DPGMM Variational Inference

Maddie Cusimano and Luke Hewitt

This document contains the evidence lower bound and updates for variational inference in a Dirichlet Process isotropic Gaussian mixture model, as well as their derivations.

1 Setup

1.1 Variables

$$
\begin{aligned}
X &: \text{data} \\
N &: \text{number of datapoints} \\
D &: \text{dimensionality of data} \\
i &: \text{datapoint index} \\
T &: \text{truncation parameter, i.e., number of clusters in variational distribution} \\
k &: \text{cluster index} \\
\phi_k &: \text{cluster stick-breaking proportion} \\
\mu_k &: \text{cluster mean} \\
\tau_k &: \text{cluster precision} \\
z_i = k &: \text{assignment of datapoint } i \text{ to cluster } k
\end{aligned}
$$

1.2 Model, P

We use the 'rate' or 'inverse-scale' parametrization of the Gamma distribution, and the stick breaking process view of the Dirichlet process.

$$
\begin{aligned}
\phi_k &\sim Beta(1, \alpha) \\
\mu_k &\sim Gaussian(0, \mathbf{I}) \\
\tau_k &\sim Gamma(\text{shape} = 1, \text{rate} = 1) \\
z_i &\sim SBP(\boldsymbol{\phi}) \\
x_i &\sim Gaussian(\mu_{z_i}, \frac{1}{\tau_{z_i}}\mathbf{I})
\end{aligned}
$$

1.3 Variational distribution, Q

We use the truncated mean-field variational approximation set out in Blei and Jordan, with the variational distribution factorising as:

$$Q(\boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{z}) = \prod_{k=1}^{T} Q_{\phi_k}(\phi_k) Q_{\mu_k}(\mu_k) Q_{\tau_k}(\tau_k) \prod_{i=1}^{N} Q_{z_i}(z_i)$$

$$\begin{aligned} \phi_k &\sim \mathit{Beta}(\lambda_{k,1}, \lambda_{k,2}) \\ \mu_k &\sim \mathit{Gaussian}(\nu_k, \mathbf{I}) \\ \tau_k &\sim \mathit{Gamma}(a_k, b_k) \\ z_i &\sim \mathit{Discrete}(\boldsymbol{\zeta}_i) \end{aligned}$$

Where the distribution is clear from context, we omit the subscript on Q and write simply $Q(\phi_k)$ or $\mathbb{E}_{\phi_k \sim Q}[\ldots]$

2 Evidence lower bound

Maximizing the evidence lower bound provides the variational updates.

$$\begin{aligned} \log P(X) \geq & \sum_{k=1}^{T} \mathop{\mathbb{E}}_{\phi_k \sim Q} \big[\log P(\phi_k) - \log Q(\phi_k) \big] \\ & + \sum_{k=1}^{T} \mathop{\mathbb{E}}_{\mu_k \sim Q} \big[\log P(\mu_k) - \log Q(\mu_k) \big] \\ & + \sum_{k=1}^{T} \mathop{\mathbb{E}}_{\tau_k \sim Q} \big[\log P(\tau_k) - \log Q(\tau_k) \big] \\ & + \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\boldsymbol{\phi}, z_i \sim Q} \big[\log P(z_i | \boldsymbol{\phi}) - \log Q(z_i) \big] \\ & + \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\tau} \sim Q} \big[\log P(x_i | \mu_{z_i}, \tau_{z_i}) \big] \end{aligned}$$

2.1 ϕ_k term

This bound is equal to the negative KL divergence between Beta distributions.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\phi_k \sim Q}\big[\log P(\phi_k) - \log Q(\phi_k)\big] &= -D_{KL}(Q_{\phi_k}||P_{\phi_k}) \\
&= \log\Gamma(1+\alpha) - \log\Gamma(\alpha) \\
&+ (\alpha - 1)(\Psi(\lambda_{k,2}) - \Psi(\lambda_{k,1} + \lambda_{k,2})) \\
&- \log\Gamma(\lambda_{k,1} + \lambda_{k,2}) + \log\Gamma(\lambda_{k,1}) + \log\Gamma(\lambda_{k,2}) \\
&- (\lambda_{k,1} - 1)(\Psi(\lambda_{k,1}) - \Psi(\lambda_{k,1} + \lambda_{k,2})) \\
&- (\lambda_{k,2} - 1)(\Psi(\lambda_{k,2}) - \Psi(\lambda_{k,1} + \lambda_{k,2}))
\end{aligned}
$$

