbol{\pi}) \ln \mathrm{Cat}(z_n = c | \boldsymbol{\pi}) \ \mathrm{d}\boldsymbol{\pi}\right)
\end{aligned}
$$

VBGMM – update for $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda})$

$$q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi})^* \propto \exp\left(\sum_{\mathbf{z}} q(\mathbf{z}) \ln p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi})\right)$$

$$= \exp\left(\sum_{\mathbf{z}} \prod_n q(z_n) \sum_n \{\ln p(x_n|z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \ln p(z_n|\boldsymbol{\pi})\} + \sum_c \ln p(\mu_c, \lambda_c) + \ln p(\boldsymbol{\pi})\right)$$

$$= \exp\left(\sum_c \sum_n \gamma_{nc} \{\ln p(x_n|z_n = c, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \ln p(z_n = c|\boldsymbol{\pi})\} + \ln p(\mu_c, \lambda_c) + \ln p(\boldsymbol{\pi})\right)$$

$$= \prod_c \left[\exp\left(\sum_n \gamma_{nc} \ln \mathcal{N}(x; \mu_c, \lambda_c^{-1})\right) p(\mu_c, \lambda_c)\right] \exp\left(\sum_c \sum_n \gamma_{nc} \ln p(z_n = c|\boldsymbol{\pi})\right) p(\boldsymbol{\pi})$$

$$\propto \prod_c q(\mu_c, \lambda_c)^* q(\boldsymbol{\pi})^*$$

- Again, we obtain induced factorization for $q(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi})$

$$q(\mu_c, \lambda_c)^* \propto \exp\left(\sum_n \gamma_{nc} \ln \mathcal{N}(x; \mu_c, \lambda_c^{-1})\right) \mathrm{NormalGamma}(\mu_c, \lambda_c|m, \kappa, a, b)$$

$$q(\boldsymbol{\pi})^* \propto \exp\left(\sum_c \sum_n \gamma_{nc} \ln \mathrm{Cat}(z_n = c|\boldsymbol{\pi})\right) \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

Flashback - Factorization over components

Example with only 3 fames (i.e $\mathbf{z} = [\mathrm{z}_1, \mathrm{z}_2, \mathrm{z}_3]$)

$$\sum_{\mathbf{z}} \prod_{n} q(z_n) \sum_{n} f(z_n) =$$

$$\sum_{z_1} \sum_{z_2} \sum_{z_3} q(z_1) q(z_2) q(z_3) f(z_1) + \sum_{z_1} \sum_{z_2} \sum_{z_3} q(z_1) q(z_2) q(z_3) f(z_2) + \sum_{z_1} \sum_{z_2} \sum_{z_3} q(z_1) q(z_2) q(z_3) f(z_3) =$$

$$\sum_{z_1} q(z_1) f(z_1) \sum_{z_2} q(z_2) \sum_{z_3} q(z_3) + \sum_{z_1} q(z_1) \sum_{z_2} q(z_2) f(z_2) \sum_{z_3} q(z_3) + \sum_{z_1} q(z_1) \sum_{z_2} q(z_2) \sum_{z_3} q(z_3) f(z_3) =$$

$$\sum_{z_1} q(z_1) f(z_1) + \sum_{z_2} q(z_2) f(z_2) + \sum_{z_3} q(z_3) f(z_3) =$$

$$\sum_{c=1}^{C} q(z_1 = c) f(z_1 = c) + \sum_{c=1}^{C} q(z_2 = c) f(z_2 = c) + \sum_{c=1}^{C} q(z_3 = c) f(z_3 = c) =$$

$$\sum_{c=1}^{C} \sum_{n} q(z_n = c) f(z_n = c)$$

VBGMM – update for $q(\mu_c, \lambda_c)$

$$q(\mu_c, \lambda_c)^* \propto \exp\left(\sum_n \gamma_{nc} \ln \mathcal{N}(x; \mu_c, \lambda_c^{-1})\right) \text{NormalGamma}(\mu_c, \lambda_c | m, \kappa, a, b)$$

$$= \prod_n \mathcal{N}(x; \mu_c, \lambda_c^{-1})^{\gamma_{nc}} \text{NormalGamma}(\mu_c, \lambda_c | m, \kappa, a, b)$$

$$\propto \text{NormalGamma}\left(\mu_c, \lambda_c | \frac{\kappa m + N_c \bar{x}_c}{\kappa + N_c}, \kappa + N_c, a + \frac{N_c}{2}, b + \frac{N_c}{2}\left(s_c + \frac{\kappa(\bar{x}_c - m)^2}{\kappa + N_c}\right)\right)$$

$$\propto \text{NormalGamma}\left(\mu_c, \lambda_c | m_c^*, \kappa_c^*, a_c^*, b_c^*, \right)$$

$$N_c = \sum_n \gamma_{nc}$$

$$\bar{x}_c = \frac{\sum_n \gamma_{nc} x_n}{\sum_n \gamma_{nc}}$$

$$s_c = \frac{\sum_n \gamma_{nc}(x_n - \bar{x}_c)^2}{\sum_n \gamma_{nc}}$$

Updating distribution $q(\mu_c, \lambda_c)$ means updating the parameters $m_c^*, \kappa_c^*, a_c^*, b_c^*$

VBGMM – update for $q(\boldsymbol{\pi})$

$$q(\boldsymbol{\pi})^* \propto \exp\left(\sum_c \sum_n \gamma_{nc} \ln \mathrm{Cat}(z_n = c|\boldsymbol{\pi})\right) \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$
$$\propto \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} + \mathbf{N})$$
$$\propto \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}^*)$$

$$\mathbf{N} = [N_1, N_2 \ldots, N_C]$$
$$N_c = \sum_n \gamma_{nc}$$

Updating distributions$q(\boldsymbol{\pi})$ means updating the vector $\boldsymbol{\alpha}^* = [\alpha_1^*, \alpha_2^*, \ldots \alpha_C^*]$

