WIKIPEDIA
The Free Encyclopedia
WIKIPEDIA

Variational Bayesian methods

Variational Bayesian methods are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning. They are typically used in complex statistical models consisting of observed variables (usually termed "data") as well as unknown parameters and latent variables, with various sorts of relationships among the three types of random variables, as might be described by a graphical model. As typical in Bayesian inference, the parameters and latent variables are grouped together as "unobserved variables". Variational Bayesian methods are primarily used for two purposes:

1. To provide an analytical approximation to the posterior probability of the unobserved variables, in order to do statistical inference over these variables.
2. To derive a lower bound for the marginal likelihood (sometimes called the *evidence*) of the observed data (i.e. the marginal probability of the data given the model, with marginalization performed over unobserved variables). This is typically used for performing model selection, the general idea being that a higher marginal likelihood for a given model indicates a better fit of the data by that model and hence a greater probability that the model in question was the one that generated the data. (See also the Bayes factor article.)

In the former purpose (that of approximating a posterior probability), variational Bayes is an alternative to Monte Carlo sampling methods—particularly, Markov chain Monte Carlo methods such as Gibbs sampling—for taking a fully Bayesian approach to statistical inference over complex distributions that are difficult to evaluate directly or sample. In particular, whereas Monte Carlo techniques provide a numerical approximation to the exact posterior using a set of samples, variational Bayes provides a locally-optimal, exact analytical solution to an approximation of the posterior.

Variational Bayes can be seen as an extension of the expectation–maximization (EM) algorithm from maximum likelihood (ML) or maximum a posteriori (MAP) estimation of the single most probable value of each parameter to fully Bayesian estimation which computes (an approximation to) the entire posterior distribution of the parameters and latent variables. As in EM, it finds a set of optimal parameter values, and it has the same alternating structure as does EM, based on a set of interlocked (mutually dependent) equations that cannot be solved analytically.

For many applications, variational Bayes produces solutions of comparable accuracy to Gibbs sampling at greater speed. However, deriving the set of equations used to update the parameters iteratively often requires a large amount of work compared with deriving the comparable Gibbs sampling equations. This is the case even for many models that are conceptually quite simple, as is demonstrated below in the case of a basic non-hierarchical model with only two parameters and no latent variables.

Mathematical derivation

Problem

In variational inference, the posterior distribution over a set of unobserved variables $\mathbf{Z} = \{Z_1 \dots Z_n\}$ given some data $\mathbf{X}$ is approximated by a so-called **variational distribution**, $Q(\mathbf{Z})$:

$$P(\mathbf{Z} \mid \mathbf{X}) \approx Q(\mathbf{Z}).$$

The distribution $Q(\mathbf{Z})$ is restricted to belong to a family of distributions of simpler form than $P(\mathbf{Z} \mid \mathbf{X})$ (e.g. a family of Gaussian distributions), selected with the intention of making $Q(\mathbf{Z})$ similar to the true posterior, $P(\mathbf{Z} \mid \mathbf{X})$.

The similarity (or dissimilarity) is measured in terms of a dissimilarity function $d(Q; P)$ and hence inference is performed by selecting the distribution $Q(\mathbf{Z})$ that minimizes $d(Q; P)$.

KL divergence

The most common type of variational Bayes uses the Kullback–Leibler divergence (KL-divergence) of Q from P as the choice of dissimilarity function. This choice makes this minimization tractable. The KL-divergence is defined as

$$D_{\mathrm{KL}}(Q \parallel P) \triangleq \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \frac{Q(\mathbf{Z})}{P(\mathbf{Z} \mid \mathbf{X})}.$$

Note that Q and P are reversed from what one might expect. This use of reversed KL-divergence is conceptually similar to the expectation–maximization algorithm. (Using the KL-divergence in the other way produces the expectation propagation algorithm.)

Intractability

Variational techniques are typically used to form an approximation for:

$$P(\mathbf{Z} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Z})}{P(\mathbf{X})} = \frac{P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{Z})}{\int_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z}')\, d\mathbf{Z}'}$$

The marginalization over $\mathbf{Z}$ to calculate $P(\mathbf{X})$ in the denominator is typically intractable, because, for example, the search space of $\mathbf{Z}$ is combinatorially large. Therefore, we seek an approximation, using $Q(\mathbf{Z}) \approx P(\mathbf{Z} \mid \mathbf{X})$.

