Chapter 21

Zhu Qian

21.5 Variational Bayes

- So far we have been concentrating on inferring latent variables z_i assuming the parameters θ of the model are known. Now suppose we want to infer the parameters themselves. If we make a fully factorized (i.e., mean field) approximation, $p(\theta|D) \approx \prod_k q(\theta|k)$, we get a method known as **variational Bayes or VB.**
- We give some examples of VB below, assuming that there are no latent variables.

21.5 Variational Bayes

- If we want to infer both latent variables and parameters, and we make an approximation of the form $p(\theta, z_{1:n}|D) \approx q(\theta) \prod_i q_i(z_i)$,we get a method known as variational Bayes EM, which we described in Section 21.6.

21.5.1 Example: VB for a univariate Gaussian

- Let us consider how to apply VB to infer the posterior over the parameters for a 1d Gaussian, $p(\mu, \lambda)$ where $\lambda = 1/\sigma^2$ is the precision.
- For convenience, we will use a conjugate prior of the form

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})\mathrm{Ga}(\lambda|a_0, b_0)$$

- However, we will use an approximate factored posterior of the form

$$q(\mu, \lambda) = q_\mu(\mu)q_\lambda(\lambda)$$

21.5.1.1 Target distribution

The unnormalized log posterior has the form

$$\begin{aligned}
\log \tilde{p}(\mu, \lambda) &= \log p(\mu, \lambda, \mathcal{D}) = \log p(\mathcal{D}|\mu, \lambda) + \log p(\mu|\lambda) + \log p(\lambda) \\
&= \frac{N}{2} \log \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} (x_i - \mu)^2 - \frac{\kappa_0 \lambda}{2} (\mu - \mu_0)^2 \\
&\quad + \frac{1}{2} \log(\kappa_0 \lambda) + (a_0 - 1) \log \lambda - b_0 \lambda + \text{const}
\end{aligned}$$

21.3 The mean field method

- In Section 21.3.1, we derive a coordinate descent method, where at each step we make the following update:

$$\log q_j(\mathbf{x}_j) = \mathbb{E}_{-q_j}\left[\log \tilde{p}(\mathbf{x})\right] + \text{const}$$

- Where $\tilde{p}(x) = p(x, D)$ is the unnormalized posterior and the notation $\mathrm{E}_{-q_j}[f(x)]$ means to take the expectation over $f(x)$with respect to all the variables except for x_j.

21.5.1.2 Updating $q_\mu(\mu)$

- The optimal form for $q_\mu(\mu)$ is obtained by averaging over λ

$$\begin{aligned}\log q_\mu(\mu) &= \mathbb{E}_{q_\lambda}\left[\log p(\mathcal{D}|\mu,\lambda) + \log p(\mu|\lambda)\right] + \text{const} \\ &= -\frac{\mathbb{E}_{q_\lambda}[\lambda]}{2}\left\{\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\right\} + \text{const}\end{aligned}$$

By completing the square one can show that $q_\mu(\mu) = N(\mu|\mu_N, \kappa_N^{-1})$, where

$$\mu_N = \frac{\kappa_0\mu_0 + N\overline{x}}{\kappa_0 + N}, \quad \kappa_N = (\kappa_0 + N)\mathbb{E}_{q_\lambda}[\lambda]$$

21.5.1.3 Updating $q_\lambda(\lambda)$

- The optimal form for $q_\lambda(\lambda)$ is given by

$$\begin{aligned}
\log q_\lambda(\lambda) &= \mathbb{E}_{q_\mu}\left[\log p(\mathcal{D}|\mu,\lambda) + \log p(\mu|\lambda) + \log p(\lambda)\right] + \text{const} \\
&= (a_0 - 1)\log\lambda - b_0\lambda + \frac{1}{2}\log\lambda + \frac{N}{2}\log\lambda \\
&\quad -\frac{\lambda}{2}\mathbb{E}_{q_\mu}\left[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\right] + \text{const}
\end{aligned}$$

- We recognize this as the log of a Gamma distribution, hence $q_\lambda(\lambda) = Ga(\lambda|a_N, b_N)$, where

$$\begin{aligned}
a_N &= a_0 + \frac{N+1}{2} \\
b_N &= b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\right]
\end{aligned}$$

