Debugging VB

How do we know when VB has converged? How do we know that our answer is right?

One check: VB lower bound

$$\mathcal{L} = \sum_Z \iiint q(Z, \pi, \mu, \Lambda) \ln \left\{ \frac{p(X, Z, \pi, \mu, \Lambda)}{q(Z, \pi, \mu, \Lambda)} \right\} d\pi\, d\mu\, d\Lambda$$

when this stops going up we can declare our inference algorithm to have converged. Dropping the $*$'s and Expectation subscripts we have

$$= \mathbb{E}[\ln p(X, Z, \pi, \mu, \Lambda)] - \mathbb{E}[\ln q(Z, \pi, \mu, \Lambda)]$$

(2) $$= \mathbb{E}[\ln p(X|Z, \mu, \Lambda)] + \mathbb{E}[\ln p(Z|\pi)] + \mathbb{E}[\ln p(\pi)] + \mathbb{E}[\ln p(\mu, \Lambda)] - \underbrace{\mathbb{E}[\ln q(Z)] - \mathbb{E}[\ln q(\pi)] - \mathbb{E}[\ln q(\mu, \Lambda)]}$$

Note these are all entropy terms that can be looked up

We avoided doing these expectations before, we'll do one here just to show / highlight the required techniques.

Foremost is a conditional expectation trick. When $p(a,b)$ naturally factorizes as $p(b|a)\,p(a) = p(a,b)$ - or is specified as such - we can perform expectations in a step wise manner

$$\boxed{\mathbb{E}_{a,b}[f(x,a,b)] = \mathbb{E}_a\left[\mathbb{E}_{b|a}[f(x,a,b)]\right]}$$

pf.
$$\begin{aligned}
&= \mathbb{E}_a\left[\sum_b f(x,a,b)\, p(b|a)\right] \\
&= \sum_a \left(\sum_b f(x,a,b)\, p(b|a)\right) p(a) \\
&= \sum_a \sum_b f(x,a,b)\, p(a,b) \quad \blacksquare
\end{aligned}$$

all w.r.t. q^* dists

Let's work on first term

$$\mathbb{E}[\ln p(X|Z,\mu,\Lambda)] = \mathbb{E}_{\Lambda}\Big[\mathbb{E}_{\mu|\Lambda}\Big[\mathbb{E}_{Z}\Big[\ln \prod_n \prod_k \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}\Big]\Big]\Big]$$

applying the log & taking the Expectation w.r.t. $q^*(z)$ gives

$$= \mathbb{E}_{\Lambda}\Big[\mathbb{E}_{\mu|\Lambda}\Big[\sum_n \sum_k \mathbb{E}[z_{nk}] \cdot \ln \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})\Big]\Big]$$

we know this, $\mathbb{E}_{q(z)}[z_{nk}] = r_{nk}$

pushing the Expectations into the sums & taking the log of the Normal

$$= \sum_n \sum_k r_{nk} \, \mathbb{E}_{\Lambda}\Big[\mathbb{E}_{\mu|\Lambda}\Big[-\tfrac{1}{2}(x_n-\mu_k)^T \Lambda_k (x_n-\mu_k) - \tfrac{D}{2}\ln(2\pi) + \tfrac{1}{2}\ln|\Lambda_k|\Big]\Big]$$

(1) $$= \sum_n \sum_k r_{nk} \Big(\mathbb{E}_{\Lambda}\Big[\mathbb{E}_{\mu|\Lambda}\Big[-\tfrac{1}{2}(x_n-\mu_k)^T \Lambda_k (x_n-\mu_k)\Big]\Big] - \underbrace{\tfrac{D}{2}\ln(2\pi)}_{\text{this term constant}} + \tfrac{1}{2}\underbrace{\mathbb{E}_{\Lambda}\big[\ln|\Lambda_k|\big]}\Big)$$

this term takes a little work

this term can be looked up for Wishart distributions B.81 pg 693 PRML

Let's generalize this in the following way

Let $\mu \sim \mathcal{N}(a, (\Sigma b)^{-1})$ and $\Sigma \sim W(\Psi, v)$

What is $\mathbb{E}_{\Sigma}\big[\mathbb{E}_{\mu|\Sigma}[(x-\mu)^T \Sigma (x-\mu)]\big]$?

this is a case where conditional expectation from previous page helps

Nice trick - recast problem slightly by adding and subtracting mean of μ, i.e.