Evidence lower bound

Given that $P(\mathbf{Z} \mid \mathbf{X}) = \frac{P(\mathbf{X}, \mathbf{Z})}{P(\mathbf{X})}$, the KL-divergence above can also be written as

$$\begin{aligned} D_{\mathrm{KL}}(Q \parallel P) &= \textstyle\sum_{\mathbf{Z}} Q(\mathbf{Z}) \left[\log \frac{Q(\mathbf{Z})}{P(\mathbf{Z},\mathbf{X})} + \log P(\mathbf{X})\right] \\ &= \textstyle\sum_{\mathbf{Z}} Q(\mathbf{Z}) \left[\log Q(\mathbf{Z}) - \log P(\mathbf{Z},\mathbf{X})\right] + \sum_{\mathbf{Z}} Q(\mathbf{Z}) \left[\log P(\mathbf{X})\right] \end{aligned}$$

Because $P(\mathbf{X})$ is a constant with respect to $\mathbf{Z}$ and $\sum_{\mathbf{Z}} Q(\mathbf{Z}) = 1$ because $Q(\mathbf{Z})$ is a distribution, we have

$$D_{\mathrm{KL}}(Q \parallel P) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \left[\log Q(\mathbf{Z}) - \log P(\mathbf{Z},\mathbf{X})\right] + \log P(\mathbf{X})$$

which, according to the definition of expected value (for a discrete random variable), can be written as follows

$$D_{\mathrm{KL}}(Q \parallel P) = \mathbb{E}_{\mathbf{Q}} \left[\log Q(\mathbf{Z}) - \log P(\mathbf{Z},\mathbf{X})\right] + \log P(\mathbf{X})$$

which can be rearranged to become

$$\begin{aligned} \log P(\mathbf{X}) &= D_{\mathrm{KL}}(Q \parallel P) - \mathbb{E}_{\mathbf{Q}} \left[\log Q(\mathbf{Z}) - \log P(\mathbf{Z},\mathbf{X})\right] \\ &= D_{\mathrm{KL}}(Q \parallel P) + \mathcal{L}(Q) \end{aligned}$$

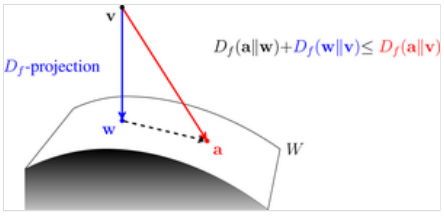
As the *log-evidence* $\log P(\mathbf{X})$ is fixed with respect to Q, maximizing the final term $\mathcal{L}(Q)$ minimizes the KL divergence of Q from P. By appropriate choice of Q, $\mathcal{L}(Q)$ becomes tractable to compute and to maximize. Hence we have both an analytical approximation Q for the posterior $P(\mathbf{Z} \mid \mathbf{X})$, and a lower bound $\mathcal{L}(Q)$ for the log-evidence $\log P(\mathbf{X})$ (since the KL-divergence is non-negative).

The lower bound $\mathcal{L}(Q)$ is known as the (negative) **variational free energy** in analogy with thermodynamic free energy because it can also be expressed as a negative energy $\mathbf{E}_Q[\log P(\mathbf{Z},\mathbf{X})]$ plus the entropy of Q. The term $\mathcal{L}(Q)$ is also known as **Evidence Lower Bound**, abbreviated as **ELBO**, to emphasize that it is a lower (worst-case) bound on the log-evidence of the data.

Proofs

By the generalized Pythagorean theorem of Bregman divergence, of which KL-divergence is a special case, it can be shown that:[1][2]

$$D_{\mathrm{KL}}(Q \parallel P) \geq D_{\mathrm{KL}}(Q \parallel Q^*) + D_{\mathrm{KL}}(Q^* \parallel P), \forall Q^* \in \mathcal{C}$$



Generalized Pythagorean theorem for Bregman divergence[2]

where $\mathcal{C}$ is a convex set and the equality holds if:

$$Q = Q^* \triangleq \arg\min_{Q \in \mathcal{C}} D_{\mathrm{KL}}(Q \parallel P).$$

In this case, the global minimizer $Q^*(\mathbf{Z}) = q^*(\mathbf{Z}_1 \mid \mathbf{Z}_2) q^*(\mathbf{Z}_2) = q^*(\mathbf{Z}_2 \mid \mathbf{Z}_1) q^*(\mathbf{Z}_1)$, with $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2\}$, can be found as follows:[1]

$$\begin{aligned} q^*(\mathbf{Z}_2) &= \frac{P(\mathbf{X})}{\zeta(\mathbf{X})} \frac{P(\mathbf{Z}_2|\mathbf{X})}{\exp(D_{\mathrm{KL}}(q^*(\mathbf{Z}_1|\mathbf{Z}_2) \| P(\mathbf{Z}_1|\mathbf{Z}_2,\mathbf{X})))} \\ &= \frac{1}{\zeta(\mathbf{X})} \exp \mathbb{E}_{q^*(\mathbf{Z}_1|\mathbf{Z}_2)} \left(\log \frac{P(\mathbf{Z},\mathbf{X})}{q^*(\mathbf{Z}_1|\mathbf{Z}_2)}\right), \end{aligned}$$

in which the normalizing constant is:

$$\begin{aligned} \zeta(\mathbf{X}) &= P(\mathbf{X}) \int_{\mathbf{Z}_2} \frac{P(\mathbf{Z}_2|\mathbf{X})}{\exp(D_{\mathrm{KL}}(q^*(\mathbf{Z}_1|\mathbf{Z}_2) \| P(\mathbf{Z}_1|\mathbf{Z}_2,\mathbf{X})))} \\ &= \int_{\mathbf{Z}_2} \exp \mathbb{E}_{q^*(\mathbf{Z}_1|\mathbf{Z}_2)} \left(\log \frac{P(\mathbf{Z},\mathbf{X})}{q^*(\mathbf{Z}_1|\mathbf{Z}_2)}\right). \end{aligned}$$

The term $\zeta(\mathbf{X})$ is often called the evidence lower bound (**ELBO**) in practice, since $P(\mathbf{X}) \geq \zeta(\mathbf{X}) = \exp(\mathcal{L}(Q^*))$,[1] as shown above.