VBGMM – update for $q(z_n)$

$$q(z_n = c)^* \propto \exp\left(\int q(\mu_c, \lambda_c) \ln \mathcal{N}(x_n; \mu_c, \lambda_c^{-1})\, \mathrm{d}\mu_c\, \mathrm{d}\lambda_c + \int q(\boldsymbol{\pi}) \ln \mathrm{Cat}(z_n = c|\boldsymbol{\pi})\, \mathrm{d}\boldsymbol{\pi}\right)$$
$$\propto \exp\left(\psi(\alpha_c^*) - \psi\left(\sum_c \alpha_c^*\right) + \frac{\psi(a_c^*) - \ln b_c^*}{2} - \frac{1}{2\kappa_c^*} - \frac{a_c^*}{2b_c^*}(x_n - m_c^*)^2\right)$$
$$= \rho_{nc}$$

$$q(z_n = c)^* = \gamma_{nc} = \frac{\rho_{nc}}{\sum_k \rho_{nk}}$$

where $\psi(.)$ is digamma function

Updating distributions $\mathrm{q}(z_n)$ means computing responsibilities γ_{nc}

Summary of VB-GMM updates

- Update distributions $\mathrm{q}(z_n)$ (i.e. the responsibilities γ_{nc}):

$$\rho_{nc} = \exp\left(\psi(\alpha_c^*) - \psi\left(\sum_c \alpha_c^*\right) + \frac{\psi(a_c^*) - \ln b_c^*}{2} - \frac{1}{2\kappa_c^*} - \frac{a_c^*}{2b_c^*}(x_n - m_c^*)^2\right)$$

$$\gamma_{nc} = \frac{\rho_{nc}}{\sum_k \rho_{nk}}$$

- For all $c = 1..C$, update parameters of $\mathrm{q}(\mu_c, \lambda_c)$ and $\mathrm{q}(\boldsymbol{\pi})$:

$$m_c^* = \frac{\kappa m + N_c \bar{x}_c}{\kappa + N_c}$$

$$\kappa_c^* = \kappa + N_c$$

$$a_c^* = a + \frac{N_c}{2}$$

$$b_c^* = b + \frac{N_c}{2}\left(s_c + \frac{\kappa(\bar{x}_c - m)^2}{\kappa + N_c}\right)$$

$$\alpha_c^* = \alpha_c + N_c$$

$$N_c = \sum_n \gamma_{nc}$$

$$\bar{x}_c = \frac{\sum_n \gamma_{nc} x_n}{\sum_n \gamma_{nc}}$$

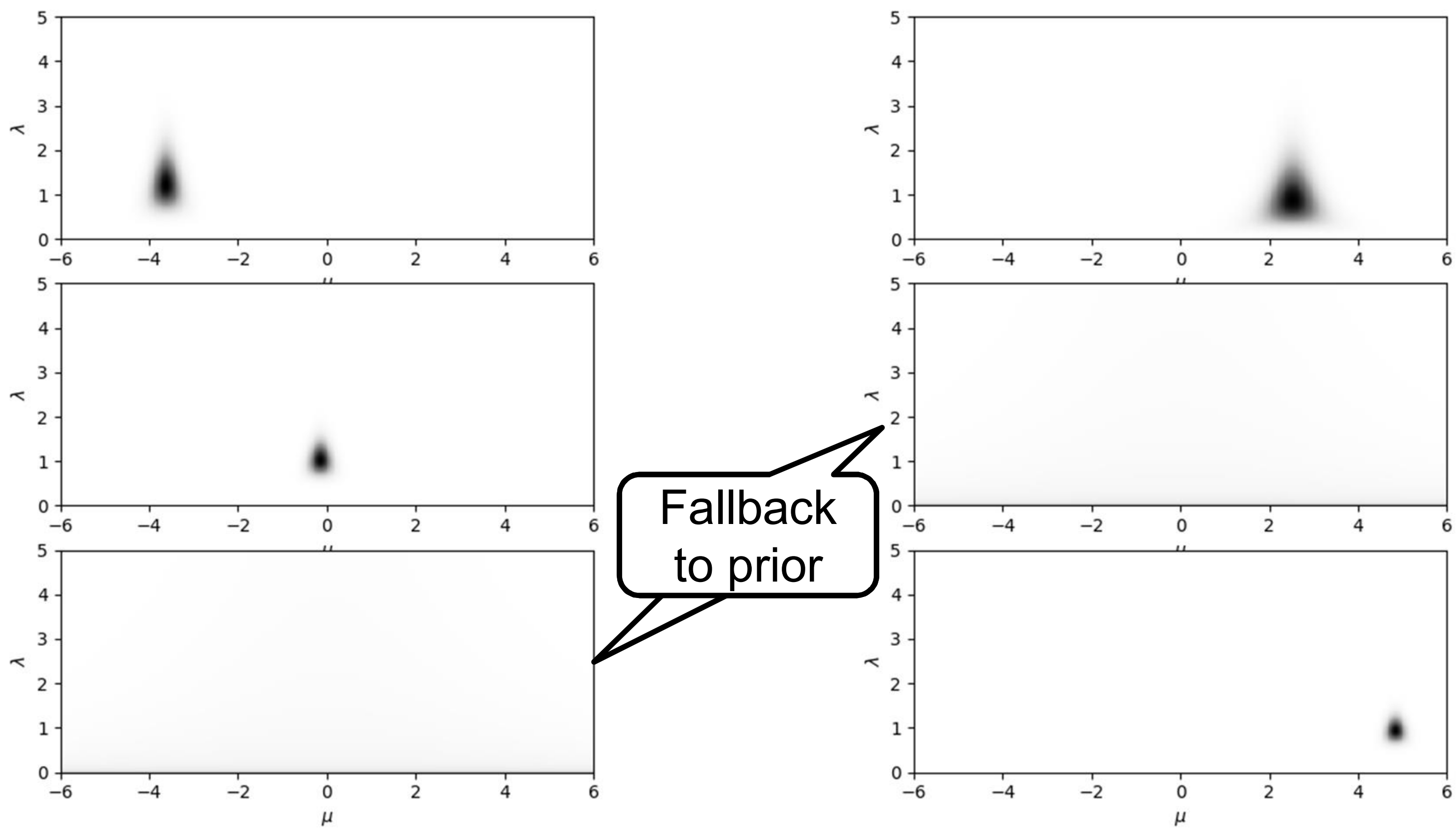
$$s_c = \frac{\sum_n \gamma_{nc}(x_n - \bar{x}_c)^2}{\sum_n \gamma_{nc}}$$