2.2 μ_k term

The bound is equal to the negative KL divergence between isotropic Gaussians.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\mu_k \sim Q}\big[\log P(\mu_k) - \log Q(\mu_k)\big] &= -D_{KL}(Q_{\mu_k}||P_{\mu_k}) \\
&= -\frac{1}{2}||\nu_k||^2
\end{aligned}
$$

2.3 τ_k term

The bound is equal to the negative KL divergence between Gamma distributions.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\tau_k \sim Q}\big[\log P(\tau_k) - \log Q(\tau_k)\big] &= -D_{KL}(Q_{\tau_k}||P_{\tau_k}) \\
&= \log\Gamma(a_k) - (a_k - 1)\Psi(a_k) - log(b_k) + a_k - \frac{a_k}{b_k}
\end{aligned}
$$

2.4 z_i term

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\phi}, z_i \sim Q}\big[\log P(z_i|\boldsymbol{\phi}) - \log Q(z_i)\big] = \sum_{k=1}^{T} \zeta_{i,k}\Big[&- \log(\zeta_{i,k}) + \Psi(\lambda_{1,k}) - \Psi(\lambda_{1,k} + \lambda_{2,k}) \\
&+ \sum_{j=1}^{k-1}\Big(\Psi(\lambda_{j,2}) - \Psi(\lambda_{1,j} + \lambda_{2,j})\Big)\Big]
\end{aligned}
$$

For convenience later, we define:

$$
\eta_{z_{i,k}} = \Psi(\lambda_{1,k}) - \Psi(\lambda_{1,k} + \lambda_{2,k}) + \sum_{j=1}^{k-1}\Big(\Psi(\lambda_{j,2}) - \Psi(\lambda_{1,j} + \lambda_{2,j})\Big)
$$

2.4.1 Derivation

The stick breaking process gives:

$$P(z_i = k) = \phi_k \prod_{j=1}^{k-1} 1 - \phi_j$$

To derive the bound, we use the following properties of a Beta-distributed variable $B \sim Beta(b_1, b_2)$: 1) $\mathbb{E}[\log B] = \Psi(b_1) - \Psi(b_1 + b_2)$, 2) $1 - B \sim Beta(b_2, b_1)$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\phi}, z_i \sim Q}\Big[\log P(z_i|\boldsymbol{\phi}) - \log Q(z_i)\Big] &= \mathop{\mathbb{E}}_{z_i \sim Q}\left[\mathop{\mathbb{E}}_{\boldsymbol{\phi} \sim Q}\left[\log \phi_{z_i} + \sum_{j=1}^{z_i - 1} \log(1 - \phi_j) - \log(\zeta_i)\right]\right] \\
&= \mathop{\mathbb{E}}_{z_i \sim Q}\left[\mathop{\mathbb{E}}_{\boldsymbol{\phi} \sim Q}[\log \phi_{z_i}] + \sum_{j=1}^{z_i - 1} \mathop{\mathbb{E}}_{\boldsymbol{\phi} \sim Q}[\log(1 - \phi_j)] - \log(\zeta_i)\right] \\
&= \mathop{\mathbb{E}}_{z_i \sim Q}\Bigg[\Psi(\lambda_{z_i,1}) - \Psi(\lambda_{z_i,1} + \lambda_{z_i,2}) \\
&\qquad + \sum_{j=1}^{z_i - 1} (\Psi(\lambda_{z_j,1}) - \Psi(\lambda_{z_j,1} + \lambda_{z_j,2})) - \log(\zeta_i)\Bigg]
\end{aligned}$$

Taking the expectation with respect to $Q(z_i)$ gives the stated bound.