By interchanging the roles of $\mathbf{Z}_1$ and $\mathbf{Z}_2$, we can iteratively compute the approximated $q^*(\mathbf{Z}_1)$ and $q^*(\mathbf{Z}_2)$ of the true model's marginals $P(\mathbf{Z}_1 \mid \mathbf{X})$ and $P(\mathbf{Z}_2 \mid \mathbf{X})$, respectively. Although this iterative scheme is guaranteed to converge monotonically,[1] the converged Q^* is only a local minimizer of $D_{\mathrm{KL}}(Q \parallel P)$.

If the constrained space $\mathcal{C}$ is confined within independent space, i.e. $q^*(\mathbf{Z}_1 \mid \mathbf{Z}_2) = q^*(\mathbf{Z}_1)$,the above iterative scheme will become the so-called mean field approximation $Q^*(\mathbf{Z}) = q^*(\mathbf{Z}_1)q^*(\mathbf{Z}_2)$,as shown below.

Mean field approximation

The variational distribution $Q(\mathbf{Z})$ is usually assumed to factorize over some partition of the latent variables, i.e. for some partition of the latent variables $\mathbf{Z}$ into $\mathbf{Z}_1 \dots \mathbf{Z}_M$,

$$Q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i \mid \mathbf{X})$$

It can be shown using the calculus of variations (hence the name "variational Bayes") that the "best" distribution q_j^* for each of the factors q_j (in terms of the distribution minimizing the KL divergence, as described above) satisfies:[3]

$$q_j^*(\mathbf{Z}_j \mid \mathbf{X}) = \frac{e^{\mathbf{E}_{q^*_{-j}}[\ln p(\mathbf{Z},\mathbf{X})]}}{\int e^{\mathbf{E}_{q^*_{-j}}[\ln p(\mathbf{Z},\mathbf{X})]}\, d\mathbf{Z}_j}$$

where $\mathbf{E}_{q^*_{-j}}[\ln p(\mathbf{Z},\mathbf{X})]$ is the expectation of the logarithm of the joint probability of the data and latent variables, taken with respect to q^* over all variables not in the partition: refer to Lemma 4.1 of[4] for a derivation of the distribution $q_j^*(\mathbf{Z}_j \mid \mathbf{X})$.

In practice, we usually work in terms of logarithms, i.e.:

$$\ln q_j^*(\mathbf{Z}_j \mid \mathbf{X}) = \mathbf{E}_{q^*_{-j}}[\ln p(\mathbf{Z},\mathbf{X})] + \text{constant}$$

The constant in the above expression is related to the normalizing constant (the denominator in the expression above for q_j^*) and is usually reinstated by inspection, as the rest of the expression can usually be recognized as being a known type of distribution (e.g. Gaussian, gamma, etc.).

Using the properties of expectations, the expression $\mathbf{E}_{q^*_{-j}}[\ln p(\mathbf{Z},\mathbf{X})]$ can usually be simplified into a function of the fixed hyperparameters of the prior distributions over the latent variables and of expectations (and sometimes higher moments such as the variance) of latent variables not in the current partition (i.e. latent variables not included in $\mathbf{Z}_j$). This creates circular dependencies between the parameters of the distributions over variables in one partition and the expectations of variables in the other partitions. This naturally suggests an iterative algorithm, much like EM (the expectation–maximization algorithm), in which the expectations (and possibly higher moments) of the latent variables are initialized in some fashion (perhaps randomly), and then the parameters of each distribution are computed in turn using the current values of the expectations, after which the expectation of the newly computed distribution is set appropriately according to the computed parameters. An algorithm of this sort is guaranteed to converge.[5]

In other words, for each of the partitions of variables, by simplifying the expression for the distribution over the partition's variables and examining the distribution's functional dependency on the variables in question, the family of the distribution can usually be determined (which in turn determines the value of the constant). The formula for the distribution's parameters will be expressed in terms of the prior distributions' hyperparameters (which are known constants), but also in terms of expectations of functions of variables in other partitions. Usually these expectations can be simplified into functions of expectations of the variables themselves (i.e. the means); sometimes expectations of squared variables (which can be related to the variance of the variables), or expectations of higher powers (i.e. higher moments) also appear. In most cases, the other variables' distributions will be from known families, and the formulas for the relevant expectations can be looked up. However, those formulas depend on those distributions' parameters, which depend in turn on the expectations about other variables. The result is that the formulas for the parameters of each variable's distributions can be expressed as a series of equations with mutual, nonlinear dependencies among the variables. Usually, it is not possible to solve this system of equations directly. However, as described above, the dependencies suggest a simple iterative algorithm, which in most cases is guaranteed to converge. An example will make this process clearer.