- Iterate until convergence

VB parameter posteriors

- Priors:
  - $p(\mu_c, \lambda_c), = \mathrm{NormalGamma}(\mu_c, \lambda_c | 0.0,\ 0.05,\ 0.05,\ 0.05),\ c = 1..C$
  - $p(\boldsymbol{\pi}), = \mathrm{Dir}(\boldsymbol{\pi} | [1, 1, 1, 1, 1, 1])$
- Posteriors:
  - $\boldsymbol{\alpha}_N$ = [17.1 8.3 32.2 1.0 1.0 46.4]
  - $q(\mu_c, \lambda_c)$ for the 6 Gaussian components

Fallback to prior



Evaluating VB-GMM

- Lower bound $\mathcal{L}\big(q(\mathbf{Y})\big)$ can be evaluated to check for the convergence
  - Formula not shown here

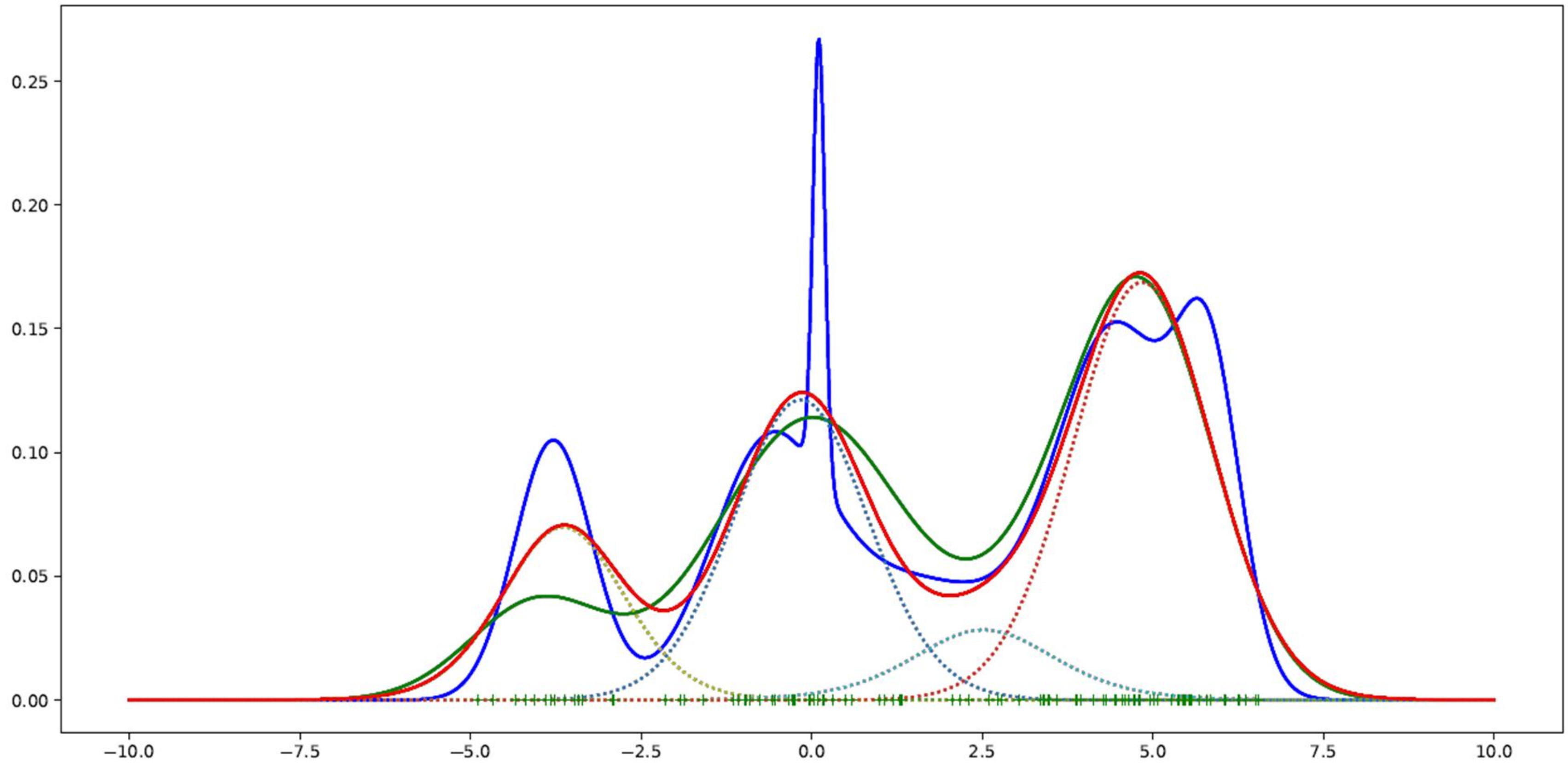
- Posterior predictive distribution is a mixture component specific posterior predictive of Student's t-distributions

$$p(x'|\mathbf{x}) = \sum\nolimits_c \mathrm{St}\left(x' \middle| m_c^*, 2\mathrm{a}_\mathrm{c}^*, \frac{a_c^* \kappa_c^*}{b_c^*(\kappa_c^* + 1)}\right) \pi_c^*$$

where mixture weights are give by categorical posterior predictive:

$$\pi_c^* = \frac{\alpha_c^*}{\sum_c \alpha_c^*}$$

VB predictive vs. ML solution



- VB was initialized from ML solution – first update of $q(\mu_c, \lambda_c)$ and $q(\boldsymbol{\pi})$ uses the responsibilities from last ML iteration
- VB recovers from ML overfitting and more robust solution closer to the true distribution for generating the training data

Approximate inference (for Bayesian GMM)

- Variational Bayes
  - Approximate intractable $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \mathbf{z}|\mathbf{X})$ with tractable $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \mathbf{z}|\mathbf{X})$
  - Iteratively tune parameters of $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \mathbf{z}|\mathbf{X})$ minimize $D_{KL}\ (q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{z}|\mathbf{X})||p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{z}|\mathbf{X}))$

- Gibbs sampling
  - Instead of obtaining $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{z}|\mathbf{X})$, we only generate samples from this distribution
  - Integrating over $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{z}|\mathbf{X})$ (e.g. for predictive distribution) can be approximated with *empirical expectations*