2.5 x_i term

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\tau} \sim Q}\Big[\log P(x_i|\mu_{z_i}, \tau_{z_i})\Big] = \sum_{k=1}^{T} \zeta_{i,k} \Bigg(&- \frac{D}{2}\Big(\log(2\pi) - \Psi(a_k) + \log(b_k)\Big) \\
&- \frac{a}{2b}\Big(||x_i - \nu_k||^2 + \frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}} 0.5^{1.5}}\Big)\Bigg)
\end{aligned}$$

For convenience later, we define:

$$\eta_{x_{i,k}} = -\frac{D}{2}\Big(\log(2\pi) - \Psi(a_k) + \log(b_k)\Big) - \frac{a}{2b}\Big(||x_i - \nu_k||^2 + \frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}} 0.5^{1.5}}\Big)$$

2.5.1 Derivation

First taking the expectation with respect to Q(z_i):

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{z},\boldsymbol{\mu},\boldsymbol{\tau}\sim Q}\left[\log P(x_i|\mu_{z_i},\tau_{z_i})\right] \\
&= \sum_{k=1}^{T}\zeta_{i,k}\mathop{\mathbb{E}}_{\tau_k\sim Q}\left[\mathop{\mathbb{E}}_{\mu_k\sim Q}[\log P(x_i|\mu_k,\tau_k)]\right] \\
&= \sum_{k=1}^{T}\zeta_{i,k}\mathop{\mathbb{E}}_{\tau_k\sim Q}\left[\mathop{\mathbb{E}}_{\mu_k\sim Q}[\log N(x_i;\mu_k,\frac{1}{\tau_k}\mathbf{I})\right] \\
&= \sum_{k=1}^{T}\zeta_{i,k}\mathop{\mathbb{E}}_{\tau_k\sim Q}\left[\int_{\mu_k}N(\mu_k;\nu_k,\mathbf{I})\log N(x_i;\mu_k,\frac{1}{\tau_k}\mathbf{I})d\mu_k\right]
\end{aligned}
\tag{1}
$$

First looking at the integral inside the expectation:

$$
\begin{aligned}
&\int_{\mu_k}N(\mu_k;\nu_k,\mathbf{I})\log N(x_i;\mu_k,\frac{1}{\tau_k}\mathbf{I})d\mu_k \\
&= \int_{\mu_k}(2\pi)^{\frac{D}{2}}\exp(-\frac{1}{2}||\mu_k-\nu_k||^2)\Big(-\frac{D}{2}\log\Big(\frac{2\pi}{\tau_k}\Big)-\frac{\tau_k}{2}||x_i-\mu_k||^2\Big)d\mu_k \\
&= -\frac{D}{2}\log\Big(\frac{2\pi}{\tau_k}\Big)+\int_{\mu_k}(2\pi)^{\frac{D}{2}}\exp(-\frac{1}{2}||\mu_k-\nu_k||^2)\Big(-\frac{\tau_k}{2}(x_i^2-2\mu_k x_i+\mu_k^2)\Big)d\mu_k
\end{aligned}
$$

To compute this integral, we use a change of variables $m=\mu_k-\nu_k$ and the equation $\int_0^\infty m^n e^{-am^2}dm=\frac{\Gamma(\frac{n+1}{2})}{2a^{\frac{n+1}{2}}}$:

$$
\begin{aligned}
&\int_{\mu_k}(2\pi)^{\frac{D}{2}}\exp(-\frac{1}{2}||\mu_k-\nu_k||^2)\Big(-\frac{\tau_k}{2}(x_i^2-2\mu_k x_i+\mu_k^2)\Big)d\mu_k \\
&= \int_{m}(2\pi)^{\frac{D}{2}}\exp(-\frac{1}{2}m^2)\Big(-\frac{\tau_k}{2}(||x_i-\nu_k||^2+(2\nu_k-2x_i)m+m^2\Big)dm \\
&= -\frac{\tau_k}{2}\Big(||x_i-\nu_k||^2+\frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}}0.5^{1.5}}\Big)
\end{aligned}
$$

The term linear in m equals 0 because it is an odd function. Overall,

$$
\begin{aligned}
&\int_{\mu_k}N(\mu_k;\nu_k,\mathbf{I})\log N(x_i;\mu_k,\frac{1}{\tau_k}\mathbf{I})d\mu_k \\
&= -\frac{D}{2}\log(\frac{2\pi}{\tau_k})-\frac{\tau_k}{2}\Big(||x_i-\nu_k||^2+\frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}}0.5^{1.5}}\Big)
\end{aligned}
\tag{2}
$$