A duality formula for variational inference

The following theorem is referred to as a duality formula for variational inference.[4] It explains some important properties of the variational distributions used in variational Bayes methods.

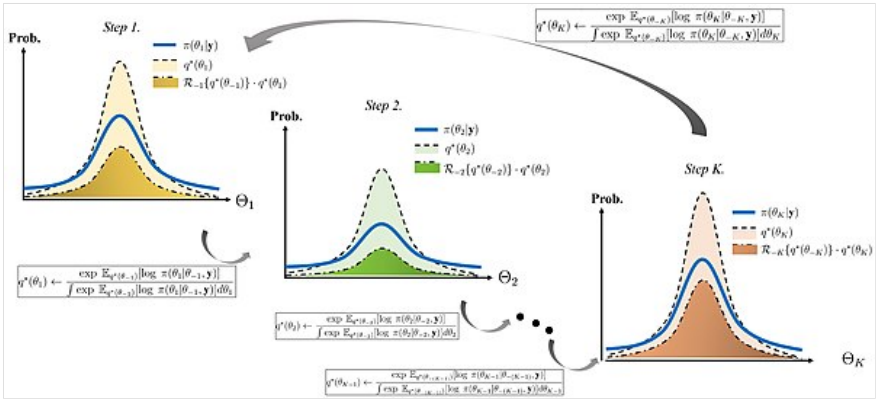
Theorem Consider two probability spaces $(\Theta, \mathcal{F}, P)$ and $(\Theta, \mathcal{F}, Q)$ with $Q \ll P$. Assume that there is a common dominating probability measure λ such that $P \ll \lambda$ and $Q \ll \lambda$. Let h denote any real-valued random variable on $(\Theta, \mathcal{F}, P)$ that satisfies $h \in L_1(P)$. Then the following equality holds

$$\log E_P[\exp h] = \sup_{Q \ll P}\{E_Q[h] - D_{\mathrm{KL}}(Q \parallel P)\}.$$

Further, the supremum on the right-hand side is attained if and only if it holds

$$\frac{q(\theta)}{p(\theta)} = \frac{\exp h(\theta)}{E_P[\exp h]},$$

almost surely with respect to probability measure Q, where $p(\theta) = dP/d\lambda$ and $q(\theta) = dQ/d\lambda$ denote the Radon–Nikodym derivatives of the probability measures P and Q with respect to λ, respectively.



Pictorial illustration of coordinate ascent variational inference algorithm by the duality formula[4]

A basic example

Consider a simple non-hierarchical Bayesian model consisting of a set of i.i.d. observations from a Gaussian distribution, with unknown mean and variance.[6] In the following, we work through this model in great detail to illustrate the workings of the variational Bayes method.

For mathematical convenience, in the following example we work in terms of the precision — i.e. the reciprocal of the variance (or in a multivariate Gaussian, the inverse of the covariance matrix) — rather than the variance itself. (From a theoretical standpoint, precision and variance are equivalent since there is a one-to-one correspondence between the two.)

The mathematical model

We place conjugate prior distributions on the unknown mean μ and precision τ, i.e. the mean also follows a Gaussian distribution while the precision follows a gamma distribution. In other words:

$$\begin{aligned} \tau &\sim \text{Gamma}(a_0, b_0) \\ \mu|\tau &\sim \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1}) \\ \{x_1, \ldots, x_N\} &\sim \mathcal{N}(\mu, \tau^{-1}) \\ N &= \text{number of data points} \end{aligned}$$

The hyperparameters μ_0, λ_0, a_0 and b_0 in the prior distributions are fixed, given values. They can be set to small positive numbers to give broad prior distributions indicating ignorance about the prior distributions of μ and τ.

We are given N data points $\mathbf{X} = \{x_1, \ldots, x_N\}$ and our goal is to infer the posterior distribution $q(\mu, \tau) = p(\mu, \tau \mid x_1, \ldots, x_N)$ of the parameters μ and τ.

The joint probability

The joint probability of all variables can be rewritten as

$$p(\mathbf{X}, \mu, \tau) = p(\mathbf{X} \mid \mu, \tau)p(\mu \mid \tau)p(\tau)$$

where the individual factors are

$$\begin{aligned} p(\mathbf{X} \mid \mu, \tau) &= \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \tau^{-1}) \\ p(\mu \mid \tau) &= \mathcal{N}\left(\mu \mid \mu_0, (\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \text{Gamma}(\tau \mid a_0, b_0) \end{aligned}$$

where

$$\begin{aligned} \mathcal{N}(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \\ \text{Gamma}(\tau \mid a, b) &= \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau} \end{aligned}$$

Factorized approximation

Assume that $q(\mu, \tau) = q(\mu)q(\tau)$, i.e. that the posterior distribution factorizes into independent factors for μ and τ. This type of assumption underlies the variational Bayesian method. The true posterior distribution does not in fact factor this way (in fact, in this simple case, it is known to be a Gaussian-gamma distribution), and hence the result we obtain will be an approximation.