21.5.1.4 Computing the expectations

- To implement the updates, we have to specify how to compute the various expectations. Since $q(\mu) = N(\mu|\mu_N, \kappa_N^{-1})$ we have

$$\begin{aligned} \mathbb{E}_{q(\mu)}[\mu] &= \mu_N \\ \mathbb{E}_{q(\mu)}[\mu^2] &= \frac{1}{\kappa_N} + \mu_N^2 \end{aligned}$$

- Since $q(\lambda) = \mathrm{Ga}(\lambda|a_N, b_N)$, we have

$$\mathbb{E}_{q(\lambda)}[\lambda] = \frac{a_N}{b_N}$$

21.5.1.4 Computing the expectations

- We can now give explicit forms for the update equations. For $q(\mu)$ we have

$$\begin{aligned} \mu_N &= \frac{\kappa_0 \mu_0 + N\overline{x}}{\kappa_0 + N} \\ \kappa_N &= (\kappa_0 + N)\frac{a_N}{b_N} \end{aligned}$$

- And for For $q(\lambda)$ we have

$$\begin{aligned} a_N &= a_0 + \frac{N+1}{2} \\ b_N &= b_0 + \kappa_0(\mathbb{E}\left[\mu^2\right] + \mu_0^2 - 2\mathbb{E}\left[\mu\right]\mu_0) + \frac{1}{2}\sum_{i=1}^{N}\left(x_i^2 + \mathbb{E}\left[\mu^2\right] - 2\mathbb{E}\left[\mu\right]x_i\right) \end{aligned}$$

21.5.1.5 Illustration

- Figure gives an example of this method in action. The green contours represent the exact posterior, which is Gaussian-Gamma. The dotted red contours represent the variational approximation over several iterations. We see that the final approximation is reasonably close to the exact solution. However, it is more “compact” than the true distribution. It is often the case that mean field inference underestimates the posterior uncertainty;

