Variational Inference

is proportional to a normal density when considered as a function of α_j. We can identify the parameters of this normal distribution by completing the square of the quadratic expression or, more intuitively from a statistical perspective, recognizing the expression as equivalent to two pieces of information, one centered at y_j with inverse-variance σ_j^{-2} and one centered at $E(\mu)$ with inverse-variance $E(\frac{1}{\tau^2})$.

We combine these by weighting the means and adding the inverse-variances, thus getting the following form for the variational Bayes component for α_j:

$$g(\alpha_j) = \mathcal{N}\left(\alpha_j \,\middle|\, \frac{\frac{1}{\sigma_j^2}y_j + E\left(\frac{1}{\tau^2}\right)E(\mu)}{\frac{1}{\sigma_j^2} + E\left(\frac{1}{\tau^2}\right)},\ \frac{1}{\frac{1}{\sigma_j^2} + E\left(\frac{1}{\tau^2}\right)}\right). \tag{13.18}$$

For μ, we inspect (13.16). Averaging over all the parameters other than μ, the expression $E\log p(\theta \mid y)$ has the form

$$-\frac{1}{2}E\left(\frac{1}{\tau^2}\right)\sum_{j=1}^{8}(E(\alpha_j)-\mu)^2 + \text{const.}$$

As above, this is the logarithm of a normal density function; the parameters of this distribution can be determined by considering it as a combination of 8 pieces of information:

$$g(\mu) = \mathcal{N}\left(\mu \,\middle|\, \frac{1}{8}\sum_{j=1}^{8}E(\alpha_j),\ \frac{1}{8E\left(\frac{1}{\tau^2}\right)}\right). \tag{13.19}$$

Finally, averaging over all parameters other than τ gives a density function that can be recognized as inverse-gamma or, in the parameterization we prefer,

$$g(\tau^2) = \text{Inv-}\chi^2\left(\tau^2 \,\middle|\, 7,\ \frac{1}{7}\sum_{j=1}^{8}E\left((\alpha_j-\mu)^2\right)\right), \tag{13.20}$$

with the expectation $E\left((\alpha_j-\mu)^2\right)$ over the approximating distribution g.

The above expressions are essentially identical to the derivations of the conditional distributions for the Gibbs sampler for the hierarchical normal model in Section 11.6 and the EM algorithm in Section 13.6, with the only difference being that in the 8-schools example we assume the data variances σ_j are known.

Determining the conditional expectations

Rewriting the above factors in generic notation, we have:

$$g(\alpha_j) = \mathcal{N}(\alpha_j \mid M_{\alpha_j}, S_{\alpha_j}^2), \quad j = 1,\ldots,8, \tag{13.21}$$

$$g(\mu) = \mathcal{N}(\mu \mid M_\mu, S_\mu^2), \tag{13.22}$$

$$g(\tau^2) = \text{Inv-}\chi^2(\tau^2 \mid T, M_\tau^2). \tag{13.23}$$

We will need these to get the conditional expectations for each of the above three steps:

- To specify the distribution for α_j in (13.18), we need $E(\mu)$, which is M_μ from (13.22), and $E\left(\frac{1}{\tau^2}\right)$, which is $\frac{1}{M_\tau^2}$ from (13.23).
- To specify the distribution for μ in (13.19), we need $E(\alpha_j)$, which is M_{α_j} from (13.21), and $E\left(\frac{1}{\tau^2}\right)$, which is $\frac{1}{M_\tau^2}$ from (13.23).
- To specify the distribution for τ in (13.20), we need $E\left((\alpha_j-\mu)^2\right)$, which is $(M_{\alpha_j}-M_\mu)^2 + S_{\alpha_j}^2 + S_\mu^2$ from (13.21) and (13.22), and using the assumption that the densities g are independent in the variational approximation.