Derivation of $q(\mu)$

Then

$$
\begin{aligned}
\ln q_\mu^*(\mu) &= \mathrm{E}_\tau[\ln p(\mathbf{X} \mid \mu, \tau) + \ln p(\mu \mid \tau) + \ln p(\tau)] + C \\
&= \mathrm{E}_\tau[\ln p(\mathbf{X} \mid \mu, \tau)] + \mathrm{E}_\tau[\ln p(\mu \mid \tau)] + \mathrm{E}_\tau[\ln p(\tau)] + C \\
&= \mathrm{E}_\tau\left[\ln \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \tau^{-1}\right)\right] + \mathrm{E}_\tau\left[\ln \mathcal{N}\left(\mu \mid \mu_0, (\lambda_0\tau)^{-1}\right)\right] + C_2 \\
&= \mathrm{E}_\tau\left[\ln \prod_{n=1}^{N} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{(x_n-\mu)^2\tau}{2}}\right] + \mathrm{E}_\tau\left[\ln \sqrt{\frac{\lambda_0\tau}{2\pi}} e^{-\frac{(\mu-\mu_0)^2\lambda_0\tau}{2}}\right] + C_2 \\
&= \mathrm{E}_\tau\left[\sum_{n=1}^{N} \left(\frac{1}{2}(\ln\tau - \ln 2\pi) - \frac{(x_n-\mu)^2\tau}{2}\right)\right] + \mathrm{E}_\tau\left[\frac{1}{2}(\ln\lambda_0 + \ln\tau - \ln 2\pi) - \frac{(\mu-\mu_0)^2\lambda_0\tau}{2}\right] + C_2 \\
&= \mathrm{E}_\tau\left[\sum_{n=1}^{N} -\frac{(x_n-\mu)^2\tau}{2}\right] + \mathrm{E}_\tau\left[-\frac{(\mu-\mu_0)^2\lambda_0\tau}{2}\right] + \mathrm{E}_\tau\left[\sum_{n=1}^{N} \frac{1}{2}(\ln\tau - \ln 2\pi)\right] + \mathrm{E}_\tau\left[\frac{1}{2}(\ln\lambda_0 + \ln\tau - \ln 2\pi)\right] + C_2 \\
&= \mathrm{E}_\tau\left[\sum_{n=1}^{N} -\frac{(x_n-\mu)^2\tau}{2}\right] + \mathrm{E}_\tau\left[-\frac{(\mu-\mu_0)^2\lambda_0\tau}{2}\right] + C_3 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right\} + C_3
\end{aligned}
$$

In the above derivation, C, C_2 and C_3 refer to values that are constant with respect to μ. Note that the term $\mathrm{E}_\tau[\ln p(\tau)]$ is not a function of μ and will have the same value regardless of the value of μ. Hence in line 3 we can absorb it into the constant term at the end. We do the same thing in line 7.

The last line is simply a quadratic polynomial in μ. Since this is the logarithm of $q_\mu^*(\mu)$, we can see that $q_\mu^*(\mu)$ itself is a Gaussian distribution.

With a certain amount of tedious math (expanding the squares inside of the braces, separating out and grouping the terms involving μ and μ^2 and completing the square over μ), we can derive the parameters of the Gaussian distribution:

$$
\begin{aligned}
\ln q_\mu^*(\mu) &= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right\} + C_3 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{\sum_{n=1}^{N}(x_n^2 - 2x_n\mu + \mu^2) + \lambda_0(\mu^2 - 2\mu_0\mu + \mu_0^2)\right\} + C_3 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{\left(\sum_{n=1}^{N} x_n^2\right) - 2\left(\sum_{n=1}^{N} x_n\right)\mu + \left(\sum_{n=1}^{N}\mu^2\right) + \lambda_0\mu^2 - 2\lambda_0\mu_0\mu + \lambda_0\mu_0^2\right\} + C_3 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{(\lambda_0+N)\mu^2 - 2\left(\lambda_0\mu_0 + \sum_{n=1}^{N} x_n\right)\mu + \left(\sum_{n=1}^{N} x_n^2\right) + \lambda_0\mu_0^2\right\} + C_3 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{(\lambda_0+N)\mu^2 - 2\left(\lambda_0\mu_0 + \sum_{n=1}^{N} x_n\right)\mu\right\} + C_4 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{(\lambda_0+N)\mu^2 - 2\left(\frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)(\lambda_0+N)\mu\right\} + C_4 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{(\lambda_0+N)\left(\mu^2 - 2\left(\frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)\mu\right)\right\} + C_4 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{(\lambda_0+N)\left(\mu^2 - 2\left(\frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)\mu + \left(\frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)^2 - \left(\frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)^2\right)\right\} + C_4 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{(\lambda_0+N)\left(\mu^2 - 2\left(\frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)\mu + \left(\frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)^2\right)\right\} + C_5 \\
&= -\frac{\mathrm{E}_\tau[\tau]}{2}\left\{(\lambda_0+N)\left(\mu - \frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)^2\right\} + C_5 \\
&= -\frac{1}{2}(\lambda_0+N)\,\mathrm{E}_\tau[\tau]\left(\mu - \frac{\lambda_0\mu_0 + \sum_{n=1}^{N} x_n}{\lambda_0+N}\right)^2 + C_5
\end{aligned}
$$

Note that all of the above steps can be shortened by using the formula for the sum of two quadratics.