Now, substituting (2) into (1) and taking the expectation with respect to $Q(\tau_k)$:

$$\sum_{k=1}^{T} \zeta_{i,k} \mathop{\mathbb{E}}_{\tau_k \sim Q}\Big[\mathop{\mathbb{E}}_{\mu_k \sim Q}[\log P(x_i|\mu_k, \tau_k)]\Big]$$
$$= \sum_{k=1}^{T} \zeta_{i,k} \mathop{\mathbb{E}}_{\tau_k \sim Q}\Big[-\frac{D}{2}\log(\frac{2\pi}{\tau_k}) - \frac{\tau_k}{2}\Big(||x_i - \nu_k||^2 + \frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}} 0.5^{1.5}}\Big)\Big]$$
$$= \sum_{k=1}^{T} \zeta_{i,k} \Bigg(-\frac{D}{2}(\log(2\pi) - \mathop{\mathbb{E}}_{\tau_k \sim Q}[\log(\tau_k)]) - \frac{\mathbb{E}_{\tau_k \sim Q}[\tau_k]}{2}\Big(||x_i - \nu_k||^2 + \frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}} 0.5^{1.5}}\Big)\Bigg)$$

For $G \sim Gamma(g_1, g_2)$ with a rate parametrization, $\mathbb{E}[G] = \frac{g_1}{g_2}$ and $\mathbb{E}[\log(G)] = \Psi(g_1) - \log(g_2)$. Using these gives the stated bound.

3 Variational updates

3.1 λ updates

Blei and Jordan present the updates for the cluster probabilities in equations (18) and (19).

$$\lambda_{k,1} = 1 + \sum_i \zeta_{i,k}$$
$$\lambda_{k,2} = \alpha + \sum_i \sum_{j>k} \zeta_{i,j}$$

3.2 ν update

$$\nu_k = \frac{\sum_{i=1}^{N} \frac{\zeta_{i,k} a_k}{b_k} x_i}{1 + \sum_{i=1}^{N} \frac{\zeta_{i,k} a_k}{b_k}}$$

3.2.1 Derivation

Taking the derivative of the evidence lower bound with respect to ν_k gives:

$$\frac{\delta L}{\delta \nu_k} = -\nu_k + \sum_{i=1}^{N} \frac{\zeta_i a_k}{b_k}(x_i - \nu_k)$$

Setting this to zero and solving for ν_k gives the update.

3.3 a and b updates

$$
\begin{aligned}
a_k &= 1 + \frac{D}{2}\sum_{i=1}^{N}\zeta_{i,k} \\
b_k &= 1 + \frac{1}{2}\sum_{i=1}^{N}\zeta_{i,k}\Big(||x_i - \nu_k||^2 + \frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}}0.5^{1.5}}\Big)
\end{aligned}
$$

3.3.1 Derivation

Rather than maximising the ELBO with respect to a and b directly (which does not yield a closed-form derivative), we instead maximise the ELBO directly with respect to Q_τ and show that the result takes the functional form of a Gamma distribution for some particular a and b.

Using the calculus of variations, the optimal Q_τ can be shown to be

$$
Q(\tau_{k^*}) \propto \exp\Big(\mathop{\mathbb{E}}_{\boldsymbol{\phi},\mathbf{z},\boldsymbol{\mu},\boldsymbol{\tau}_{-k^*}\sim Q} P(\boldsymbol{\phi},\mathbf{z},\boldsymbol{\mu},\boldsymbol{\tau},X)\Big)
$$