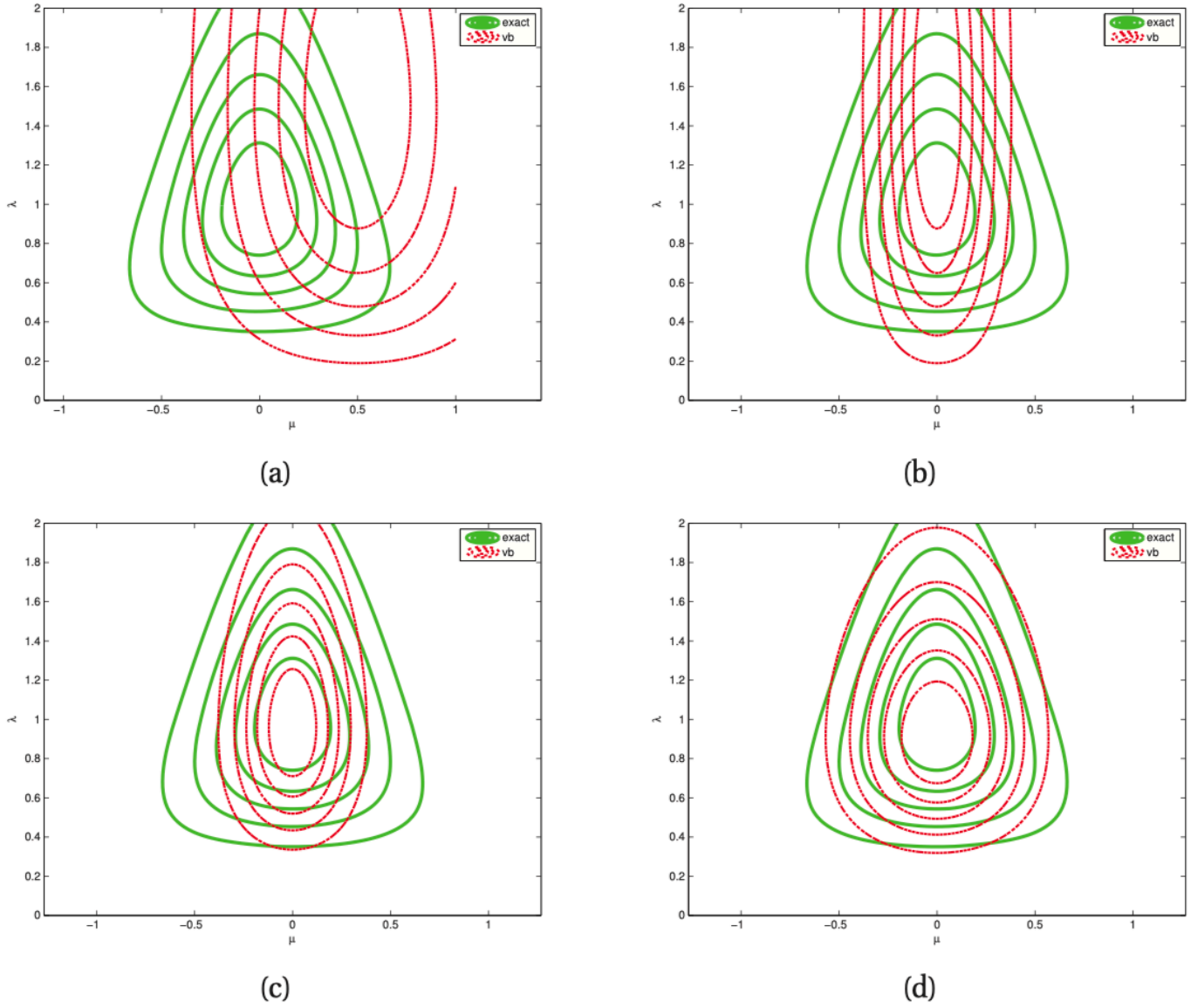


Figure 21.5 Factored variational approximation (red) to the Gaussian-Gamma distribution (green). (a) Initial guess. (b) After updating q_μ. (c) After updating q_λ. (d) At convergence (after 5 iterations). Based on 10.4 of (Bishop 2006b). Figure generated by `unigaussVbDemo`.

21.5.1.6 Lower bound

- In VB, we are maximizing $L(q)$, which is a lower bound on the log marginal likelihood:

$$L(q) \leq \log p(\mathcal{D}) = \log \int \int p(\mathcal{D}|\mu, \lambda) p(\mu, \lambda) d\mu d\lambda$$

- It is very useful to compute the lower bound itself, for three reasons. First, it can be used to assess convergence of the algorithm. Second, it can be used to assess the correctness of one's code: as with EM, if the bound does not increase monotonically, there must be a bug. Third, the bound can be used as an approximation to the marginal likelihood, which can be used for Bayesian model selection.

21.5.1.6 Lower bound

- For this model, L(q) can be computed as follows:

$$\begin{aligned} L(q) &= \int\int q(\mu,\lambda)\log\frac{p(\mathcal{D},\mu,\lambda)}{q(\mu,\lambda)}d\mu d\lambda \\ &= \mathbb{E}\left[\log p(\mathcal{D}|\mu,\lambda)\right] + \mathbb{E}\left[\log p(\mu|\lambda)\right] + \mathbb{E}\left[\log p(\lambda)\right] \\ &\quad -\mathbb{E}\left[\log q(\mu)\right] - \mathbb{E}\left[\log q(\lambda)\right] \end{aligned}$$

- where all expectations are wrt $q(\mu,\lambda)$. We recognize the last two terms as the entropy of a Gaussian and the entropy of a Gamma distribution, which are given by

$$\begin{aligned} \mathbb{H}\left(\mathcal{N}(\mu_N,\kappa_N^{-1})\right) &= -\frac{1}{2}\log\kappa_N + \frac{1}{2}(1+\log(2\pi)) \\ \mathbb{H}\left(\mathrm{Ga}(a_N,b_N)\right) &= \log\Gamma(a_N) - (a_N-1)\psi(a_N) - \log(b_N) + a_N \end{aligned}$$

- where $\psi()$ is the digamma function.