- …

Gibbs Sampling

- Assume we cannot sample from complex joint distribution $p(z_1, z_2)$ but it is possible to sample from conditional distributions $p(z_1|z_2)$ and $p(z_2|z_1)$
  1. Given z_1^* and generate $\mathrm{z}_2^* \sim p(z_2|z_1)$
  2. Given z_2^* and generate $\mathrm{z}_1^* \sim p(z_1|z_2)$
  3. Iterate previous two steps

- After several iterations (burn-in) the algorithm starts generating samples from $p(z_1, z_2)$
- It can be extended to more than two variables

Gibbs Sampling for Bayesian GMM

- Using sampled values of $\{\mu_c^*, \lambda_c^*\}$ and $\boldsymbol{\pi}^*$, generate new samples (hard assignments of observations to GMM components) from posterior over z_n^*
  - The distribution is just like the responsibilities from EM:

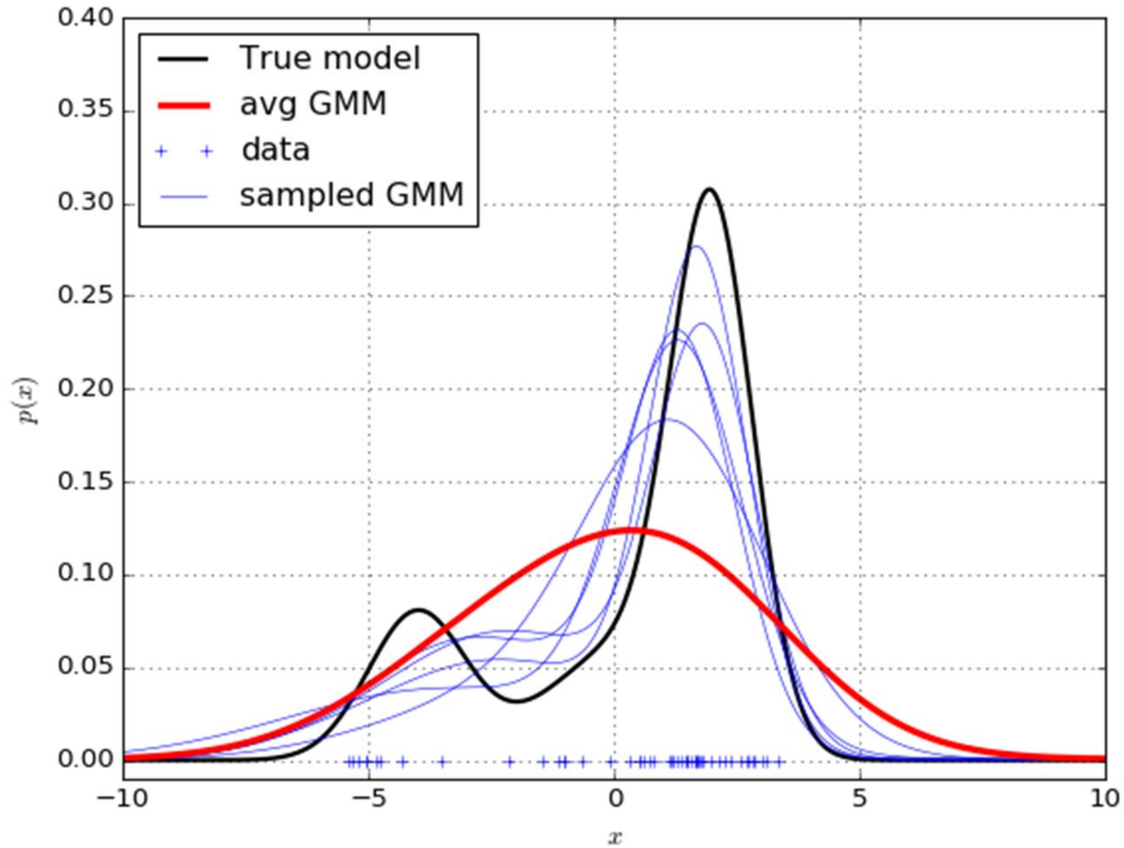
$$P(z_n = c|\mathbf{x}_n) = \frac{p(x_n|z_n = c)P(z_n = c)}{\sum_k p(x_n|k)P(k)} = \frac{\mathcal{N}(x_n|\mu_c^*, \lambda_c^{*-1})\pi_c^*}{\sum_k \mathcal{N}(\mathrm{x}_i|\mu_k^*, \lambda_k^{*-1})\pi_k^*}$$

- Using the sampled values z_n^* , for each component c, generate new samples of GMM parameters μ_c^*, λ_c^* from posteriors $p(\mu_c, \lambda_c|\mathbf{x}, \mathbf{z}^*)$
  - Estimate sufficient statistics $N_c^*, \bar{x}_c^*, s_c^*$ using the observations $\{\mathrm{x}_n: z_n = c\}$ (i.e. those hard assigned to the component c) and calculated the posterior as:

$$p(\mu_c, \lambda_c|\mathbf{x}) = \mathrm{NormalGamma}\left(\mu_c, \lambda_c \middle| \frac{\kappa m + N_c\bar{x}_c}{\kappa + N_c}, \kappa + N_c, a + \frac{N_c}{2}, b + \frac{N_c}{2}\left(s_c + \frac{\kappa(\bar{x}_c - m)^2}{\kappa + N_c}\right)\right)$$

- Sample $\boldsymbol{\pi}^*$ from posterior $p(\boldsymbol{\pi}|\mathbf{z}^*) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} + \mathbf{N}^*)$ where the vector of component occupation counts $\mathbf{N}^* = [N_1^*, N_2^*, \dots, N_C^*]$ is given by $\mathbf{z}^*$