Collecting all constant factors with respect to τ_{k^*}, we get:

$$
\begin{aligned}
\log Q(\tau_{k^*}) &= \mathop{\mathbb{E}}_{\mathbf{z},\boldsymbol{\mu},\boldsymbol{\tau}_{-k^*}\sim Q}[\log(P(\tau_{k^*}) + \log(P(X|z,\mu,\tau_{-k^*},\tau_{k^*})] + \text{Const.} \\
&= \mathop{\mathbb{E}}_{\mathbf{z},\boldsymbol{\mu},\boldsymbol{\tau}_{-k^*}\sim Q}[\log Gamma(\tau_{k^*};1,1) + \sum_{i=1}^{N}\log(N(x_i;\mu_{z_i},\frac{1}{\tau_{z_i}}\mathbf{I}))] + \text{Const.} \\
&= \log Gamma(\tau_{k^*};1,1) + \sum_{i=1}^{N}\sum_{k=1}^{T}\zeta_{i,k}\mathop{\mathbb{E}}_{\boldsymbol{\mu},\boldsymbol{\tau}_{-k^*}\sim Q}[\log N(x_i;\mu_{z_i},\frac{1}{\tau_{z_i}}\mathbf{I}] + \text{Const.} \\
&= \log Gamma(\tau_{k^*};1,1) + \sum_{i=1}^{N}\zeta_{i,k^*}\mathop{\mathbb{E}}_{\mu\sim Q}[\log N(x_i;\mu_{k^*},\frac{1}{\tau_{k^*}}\mathbf{I}] + \text{Const.}
\end{aligned}
$$

Substituting in equation (2):

$$
\begin{aligned}
\log Q(\tau_{k^*}) = \log Gamma(\tau_{k^*}; 1, 1) \\
&+ \sum_{i=1}^{N} \zeta_{i,k^*} \Bigg(-\frac{D}{2} \log\Big(\frac{2\pi}{\tau_{k^*}}\Big) - \frac{\tau_{k^*}}{2}\Big(||x_i - \nu_{k^*}||^2 + \frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}} 0.5^{1.5}} \Big)\Bigg) \\
&+ \text{Const.} \\
= -\tau_{k^*} - \tau_{k^*} \sum_{i=1}^{N} \frac{\zeta_{i,k^*}}{2}\Big(||x_i - \nu_{k^*}||^2 - \frac{\Gamma(1.5)}{(2\pi)^{\frac{D}{2}} (0.5)^{1.5}} \Big) \\
&+ \log(\tau_{k^*}) \sum_{i=1}^{N} \frac{\zeta_{i,k*} D}{2} + \text{Const.}
\end{aligned}
$$

Gathering up the terms containing τ_k and $\log(\tau_k)$, one can show that this is the expression for the log of a Gamma distribution with shape parameter a and rate parameter b equal to the updates above.

3.4 ζ update

Using the η terms defined in the sections on the z and x terms of the ELBO, which are constant with respect to ζ:

$$
\zeta_{i,k} = \frac{\exp(\eta_{z_{i,k}} + \eta_{x_{i,k}} - 1)}{\sum_{j=1}^{T} \exp(\eta_{z_{i,j}} + \eta_{x_{i,j}} - 1)}
$$

3.4.1 Derivation

To maximize the ELBO with respect to $\zeta_{\mathbf{i}}$ subject to $\sum_k \zeta_{i,k} = 1$, we use Lagrange multipliers. (The λ below is not the variational parameter.)

$$
\begin{aligned}
\mathcal{L}(\zeta_{\mathbf{i}}, \lambda) &= \Bigg(\sum_{j=1}^{T} \zeta_{i,j} \Big[-\log(\zeta_{i,j}) + \eta_{z_{i,j}} \Big] + \sum_{j=1}^{T} \zeta_{i,j} \eta_{x_{i,j}} \Bigg) - \lambda(\sum_k \zeta_{i,k} - 1) \\
\frac{\delta\, \mathcal{L}(\zeta_{\mathbf{i}}, \lambda)}{\delta \zeta_{i,k}} &= \Bigg(-1 - \log(\zeta_{i,k}) + \eta_{z_{i,k}} + \eta_{x_{i,k}} \Bigg) - \lambda \\
\frac{\delta\, \mathcal{L}(\zeta_{\mathbf{i}}, \lambda)}{\delta \lambda} &= 1 - \sum_{j=1}^{T} \zeta_{i,j}
\end{aligned}
$$

Setting these partial derivatives to 0 and solving for $\zeta_{i,k}$ gives the stated update.