21.5.1.6 Lower bound

- To compute the other terms, we need the following facts:

$$\begin{aligned}
\mathbb{E}\left[\log x | x \sim \mathrm{Ga}(a,b)\right] &= \psi(a) - \log(b) \\
\mathbb{E}\left[x | x \sim \mathrm{Ga}(a,b)\right] &= \frac{a}{b} \\
\mathbb{E}\left[x | x \sim \mathcal{N}(\mu,\sigma^2)\right] &= \mu \\
\mathbb{E}\left[x^2 | x \sim \mathcal{N}(\mu,\sigma^2)\right] &= \mu + \sigma^2
\end{aligned}$$

- For the expected log likelihood, one can show that

$$\begin{aligned}
&\mathbb{E}_{q(\mu,\lambda)}\left[\log p(\mathcal{D}|\mu,\lambda)\right] \\
&= -\frac{N}{2}\log(2\pi) + \frac{N}{2}\mathbb{E}_{q(\lambda)}\left[\log\lambda\right] - \frac{\mathbb{E}\left[\lambda\right]_{q(\lambda)}}{2}\sum_{i=1}^{N}\mathbb{E}_{q(\mu)}\left[(x_i-\mu)^2\right] \\
&= -\frac{N}{2}\log(2\pi) + \frac{N}{2}\left(\psi(a_N) - \log b_N\right) \\
&\quad -\frac{N a_N}{2b_N}\left(\hat{\sigma}^2 + \overline{x}^2 - 2\mu_N\overline{x} + \mu_N^2 + \frac{1}{\kappa_N}\right)
\end{aligned}$$

21.5.1.6 Lower bound

- For the expected log prior of λ, we have

$$\begin{aligned}\mathbb{E}_{q(\lambda)}\left[\log p(\lambda)\right] &= (a_0-1)\mathbb{E}\left[\log\lambda\right] - b_0\mathbb{E}\left[\lambda\right] + a_0\log b_0 - \log\Gamma(a_0) \\ &= (a_0-1)(\psi(a_N)-\log b_N) - b_0\frac{a_N}{b_N} + a_0\log b_0 - \log\Gamma(a_0)\end{aligned}$$

- For the expected log prior of μ, one can show that

$$\begin{aligned}\mathbb{E}_{q(\mu,\lambda)}\left[\log p(\mu|\lambda)\right] &= \frac{1}{2}\log\frac{\kappa_0}{2\pi} + \frac{1}{2}\mathbb{E}\left[\log\lambda\right]q(\lambda) - \frac{1}{2}\mathbb{E}_{q(\mu,\lambda)}\left[(\mu-\mu_0)^2\kappa_0\lambda\right] \\ &= \frac{1}{2}\log\frac{\kappa_0}{2\pi} + \frac{1}{2}\left(\psi(a_N)-\log b_N\right) \\ &\quad -\frac{\kappa_0}{2}\frac{a_N}{b_N}\left[\frac{1}{\kappa_N} + (\mu_N-\mu_0)^2\right]\end{aligned}$$

21.5.1.6 Lower bound

- Putting it altogether, one can show that

$$L(q) = \frac{1}{2}\log\frac{1}{\kappa_N} + \log\Gamma(a_N) - a_N \log b_N + \text{const}$$

- This quantity monotonically increases after each VB update.

21.6 Variational Bayes EM

- Now consider latent variable models of the form $z_i \rightarrow x_i \leftarrow \theta$. This includes mixtures models, PCA, HMMs, etc. There are now two kinds of unknowns: parameters, θ, and latent variables, z_i. As we saw in Section 11.4, it is common to fit such models using EM, where in the E step we infer the posterior over the latent variables, $p(z_i|x_i, \theta)$, and in the M step, we compute a point estimate of the parameters, θ. The justification for this is two-fold. First, it results in simple algorithms. Second, the posterior uncertainty in θ is usually less than in z_i, since the θ are informed by all N data cases, whereas z_i is only informed by x_i ; this makes a MAP estimate of θ more reasonable than a MAP estimate of z_i.

21.6 Variational Bayes EM

- However, VB provides a way to be "more Bayesian", by modeling uncertainty in the parameters θ as well in the latent variables z_i, at a computational cost that is essentially the same as EM. This method is known as variational Bayes EM or VBEM. The basic idea is to use mean field, where the approximate posterior has the form

$$p(\boldsymbol{\theta}, \mathbf{z}_{1:N}|\mathcal{D}) \approx q(\boldsymbol{\theta})q(\mathbf{z}) = q(\boldsymbol{\theta}) \prod_i q(\mathbf{z}_i)$$

- The first factorization, between θ and z, is a crucial assumption to make the algorithm tractable. The second factorization follows from the model, since the latent variables are iid conditional on θ.