In other words:

$$
\begin{aligned}
q_\mu^*(\mu) &\sim \mathcal{N}(\mu \mid \mu_N, \lambda_N^{-1}) \\
\mu_N &= \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N} \\
\lambda_N &= (\lambda_0 + N)\,\mathrm{E}_\tau[\tau] \\
\bar{x} &= \frac{1}{N}\sum_{n=1}^{N} x_n
\end{aligned}
$$

Derivation of q(τ)

The derivation of $q_\tau^*(\tau)$ is similar to above, although we omit some of the details for the sake of brevity.

$$
\begin{aligned}
\ln q_\tau^*(\tau) &= \mathrm{E}_\mu[\ln p(\mathbf{X} \mid \mu, \tau) + \ln p(\mu \mid \tau)] + \ln p(\tau) + \text{constant} \\
&= (a_0 - 1)\ln\tau - b_0\tau + \frac{1}{2}\ln\tau + \frac{N}{2}\ln\tau - \frac{\tau}{2}\,\mathrm{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] + \text{constant}
\end{aligned}
$$

Exponentiating both sides, we can see that $q_\tau^*(\tau)$ is a gamma distribution. Specifically:

$$
\begin{aligned}
q_\tau^*(\tau) &\sim \mathrm{Gamma}(\tau \mid a_N, b_N) \\
a_N &= a_0 + \frac{N+1}{2} \\
b_N &= b_0 + \frac{1}{2}\,\mathrm{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]
\end{aligned}
$$

Algorithm for computing the parameters

Let us recap the conclusions from the previous sections:

$$
\begin{aligned}
q_\mu^*(\mu) &\sim \mathcal{N}(\mu \mid \mu_N, \lambda_N^{-1}) \\
\mu_N &= \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N} \\
\lambda_N &= (\lambda_0 + N)\,\mathrm{E}_\tau[\tau] \\
\bar{x} &= \frac{1}{N}\sum_{n=1}^{N} x_n
\end{aligned}
$$

and

$$
\begin{aligned}
q_\tau^*(\tau) &\sim \mathrm{Gamma}(\tau \mid a_N, b_N) \\
a_N &= a_0 + \frac{N+1}{2} \\
b_N &= b_0 + \frac{1}{2}\,\mathrm{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]
\end{aligned}
$$

In each case, the parameters for the distribution over one of the variables depend on expectations taken with respect to the other variable. We can expand the expectations, using the standard formulas for the expectations of moments of the Gaussian and gamma distributions:

$$
\begin{aligned}
\mathrm{E}[\tau \mid a_N, b_N] &= \frac{a_N}{b_N} \\
\mathrm{E}\left[\mu \mid \mu_N, \lambda_N^{-1}\right] &= \mu_N \\
\mathrm{E}\left[X^2\right] &= \mathrm{Var}(X) + (\mathrm{E}[X])^2 \\
\mathrm{E}\left[\mu^2 \mid \mu_N, \lambda_N^{-1}\right] &= \lambda_N^{-1} + \mu_N^2
\end{aligned}
$$

Applying these formulas to the above equations is trivial in most cases, but the equation for b_N takes more work:

$$
\begin{aligned}
b_N &= b_0 + \frac{1}{2}\,\mathrm{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] \\
&= b_0 + \frac{1}{2}\,\mathrm{E}_\mu\left[(\lambda_0 + N)\mu^2 - 2\left(\lambda_0\mu_0 + \sum_{n=1}^{N} x_n\right)\mu + \left(\sum_{n=1}^{N} x_n^2\right) + \lambda_0\mu_0^2\right] \\
&= b_0 + \frac{1}{2}\left[(\lambda_0 + N)\,\mathrm{E}_\mu[\mu^2] - 2\left(\lambda_0\mu_0 + \sum_{n=1}^{N} x_n\right)\mathrm{E}_\mu[\mu] + \left(\sum_{n=1}^{N} x_n^2\right) + \lambda_0\mu_0^2\right] \\
&= b_0 + \frac{1}{2}\left[(\lambda_0 + N)\left(\lambda_N^{-1} + \mu_N^2\right) - 2\left(\lambda_0\mu_0 + \sum_{n=1}^{N} x_n\right)\mu_N + \left(\sum_{n=1}^{N} x_n^2\right) + \lambda_0\mu_0^2\right]
\end{aligned}
$$

We can then write the parameter equations as follows, without any expectations:

$$\mu_N = \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N}$$
$$\lambda_N = (\lambda_0 + N)\frac{a_N}{b_N}$$
$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$
$$a_N = a_0 + \frac{N+1}{2}$$
$$b_N = b_0 + \frac{1}{2}\left[(\lambda_0 + N)\left(\lambda_N^{-1} + \mu_N^2\right) - 2\left(\lambda_0\mu_0 + \sum_{n=1}^{N} x_n\right)\mu_N + \left(\sum_{n=1}^{N} x_n^2\right) + \lambda_0\mu_0^2\right]$$

Note that there are circular dependencies among the formulas for λ_N and b_N. This naturally suggests an EM-like algorithm:

1. Compute $\sum_{n=1}^{N} x_n$ and $\sum_{n=1}^{N} x_n^2$. Use these values to compute μ_N and a_N.
2. Initialize λ_N to some arbitrary value.
3. Use the current value of λ_N, along with the known values of the other parameters, to compute b_N.
4. Use the current value of b_N, along with the known values of the other parameters, to compute λ_N.
5. Repeat the last two steps until convergence (i.e. until neither value has changed more than some small amount).