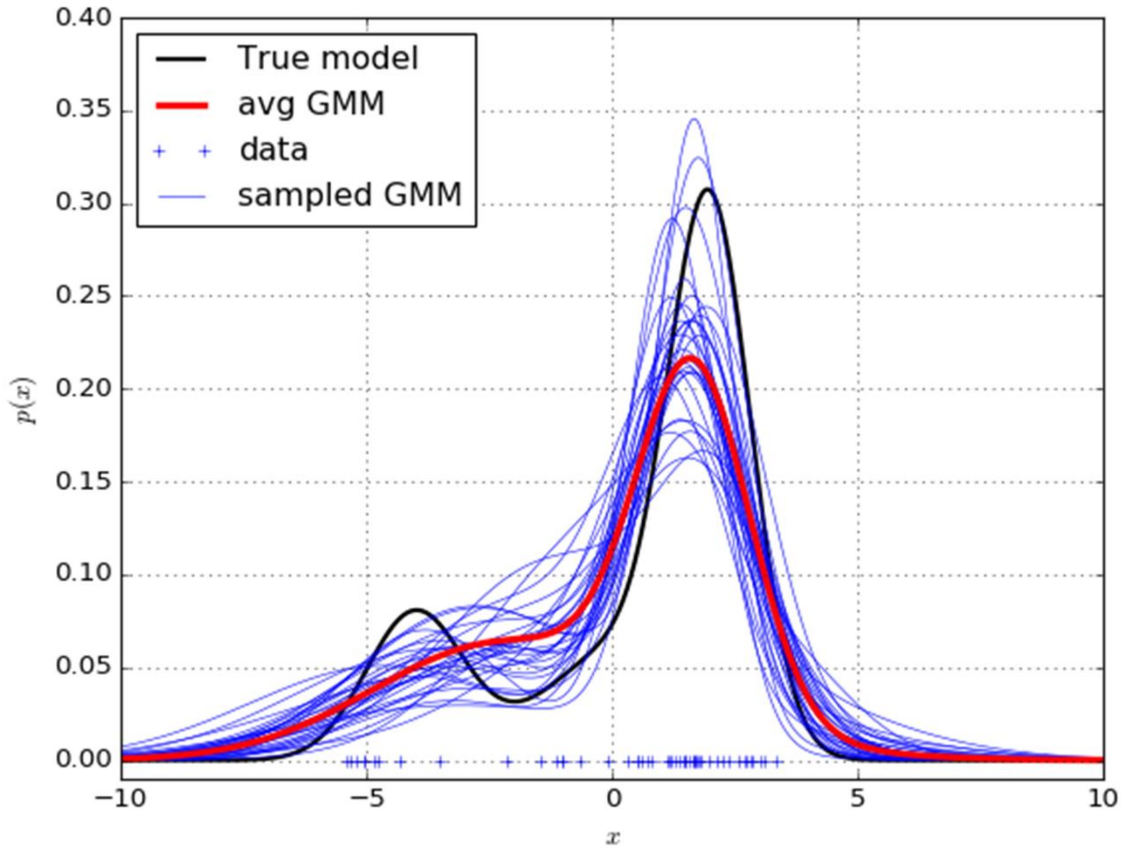
First 5-iterations of GS



Predictive distributions can be approximated by empirical expectations using the samples from the posterior distribution $\hat{\boldsymbol{\eta}}_l$:

$$p(x'|\mathbf{X}) = \int p(x'|\boldsymbol{\eta})p(\boldsymbol{\eta}|\mathbf{X})\mathrm{d}\,\boldsymbol{\eta} \approx \frac{\mathbf{1}}{\boldsymbol{L}}\sum_{\boldsymbol{l}} p(x'|\hat{\boldsymbol{\eta}}_l)$$

First 30-iterations of GS



Predictive distributions can be approximated by empirical expectations using the samples from the posterior distribution $\hat{\boldsymbol{\eta}}_l$:

$$p(x'|\mathbf{X}) = \int p(x'|\boldsymbol{\eta})p(\boldsymbol{\eta}|\mathbf{X})\mathrm{d}\,\boldsymbol{\eta} \approx \frac{1}{L}\sum_{l} p(x'|\hat{\boldsymbol{\eta}}_l)$$

Collapsed GS for Bayesian GMM

- Sampling discrete latent variables like z_n is fine as they have limited number of possible values
- For continuous latent variables like $\boldsymbol{\pi}, \mu_c, \lambda_c$, however, we might need too many samples to get a reasonable representation of their posterior distributions (especially for multivariate higher dimensional variables).
- Collapsed Gibbs Sampling
  - Iterates over (and samples only from) a subset of the latent variables in the model (e.g. the discrete ones)
  - integrates (marginalizes) over the remaining (continuous) variables
- CBS for Bayesian GMM:

for $i = 1..N$

$$z_i^* \sim P(z_i|\mathbf{x}, \mathbf{z}_{\setminus i})$$

where

$\mathbf{z}_{\setminus \mathrm{i}}$ is $\mathbf{z}$ with z_i removed

$\mathbf{x}_{\setminus \mathrm{i}}$ is $\mathbf{x}$ with x_i removed

CGS for BGMM - $P\left(z_i \middle| \mathbf{z}_{\setminus \mathrm{i}}\right)$

- How do we obtain $p(z_i|\mathbf{x}, \mathbf{z}_{\setminus \mathrm{i}})$?
- Lets first introduce some useful distributions
- Posterior distribution of weights $\boldsymbol{\pi}$ given $\mathbf{z}_{\setminus \mathrm{i}}$ (or corresponding vector of component occupation counts $\mathbf{N}_{\setminus \mathrm{i}}$)

$$p(\boldsymbol{\pi}|\mathbf{z}_{\setminus i}) \propto \prod_{n \neq i} P(z_n|\boldsymbol{\pi}) p(\boldsymbol{\pi}) = \prod_{n \neq i} \mathrm{Cat}(z_n|\boldsymbol{\pi}) \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \propto \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} + \mathbf{N}_{\setminus i})$$

- Posterior predictive distribution for z_i given $\mathbf{z}_{\setminus \mathrm{i}}$

$$P(z_i|\mathbf{z}_{\setminus i}) = \int P(z_i|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\mathbf{z}_{\setminus i}) \, \mathrm{d}\boldsymbol{\pi} = \int \mathrm{Cat}(z_i|\boldsymbol{\pi}) \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} + \mathbf{N}_{\setminus i}) \, \mathrm{d}\boldsymbol{\pi}$$

$$= \mathrm{Cat}\left(z_i \middle| \frac{\boldsymbol{\alpha} + \mathbf{N}_{\setminus i}}{\sum_c \alpha_c + N - 1}\right)$$