What is $\mathbb{E}_{\Sigma,\mu}\big[(x-a+a-\mu)^T \Sigma (x-a+a-\mu)\big]$?

This can be expanded like

$$= \mathbb{E}_\Sigma\Big[\mathbb{E}_{\mu|\Sigma}\Big[(x-a)^T\Sigma(x-a) + (x-a)^T\Sigma(a-\mu) + (a-\mu)^T\Sigma(x-a) + (a-\mu)^T\Sigma(a-\mu)\Big]\Big]$$

which has some nice properties, namely

$$= \mathbb{E}_\Sigma\Big[(x-a)^T\Sigma(x-a) + \cancel{(x-a)^T\Sigma(a-\mathbb{E}[\mu])} \to 0 + \underbrace{\cancel{(a-\mathbb{E}(\mu))^T\Sigma(x-a)}}_{0} + \underbrace{\mathbb{E}_{\mu|\Sigma}\big[(a-\mu)^T\Sigma(a-\mu)\big]}_{\text{almost a } \chi^2 \text{ RV}}\Big]$$

$$= \mathbb{E}_\Sigma\Big[(x-a)^T\Sigma(x-a) + \frac{1}{b}\underbrace{\mathbb{E}_{\mu|\Sigma}\big[(a-\mu)^T(b\Sigma)(a-\mu)\big]}\Big]$$

Now this is the expectation of a χ^2 RV. $\mathbb{E}[\psi]=D$ where $\psi \sim \chi^2_D$ and D is dimension of μ

$$= (x-a)^T\,\mathbb{E}_\Sigma[\Sigma]\,(x-a) + \frac{D}{b}$$

this is the mean of a Wishart distribution, here $\Psi\nu$

$$= \nu\,(x-a)^T\Psi(x-a) + \frac{D}{b}$$

In our case ~~the~~, plugging this in yields

$$\mathbb{E}_{\Lambda_k}\Big[\mathbb{E}_{\mu_k|\Lambda_k}\Big[-\frac{1}{2}(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)\Big]\Big]$$

$\Lambda_k \sim W(W_k,\nu_k)$

$\mu_k|\Lambda_k \sim \mathcal{N}(m_k,(\beta_k\Lambda_k)^{-1})$

$$= -\frac{1}{2}\Big[(x_n-m_k)^T W_k(x_n-m_k)\nu_k + \frac{D}{\beta_k}\Big]$$

If we go back to ① and plug in everything we arrive at

$$\mathbb{E}[\ln p(x|z,\mu,\Lambda)] = \frac{1}{2}\sum_{k=1}^{K} N_k \Big\{ \ln \tilde{\Lambda}_k - D\beta_k^{-1} - v_k \mathrm{Tr}(S_k W_k)$$

$$- v_k (\bar{x}_k - m_k)^T W_k (\bar{x}_k - m_k) - D\cdot\ln(2\pi) \Big\}$$

where

$$\ln \tilde{\Lambda}_k \equiv \mathbb{E}[\ln|\Lambda_k|] = \sum_{i=1}^{D} \psi\left(\frac{v_k+1-i}{2}\right) + D\ln 2 + \ln|W_k|$$

and $\beta_k, v_k, S_k, W_k, \bar{x}_k, m_k, N_k$ are defined as before

If we go back to ② we see that we have only accounted for the first term in the sum of expectations. The rest are given on pg's 481-482 in PRML. For completeness the remaining terms are:

$$\mathbb{E}_{\pi,z}[\ln p(z|\pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \tilde{\pi}_k$$

where $\ln \tilde{\pi}_k = \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha})$

where, as before $\alpha_k = \alpha_0 + N_k$ and $\hat{\alpha} = K\alpha_0 + N$

$$\mathbb{E}_\pi[\ln p(\pi)] = \ln C(\alpha_0) + (\alpha_0 - 1)\sum_{k=1}^{K} \ln \tilde{\pi}_k$$

$$\mathbb{E}_{\mu,\Lambda}[\ln p(\mu,\Lambda)] = \frac{1}{2}\sum_{k=1}^{K} \Big\{ D\ln(\beta_0/2\pi) + \ln\tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k}$$

$$- \beta_0 v_k (m_k - m_0)^T W_k (m_k - m_0) \Big\}$$

$$+ K\ln B(W_0, v_0)$$

$$+ \frac{(v_0 - D - 1)}{2}\sum_{k=1}^{K} \ln\tilde{\Lambda}_k - \frac{1}{2}\sum_{k=1}^{K} v_k \mathrm{Tr}(W_0^{-1} W_k)$$

$$\mathbb{E}[\ln q(z)] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln r_{nk}$$

$$\mathbb{E}[\ln q(\pi)] = \sum_{k=1}^{K} (\alpha_k - 1) \ln \tilde{\pi}_k + \ln C(\alpha)$$

$$\mathbb{E}[\ln q(\mu, \Lambda)] = \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \tilde{\Lambda}_k + \frac{D}{2} \ln\left(\frac{\beta_k}{2\pi}\right) - \frac{D}{2} - H[q(\Lambda_k)] \right]$$

The H & B terms can be looked up in the PRML appendix.

Summary: the variational lower bound is a lower bound on the evidence of the data under the model.

a) This can be useful for model comparison
b) Is useful for debugging

Overall summary:

VB re-estimation equations can be cycled through, factor by factor. In optimizing each factor we arrive at parameters ~~###~~ of distributions necessary to optimize other factors.

<u>Other Uses of Variational approximation to the posterior distribution</u>:

<u>Prediction</u>

Let $\hat{x}$ be a new data point, what is the predictive distribution (posterior) for $\hat{x}$

$$p(\hat{x}|X) = \sum_{\hat{z}} \iiint p(\hat{x}|\hat{z},\mu,\Lambda)\, p(\hat{z}|\pi)\, \underbrace{p(\pi,\mu,\Lambda|X)}_{\text{posterior}}\, d\pi\, d\mu\, d\Lambda$$

Remember, $q(\theta) = p(\theta|X)$ is the approximate posterior distribution so we can plug it into expressions like this a derive an approximate posterior predictive

For instance

$$p(\hat{x} \mid X) = \sum_{k=1}^{K} \iiint \pi_k \mathcal{N}(\hat{x} \mid \mu_k, \Lambda_k^{-1}) q(\pi) q(\mu_k, \Lambda_k) d\pi d\mu_k d\Lambda_k$$

where several integrals have been performed implicitly to get to this step.

The remaining integrals can be performed analytically to arrive at a mixture of Student-t dist's.

$$p(\hat{x} \mid X) = \frac{1}{\hat{\alpha}} \sum_{k=1}^{K} \alpha_k \, St(\hat{x} \mid m_k, L_k, \nu_k + 1 - D)$$

where $L_k = \frac{(\nu_k + 1 - D)\beta_k}{(1+\beta_k)} W_k$

Last note: learning the # of components. The prior on π allows effective control over the "sparsity" of the model.

<u>Induced factorizations</u>

Factorization of approximating dist. factorized further as a consequence of the conditional independencies ~~implied~~ by embodied by the graphical model.

Computationally important to utilize these addition factorizations (space and time)