21.6 Variational Bayes EM

- In VBEM, we alternate between updating $q(z_i|D)$(the variational E step) and updating $q(\theta|D)$ (the variational M step).
- The variational E step is similar to a standard E step, except instead of plugging in a MAP estimate of the parameters and computing $q\left(z_i\middle|D,\hat{\theta}\right)$, we need to average over the parameters.
- Roughly speaking, this can be computed by plugging in the posterior mean of the parameters instead of the MAP estimate, and then computing $q(z_i\middle|D,\bar{\theta}\ )$ using standard algorithms, such as forwards-backwards.

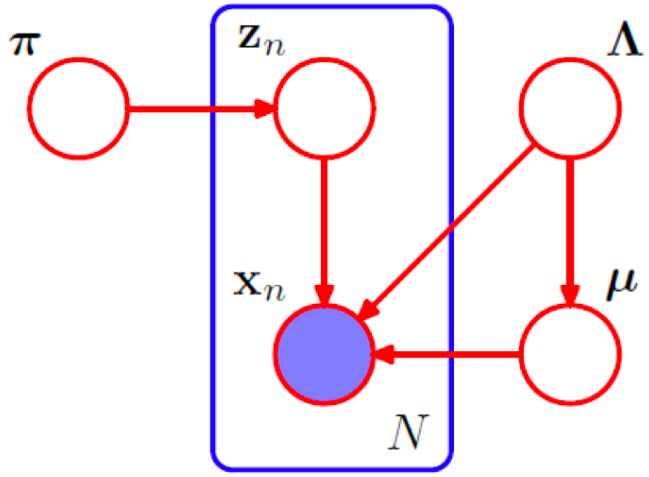
21.6 Variational Bayes EM

- The variational M step is similar to a standard M step, except instead of computing a point estimate of the parameters, we update the hyper-parameters, using the expected sufficient statis-tics. This process is usually very similar to MAP estimation in regular EM. Again, the details on how to do this depend on the form of the model.
- The principle advantage of VBEM over regular EM is that by marginalizing out the parameters, we can compute a lower bound on the marginal likelihood, which can be used for model selection. We will see an example of this in Section 21.6.1.6. VBEM is also "egalitarian", since it treats parameters as "first class citizens", just like any other unknown quantity, whereas EM makes an artificial distinction between parameters and latent variables.

21.6.1 Example: VBEM for mixtures of Gaussians

- The likelihood function is the usual one for Gaussian mixture models:

$$p(\mathbf{z}, \mathbf{X}|\boldsymbol{\theta}) = \prod_i \prod_k \pi_k^{z_{ik}} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{ik}}$$



where $z_{ik} = 1$ if data point i belongs to cluster k, and $z_{ik} = 0$ otherwise. We will assume the following factored conjugate prior

$$p(\boldsymbol{\theta}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) \prod_k \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \text{Wi}(\boldsymbol{\Lambda}_k|\mathbf{L}_0, \nu_0)$$

where Λ_k is the precision matrix for cluster k. The subscript 0 means these are parameters of the prior; we assume all the prior parameters are the same for all clusters.

21.6.1 Example: VBEM for mixtures of Gaussians

- We will try to approximate the volume around one of these modes.

$$p(\boldsymbol{\theta}, \mathbf{z}_{1:N}|\mathcal{D}) \approx q(\boldsymbol{\theta}) \prod_i q(\mathbf{z}_i)$$

- At this stage we have not specified the forms of the q functions; these will be determined by the form of the likelihood and prior. Below we will show that the optimal form is as follows:

$$\begin{aligned} q(\mathbf{z}, \boldsymbol{\theta}) &= q(\mathbf{z}|\boldsymbol{\theta})q(\boldsymbol{\theta}) = \left[\prod_i \mathrm{Cat}(\mathbf{z}_i|\mathbf{r}_i)\right] \\ &\left[\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_k \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathrm{Wi}(\boldsymbol{\Lambda}_k|\mathbf{L}_k, \nu_k)\right] \end{aligned}$$