CGS for BGMM - $p\left(x_i \middle| z_i, \mathbf{x}_{\setminus\mathrm{i}}, \mathbf{z}_{\setminus\mathrm{i}}\right)$

- Let $S_{c\setminus i}$ define the subset of observations assigned by $\mathbf{z}_{\setminus\mathrm{i}}$ to component c
- Posterior distribution of μ_c, λ_c given $\mathbf{x}_{\setminus\mathrm{i}}, \mathbf{z}_{\setminus\mathrm{i}}$ is estimated in the usual way using only the observations $S_{c\setminus i}$

$$p(\mu_c, \lambda_c | \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i}) \propto \prod_{n \in S_{c\setminus i}} p(x_n | z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) p(\boldsymbol{\mu}, \boldsymbol{\lambda})$$

$$= \prod_{n \in S_{c\setminus i}} \mathcal{N}(x_n | \mu_c, \lambda_c^{-1}) \mathrm{NormalGamma}(\mu_c, \lambda_c | m, \kappa, a, b)$$

$$\propto \prod_{n \in S_{c\setminus i}} \mathrm{NormalGamma}(\mu_k, \lambda_k | m^*_{c\setminus i}, \kappa^*_{c\setminus i}, a^*_{c\setminus i}, b^*_{c\setminus i})$$

- Posterior predictive distrib. of x_i for component c given observations $S_{c\setminus i}$

$$p(x_i | z_i = c, \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i}) = \int p(x_i | z_i = c, \mu_c, \lambda_c) p(\mu_c, \lambda_c | \mathbf{x}_{\setminus i}, \mathbf{x}_{\setminus i}) \, \mathrm{d}\mu_c, \, \mathrm{d}\lambda_c$$

$$= \int \mathcal{N}(x_i | \mu_c, \lambda_c^{-1}) \mathrm{NormalGamma}(\mu_c, \lambda_c | m^*_{c\setminus i}, \kappa^*_{c\setminus i}, a^*_{c\setminus i}, b^*_{c\setminus i}) \, \mathrm{d}\mu_c \, \mathrm{d}\lambda_c$$

$$= \mathrm{St}\left(x_i | m^*_{c\setminus i}, 2a^*_{c\setminus i}, \frac{a^*_{c\setminus i} \kappa^*_{c\setminus i}}{b^*_{c\setminus i}(\kappa^*_{c\setminus i} + 1)} \right)$$

CGS for BGMM - $p(x_i|\mathbf{x}, \mathbf{z}_{\setminus \mathrm{i}})$

- Finally, using Bayes rule

$$P(z_i|\mathbf{x}, \mathbf{z}_{\setminus i}) = P(z_i|x_i, \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i}) = \frac{p(x_i|z_i, \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i})P(z_i|\mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i})}{\sum_c p(x_i|z_i = c, \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i})P(z_i = c|\mathbf{z}_{\setminus i})}$$

- The Collapsed Gibbs sampling iterations

$$\text{for } i = 1..N$$
$$z_i^* \sim P(z_i|\mathbf{x}, \mathbf{z}_{\setminus i})$$

gives us samples from $\mathbf{z}^* \sim p(\mathbf{z}|\mathbf{x})$. What can we do with that?

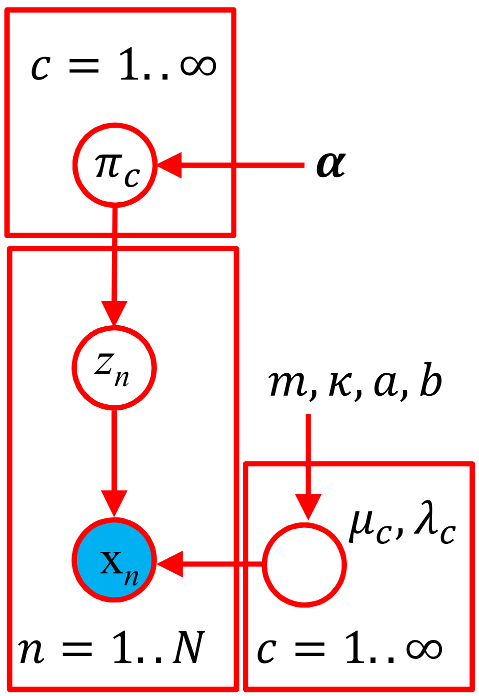
- GMM posterior predictive distribution for new x' given $\mathbf{x}$ and (sampled) $\mathbf{z}$

$$p(x'|\mathbf{x}, \mathbf{z}) = \sum_c p(x'|z = c, \mathbf{x}, \mathbf{z})P(z = c|\mathbf{x})$$

- Full predictive distribution can be approximated using the samples $\mathbf{z}_l^*$ as

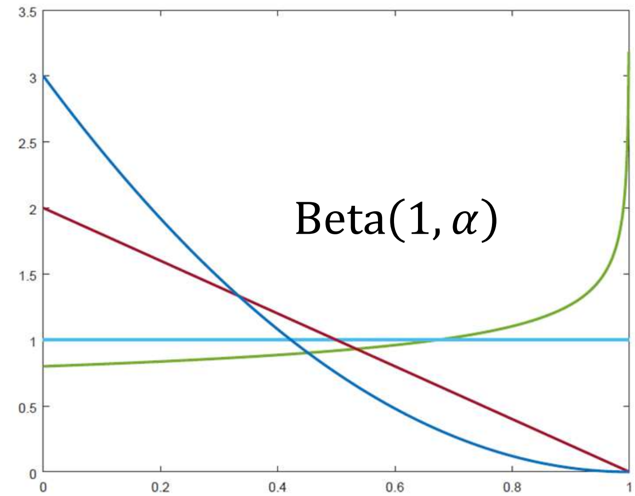
$$p(x|\mathbf{x}) = \sum_{\mathbf{z}} p(x'|\mathbf{x}, \mathbf{z})p(\mathbf{z}|\mathbf{x}) \approx \frac{1}{L}\sum_l p(x'|\mathbf{x}, \mathbf{z}_l^*)$$