We then have values for the hyperparameters of the approximating distributions of the posterior parameters, which we can use to compute any properties we want of the posterior — e.g. its mean and variance, a 95% highest-density region (the smallest interval that includes 95% of the total probability), etc.

It can be shown that this algorithm is guaranteed to converge to a local maximum.

Note also that the posterior distributions have the same form as the corresponding prior distributions. We did *not* assume this; the only assumption we made was that the distributions factorize, and the form of the distributions followed naturally. It turns out (see below) that the fact that the posterior distributions have the same form as the prior distributions is not a coincidence, but a general result whenever the prior distributions are members of the exponential family, which is the case for most of the standard distributions.

Further discussion

Step-by-step recipe

The above example shows the method by which the variational-Bayesian approximation to a posterior probability density in a given Bayesian network is derived:

1. Describe the network with a graphical model, identifying the observed variables (data) $\mathbf{X}$ and unobserved variables (parameters $\boldsymbol{\Theta}$ and latent variables $\mathbf{Z}$) and their conditional probability distributions. Variational Bayes will then construct an approximation to the posterior probability $p(\mathbf{Z}, \boldsymbol{\Theta} \mid \mathbf{X})$. The approximation has the basic property that it is a factorized distribution, i.e. a product of two or more independent distributions over disjoint subsets of the unobserved variables.
2. Partition the unobserved variables into two or more subsets, over which the independent factors will be derived. There is no universal procedure for doing this; creating too many subsets yields a poor approximation, while creating too few makes the entire variational Bayes procedure intractable. Typically, the first split is to separate the parameters and latent variables; often, this is enough by itself to produce a tractable result. Assume that the partitions are called $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$.
3. For a given partition $\mathbf{Z}_j$, write down the formula for the best approximating distribution $q_j^*(\mathbf{Z}_j \mid \mathbf{X})$ using the basic equation $\ln q_j^*(\mathbf{Z}_j \mid \mathbf{X}) = \mathrm{E}_{i \neq j}[\ln p(\mathbf{Z}, \mathbf{X})] + \text{constant}$.
4. Fill in the formula for the joint probability distribution using the graphical model. Any component conditional distributions that don't involve any of the variables in $\mathbf{Z}_j$ can be ignored; they will be folded into the constant term.
5. Simplify the formula and apply the expectation operator, following the above example. Ideally, this should simplify into expectations of basic functions of variables not in $\mathbf{Z}_j$ (e.g. first or second raw moments, expectation of a logarithm, etc.). In order for the variational Bayes procedure to work well, these expectations should generally be expressible analytically as functions of the parameters and/or hyperparameters of the distributions of these variables. In all cases, these expectation terms are constants with respect to the variables in the current partition.
6. The functional form of the formula with respect to the variables in the current partition indicates the type of distribution. In particular, exponentiating the formula generates the probability density function (PDF) of the distribution (or at least, something proportional to it, with unknown normalization constant). In order for the overall method to be tractable, it should be possible to recognize the functional form as belonging to a known distribution. Significant mathematical manipulation may be required to convert the formula into a form that matches the PDF of a known distribution. When this can be done, the normalization constant can be reinstated by definition, and equations for the parameters of the known distribution can be derived by extracting the appropriate parts of the formula.
7. When all expectations can be replaced analytically with functions of variables not in the current partition, and the PDF put into a form that allows identification with a known distribution, the result is a set of equations expressing the values of the optimum parameters as functions of the parameters of variables in other partitions.

8. When this procedure can be applied to all partitions, the result is a set of mutually linked equations specifying the optimum values of all parameters.
9. An expectation–maximization (EM) type procedure is then applied, picking an initial value for each parameter and the iterating through a series of steps, where at each step we cycle through the equations, updating each parameter in turn. This is guaranteed to converge.