21.6.1.2 Derivation of $q(z)$(variational E step)

- The form for $q(z)$ can be obtained by looking at the complete data log joint, ignoring terms that do not involve z, and taking expectations of what's left over wrt all the hidden variables except for z. We have

$$\begin{aligned} \log q(\mathbf{z}) &= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})\right] + \text{const} \\ &= \sum_i \sum_i z_{ik} \log \rho_{ik} + \text{const} \end{aligned}$$

- where we define

$$\begin{aligned} \log \rho_{ik} \triangleq\ & \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log \pi_k\right] + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log |\boldsymbol{\Lambda}_k|\right] - \frac{D}{2}\log(2\pi) \\ & -\frac{1}{2}\mathbb{E}_{q(\boldsymbol{\theta})}\left[(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\mathbf{x}_i - \boldsymbol{\mu}_k)\right] \end{aligned}$$

21.6.1.2 Derivation of $q(z)$(variational E step)

- Using the fact that q(π) = Dir(π), we have

$$\log \tilde{\pi}_k \triangleq \mathbb{E}\left[\log \pi_k\right] = \psi(\alpha_k) - \psi(\sum_{k'} \alpha_{k'})$$

Where $\psi()$ is the digamma function.

21.6.1.2 Derivation of $q(z)$(variational E step)

- Next, we use the fact that

$$q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathrm{Wi}(\boldsymbol{\Lambda}_k | \mathbf{L}_k, \nu_k)$$

to get

$$\log \tilde{\Lambda}_k \triangleq \mathbb{E}\left[\log |\boldsymbol{\Lambda}_k|\right] = \sum_{j=1}^{D} \psi\left(\frac{\nu_k + 1 - j}{2}\right) + D \log 2 + \log |\boldsymbol{\Lambda}_k|$$

- Finally, for the expected value of the quadratic form, we get

$$\mathbb{E}\left[(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\mathbf{x}_i - \boldsymbol{\mu}_k)\right] = D\beta_k^{-1} + \nu_k (\mathbf{x}_i - \mathbf{m}_k)^T \boldsymbol{\Lambda}_k (\mathbf{x}_i - \mathbf{m}_k)$$

21.6.1.2 Derivation of $q(z)$(variational E step)

- Putting it altogether, we get that the posterior responsibility of cluster *k* for datapoint *i* is

$$r_{ik} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{\frac{1}{2}} \exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\mathbf{x}_i - \mathbf{m}_k)^T \mathbf{\Lambda}_k (\mathbf{x}_i - \mathbf{m}_k)\right)$$

Compare this to the expression used in regular EM:

$$r_{ik}^{EM} \propto \hat{\pi}_k |\hat{\mathbf{\Lambda}}|_k^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T \hat{\mathbf{\Lambda}}_k (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)\right)$$

- The significance of this difference is discussed further in Section 21.6.1.7.

21.6.1.3 Derivation of q(θ) (variational M step)

- Using the mean field recipe, we have

$$\begin{aligned}\log q(\boldsymbol{\theta}) &= \log p(\boldsymbol{\pi}) + \sum_k \log p(\mu_k, \boldsymbol{\Lambda}_k) + \sum_i \mathbb{E}_{q(\mathbf{z})}\left[\log p(\mathbf{z}_i|\boldsymbol{\pi})\right] \\ &\quad + \sum_k \sum_i \mathbb{E}_{q(\mathbf{z})}\left[z_{ik}\right] \log \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + \text{const}\end{aligned}$$

- We see this factorizes into the form

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\pi}) \prod_k q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$$

21.6.1.3 Derivation of q(θ) (variational M step)

- For the π term, we have

$$\log q(\boldsymbol{\pi}) = (\alpha_0 - 1) \sum_k \log \pi_k + \sum_k \sum_i r_{ik} \log \pi_k + \text{const}$$

- Exponentiating, we recognize this as a Dirichlet distribution:

$$\begin{aligned} q(\boldsymbol{\pi}) &= \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \\ \alpha_k &= \alpha_0 + N_k \\ N_k &= \sum_i r_{ik} \end{aligned}$$