Infinite Bayesian GMM



- Lets consider Bayesian GMM with an infinite number of Gaussian components $c = 1..\infty$
- The priors for μ_c, λ_c for Gaussian component $c = 1 \ldots \infty$ can be defined as before:
  - $p(\mu_c, \lambda_c) \sim \mathrm{NormalGamma}(\mu_c, \lambda_c | \, m, \kappa, a, \mathrm{b})$
- However, we need an infinite number of mixture weights $\boldsymbol{\pi}$ = $[\pi_1, \pi_2, \ldots\ ]$ so that $\sum_{c=1}^{\infty} \pi_c = 1$
- We also need a suitable prior distribution for $\boldsymbol{\pi}$

Stick breaking process - GEM

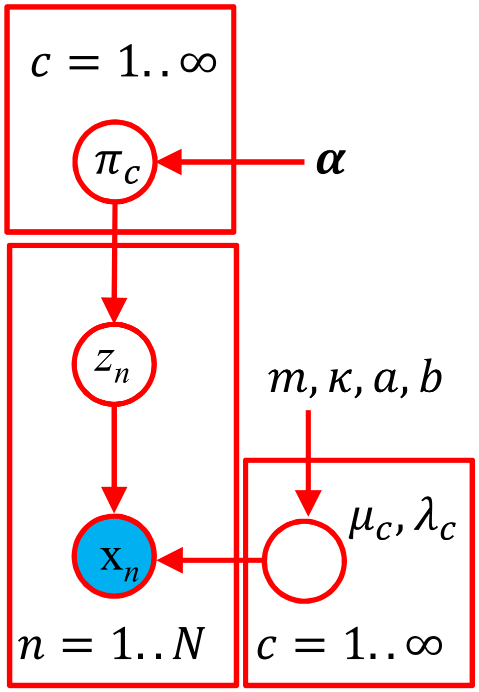


- for $c = 1,2,\ldots,\infty$
  $v_c \sim \text{Beta}(1,\alpha)$
  $\pi_c = v_c \prod_{k=1}^{c-1} (1 - v_k)$

- Take a unit length stick
  For $c = 1,2,\ldots,\infty$
  - Generate v_c in range $(0,1)$ from $\text{Beta}(1,\alpha)$
  - Break the stick into two pieces with proportions $v_c : 1 - v_c$
  - The length of the first piece corresponds to π_c
  - The second piece is the stick to be broken in further iterations

- The resulting infinite dimensional vector of weights is a sample from the stick breaking process $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$ (Griffiths, Engen and McCloskey)
- $\text{GEM}(\alpha)$ can be used as a prior for infinite number of component weights
- With small concentration parameter α, only few weights will be non-negligable

Infinite Bayesian GMM



- We assume that the observed data were generated as follows:
  - $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots] \sim \mathrm{GEM}(\alpha)$
  - For Gaussian component $c = 1 \dots \infty$
    - $\mu_c, \lambda_c \sim \mathrm{NormalGamma}(\mu_c, \lambda_c | m, \kappa, a, \mathrm{b})$
  - For each observation $i = 1 \dots N$
    - $z_n \sim \mathrm{P}(z_n | \boldsymbol{\pi}) = \mathrm{Cat}(z_n | \boldsymbol{\pi})$
    - $x_n \sim p(x_n | z_n, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathcal{N}\left(x_n | \mu_{z_n}, \ \lambda_{z_n}^{-1}\right)$

- Obviously the observed data can be generated from at most N Gaussian components.

- Again, the task is to infer the posterior distribution of parameters $\mathrm{p}(\boldsymbol{\pi}, \mu_1, \lambda_1, \mu_2, \lambda_2 \dots | \mathbf{x})$ given some observed data $\mathbf{x} = [x_1, x_2, \dots, x_N]$

CGS for infinite Bayesian GMM

- We can use the same Collapsed Gibbs sampling iterations that we used in the case of the BGMM with fixed number of Gaussian

for $i = 1..N$

$$z_i^* \sim P(z_i | \mathbf{x}, \mathbf{z}_{\setminus i})$$

where again $$P(z_i | \mathbf{x}, \mathbf{z}_{\setminus i}) = \frac{p(x_i | z_i, \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i}) P(z_i | \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i})}{\sum_c p(x_i | z_i = c, \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i}) P(z_i = c | \mathbf{z}_{\setminus i})}$$

and the component posterior predictive

$$p(x_i | z_i = c, \mathbf{x}_{\setminus i}, \mathbf{z}_{\setminus i}) = \mathrm{St}\left(x_i | m_{c\setminus i}^*, 2a_{c\setminus i}^*, \frac{a_{c\setminus i}^* \kappa_{c\setminus i}^*}{b_{c\setminus i}^* (\kappa_{c\setminus i}^* + 1)} \right)$$

- The only difference will be in $P(z_i | \mathbf{z}_{\setminus \mathrm{i}})$, which is evaluated using Chinese Restaurant Process (CRP)

Chinese Restaurant Process

- Let the prior on the infinite weight vector be $p(\boldsymbol{\pi}) = \mathrm{GEM}(\pi|\alpha)$
- Let $z_n, \;\; n = 1..N$ be samples generated from an (unknown) “infinite categorical distribution” $\mathrm{Cat}(z_n|\boldsymbol{\pi})$
- The posterior $p(\boldsymbol{\pi}|\mathbf{z}) \propto \;\prod_n p(z_n|\boldsymbol{\pi})\, p(\boldsymbol{\pi})$ is intractable
  - We cannot even easily sample from it as the sample would be infinite vector of weights
- However the predictive posterior $P(z'|\mathbf{z}) = \int P(z'|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\mathbf{z}) \mathrm{d}\boldsymbol{\pi}$ can be evaluated as

$$P(z' = c|\mathbf{z}) = \frac{N_c}{\alpha + N}$$

$$P(z' = C + 1|\mathbf{z}) = \frac{\alpha}{\alpha + N}$$

where N_c is the number of observations assigned by $\mathbf{z}$ to category c and $C + 1$ is a new so far not seen category.