21.6.1.3 Derivation of q(θ) (variational M step)

- For the μ_k and Λ_k terms, we have

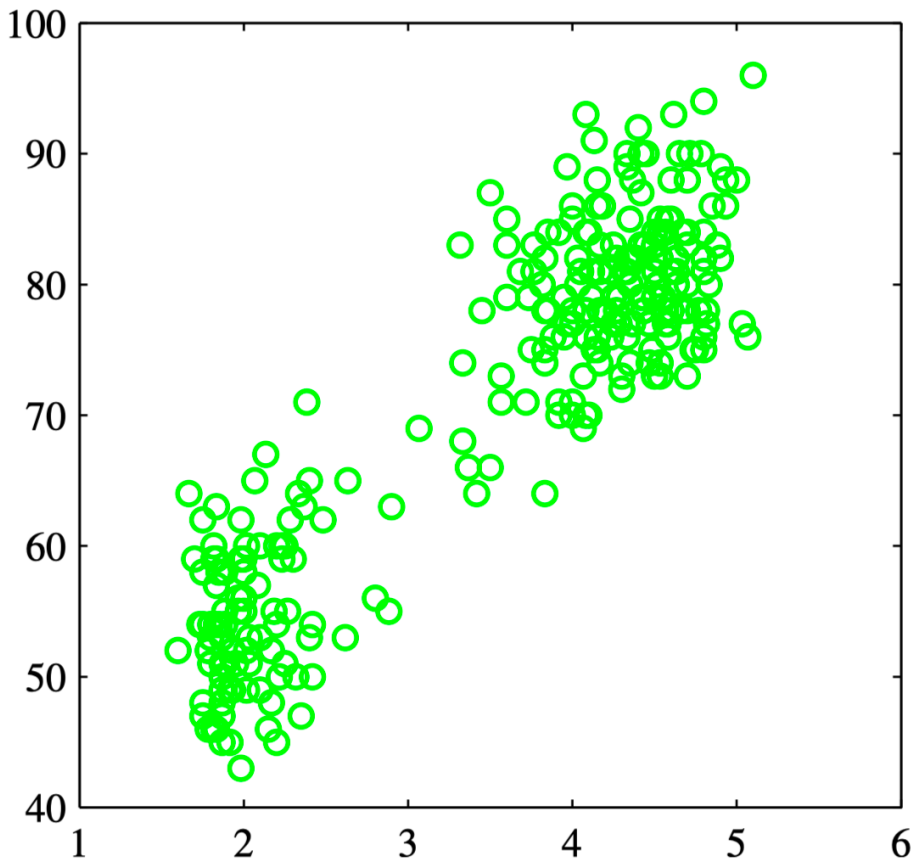
$$
\begin{aligned}
q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathrm{Wi}(\boldsymbol{\Lambda}_k | \mathbf{L}_k, \nu_k) \\
\beta_k &= \beta_0 + N_k \\
\mathbf{m}_k &= (\beta_0 \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k)/\beta_k \\
\mathbf{L}_k^{-1} &= \mathbf{L}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^T \\
\nu_k &= \nu_0 + N_k + 1 \\
\overline{\mathbf{x}}_k &= \frac{1}{N_k} \sum_i r_{ik} \mathbf{x}_i \\
\mathbf{S}_k &= \frac{1}{N_k} \sum_i r_{ik} (\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T
\end{aligned}
$$

Example: Old Faithful dataset

- Old Faithful, shown in Figure A.4, is a hydrothermal geyser in Yellowstone National Park in the state of Wyoming, U.S.A., and is a popular tourist attraction. Its name stems from the supposed regularity of its eruptions.
- The data set comprises 272 observations, each of which represents a single eruption and contains two variables corresponding to the duration in minutes of the erup-tion, and the time until the next eruption, also in minutes. Figure A.5 shows a plot of the time to the next eruption versus the duration of the eruptions. It can be seen that the time to the next eruption varies considerably, although knowledge of the duration of the current eruption allows it to be predicted more accurately. Note that there exist several other data sets relating to the eruptions of Old Faithful.

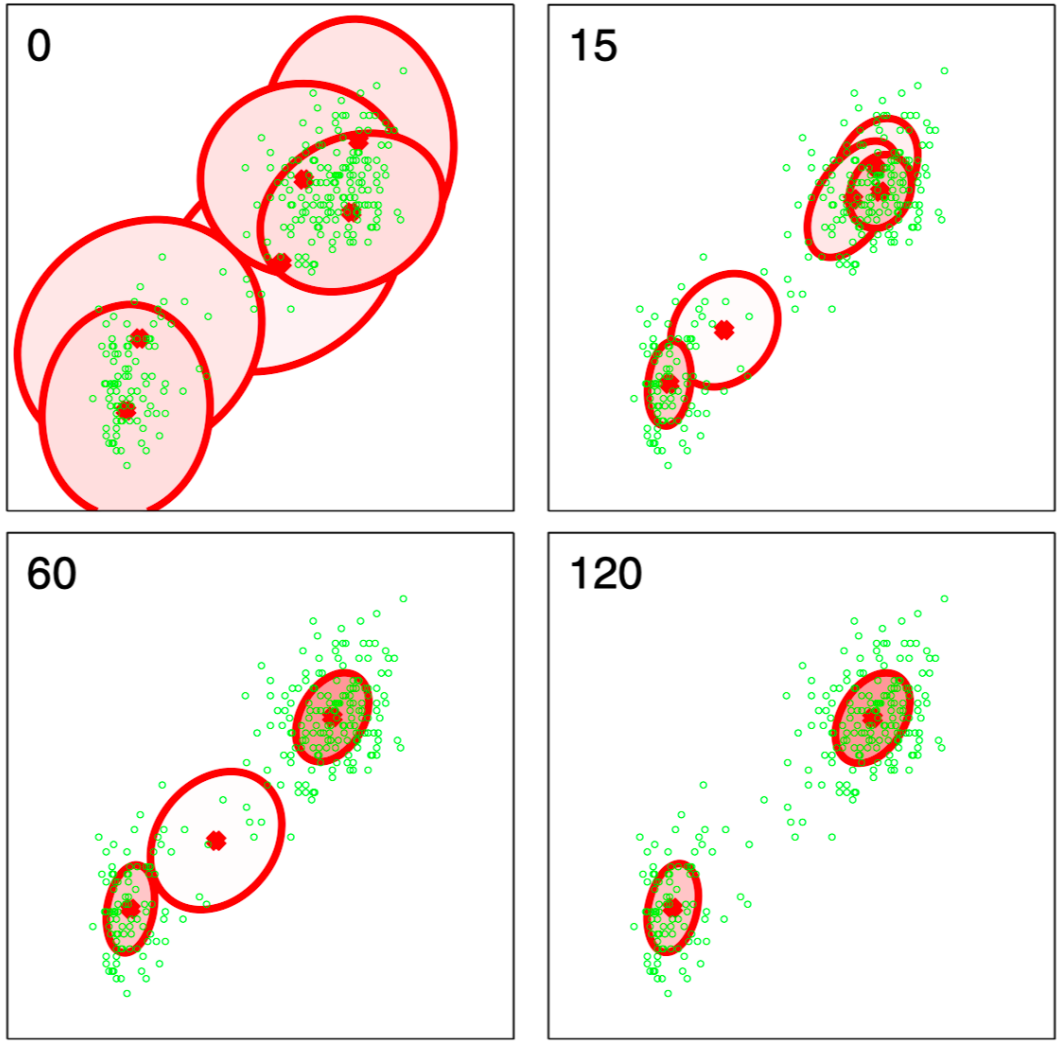
Example: Old Faithful dataset

Figure A.5 Plot of the time to the next eruption in minutes (vertical axis) versus the duration of the eruption in minutes (horizontal axis) for the Old Faithful data set.



Example: Old Faithful dataset

Figure 10.6 Variational Bayesian mixture of $K = 6$ Gaussians applied to the Old Faithful data set, in which the ellipses denote the one standard-deviation density contours for each of the components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component. The number in the top left of each diagram shows the number of iterations of variational inference. Components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted.



Example: Old Faithful dataset

- We see that after convergence, there are only two components for which the expected values of the mixing coefficients are numerically distinguishable from their prior values. This effect can be understood qualitatively in terms of the automatic trade-off in a Bayesian model between fitting the data and the complexity of the model, in which the complexity penalty arises from components whose parameters are pushed away from their prior values.
- Components that take essentially no responsibility for explaining the data points have $r_{nk} \cong 0$ and hence $N_k \cong 0$.