Chinese Restaurant Process

- Imagine Chinese Restaurant with an infinite number of tables, each with infinite capacity
- The first customer sits at the first table
- Every new customer:
  - Joins already occupied table with probability proportional to the number of customers sitting at that table

$$P(z' = c|\mathbf{z}) = \frac{N_c}{\alpha + N}$$

  - or starts a new table with probability proportional to concentrarion parameter α

$$P(z' = C + 1|\mathbf{z}) = \frac{\alpha}{\alpha + N}$$

Dirichlet Process

We have defined Infinite BGMM as (for simplicity assuming the same σ for all Gaussian component variances σ and conjugate prior $p(\mu_c) = \mathcal{N}(\mu_c|\mu_0, \sigma_0)$):

$$
\begin{aligned}
&\boldsymbol{\pi} = [\pi_1, \pi_2, \dots] \sim \mathrm{GEM}(\alpha) \\
&\mu_c \sim \mathcal{N}(\mu_c|\mu_0, \sigma_0), \qquad c = 1..\infty \\
&z_i \sim \boldsymbol{\pi}, \qquad i = 1..N \\
&x_i \sim \mathcal{N}(x_i|\mu_c, \sigma), \qquad i = 1..N
\end{aligned}
$$

Alternative definition using $\delta_\mu(\tilde{\mu}) = \begin{cases} 1, & \mu_c = \tilde{\mu} \\ 0, & \mu_c \neq \tilde{\mu} \end{cases}$

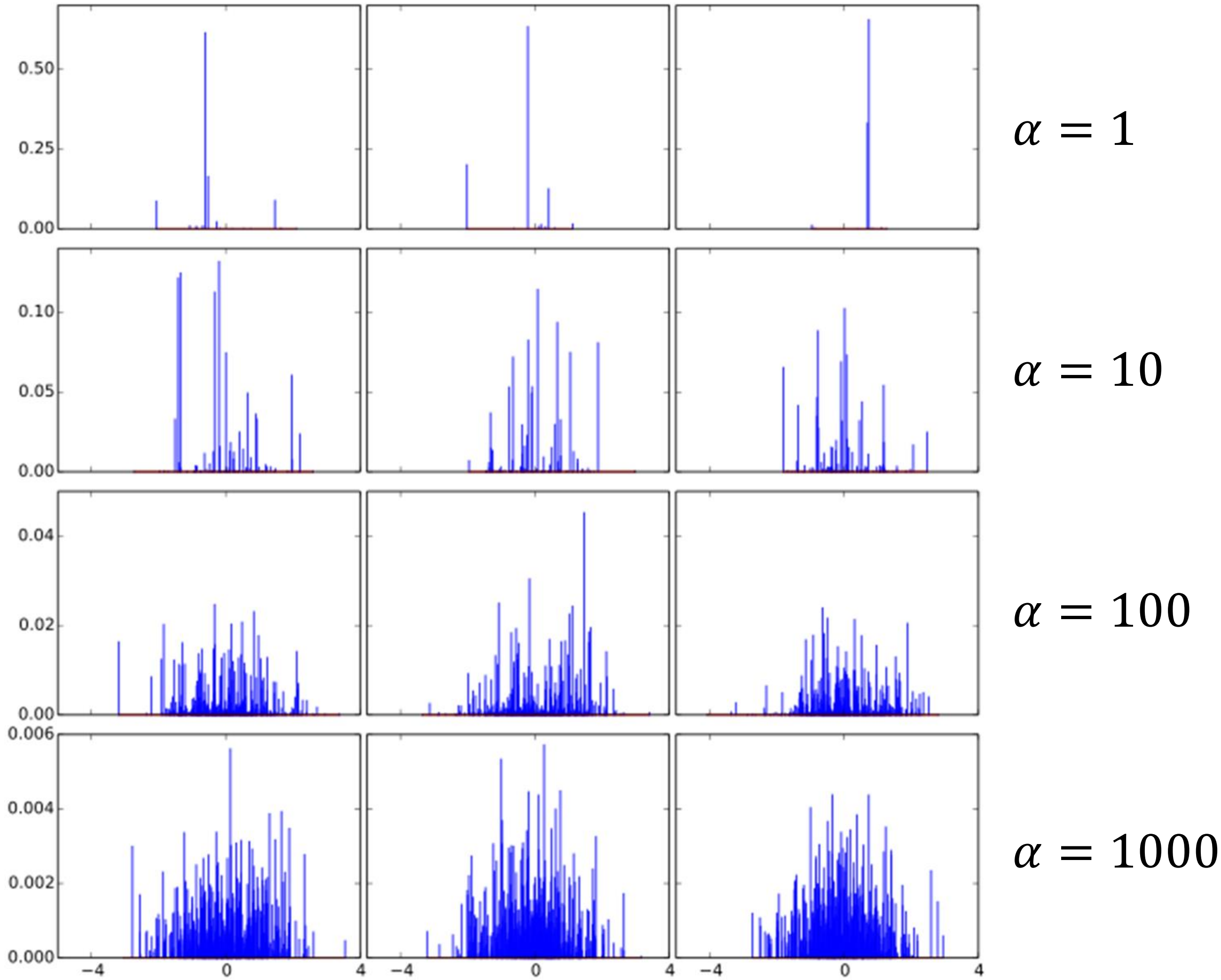
$$
\begin{aligned}
&\boldsymbol{\pi} = [\pi_1, \pi_2, \dots] \sim \mathrm{GEM}(\alpha) \\
&\mu_c \sim \mathcal{N}(\mu_c|\mu_0, \sigma_0), \qquad c = 1..\infty \\
&\tilde{\mu}_i \sim G = \sum_{c=1}^{\infty} \pi_c \delta_{\mu_c}(\tilde{\mu}_i), \qquad i = 1..N \\
&x_i \sim \mathcal{N}(x_i|\tilde{\mu}_i), \qquad i = 1..N
\end{aligned}
$$

or using Dirichlet Process with base distribution $H = \mathcal{N}(\mu_0, \sigma_0)$ and concentration parameter α

$$
\begin{aligned}
&G \sim DP(H, \alpha) \\
&\tilde{\mu}_i \sim G, \qquad i = 1..N \\
&x_i \sim \mathcal{N}(x_i|\tilde{\mu}_i), \qquad i = 1..N
\end{aligned}
$$

Dirichlet process

Samples $G \sim \mathrm{DP}(\mathcal{N}(0,1), \alpha)$



G is discrete distribution with continuous support

$\mathrm{DP}(\mathcal{N}(0,1), \alpha)$ is distribution over discrete distributions with continuous support