Most important points

Due to all of the mathematical manipulations involved, it is easy to lose track of the big picture. The important things are:

1. The idea of variational Bayes is to construct an analytical approximation to the posterior probability of the set of unobserved variables (parameters and latent variables), given the data. This means that the form of the solution is similar to other Bayesian inference methods, such as Gibbs sampling — i.e. a distribution that seeks to describe everything that is known about the variables. As in other Bayesian methods — but unlike e.g. in expectation–maximization (EM) or other maximum likelihood methods — both types of unobserved variables (i.e. parameters and latent variables) are treated the same, i.e. as random variables. Estimates for the variables can then be derived in the standard Bayesian ways, e.g. calculating the mean of the distribution to get a single point estimate or deriving a credible interval, highest density region, etc.
2. "Analytical approximation" means that a formula can be written down for the posterior distribution. The formula generally consists of a product of well-known probability distributions, each of which *factorizes* over a set of unobserved variables (i.e. it is conditionally independent of the other variables, given the observed data). This formula is not the true posterior distribution, but an approximation to it; in particular, it will generally agree fairly closely in the lowest moments of the unobserved variables, e.g. the mean and variance.
3. The result of all of the mathematical manipulations is (1) the identity of the probability distributions making up the factors, and (2) mutually dependent formulas for the parameters of these distributions. The actual values of these parameters are computed numerically, through an alternating iterative procedure much like EM.

Compared with expectation–maximization (EM)

Variational Bayes (VB) is often compared with expectation–maximization (EM). The actual numerical procedure is quite similar, in that both are alternating iterative procedures that successively converge on optimum parameter values. The initial steps to derive the respective procedures are also vaguely similar, both starting out with formulas for probability densities and both involving significant amounts of mathematical manipulations.

However, there are a number of differences. Most important is *what* is being computed.

- EM computes point estimates of posterior distribution of those random variables that can be categorized as "parameters", but only estimates of the actual posterior distributions of the latent variables (at least in "soft EM", and often only when the latent variables are discrete). The point estimates computed are the modes of these parameters; no other information is available.
- VB, on the other hand, computes estimates of the actual posterior distribution of all variables, both parameters and latent variables. When point estimates need to be derived, generally the mean is used rather than the mode, as is normal in Bayesian inference. Concomitant with this, the parameters computed in VB do *not* have the same significance as those in EM. EM computes optimum values of the parameters of the Bayes network itself. VB computes optimum values of the parameters of the distributions used to approximate the parameters and latent variables of the Bayes network. For example, a typical Gaussian mixture model will have parameters for the mean and variance of each of the mixture components. EM would directly estimate optimum values for these parameters. VB, however, would first fit a distribution to these parameters — typically in the form of a prior distribution, e.g. a normal-scaled inverse gamma distribution — and would then compute values for the parameters of this prior distribution, i.e. essentially hyperparameters. In this case, VB would compute optimum estimates of the four parameters of the normal-scaled inverse gamma distribution that describes the joint distribution of the mean and variance of the component.

A more complex example

Imagine a Bayesian Gaussian mixture model described as follows:[3]

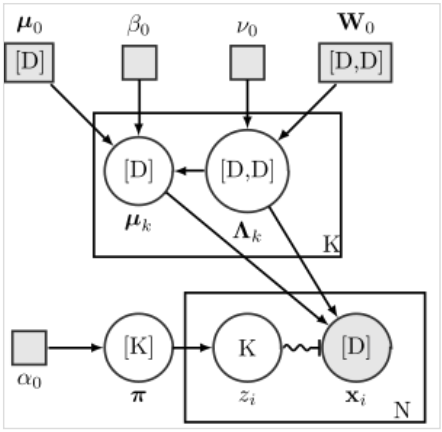
$$
\begin{aligned}
\boldsymbol{\pi} &\sim \operatorname{SymDir}(K, \alpha_0) \\
\boldsymbol{\Lambda}_{i=1\dots K} &\sim \mathcal{W}(\mathbf{W}_0, \nu_0) \\
\boldsymbol{\mu}_{i=1\dots K} &\sim \mathcal{N}(\mu_0, (\beta_0\boldsymbol{\Lambda}_i)^{-1}) \\
\mathbf{z}[i=1\dots N] &\sim \operatorname{Mult}(1, \boldsymbol{\pi}) \\
\mathbf{x}_{i=1\dots N} &\sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, {\boldsymbol{\Lambda}_{z_i}}^{-1}) \\
K &= \text{number of mixing components} \\
N &= \text{number of data points}
\end{aligned}
$$

Note:

- SymDir() is the symmetric Dirichlet distribution of dimension K, with the hyperparameter for each component set to α_0. The Dirichlet distribution is the conjugate prior of the categorical distribution or multinomial distribution.
- $\mathcal{W}()$ is the Wishart distribution, which is the conjugate prior of the precision matrix (inverse covariance matrix) for a multivariate Gaussian distribution.
- Mult() is a multinomial distribution over a single observation (equivalent to a categorical distribution). The state space is a "one-of-K" representation, i.e., a K-dimensional vector in which one of the elements is 1 (specifying the identity of the observation) and all other elements are 0.
- $\mathcal{N}()$ is the Gaussian distribution, in this case specifically the multivariate Gaussian distribution.

The interpretation of the above variables is as follows:

- $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is the set of N data points, each of which is a D-dimensional vector distributed according to a multivariate Gaussian distribution.
- $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ is a set of latent variables, one per data point, specifying which mixture component the corresponding data point belongs to, using a "one-of-K" vector representation with components z_{nk} for $k = 1 \dots K$, as described above.
- π is the mixing proportions for the K mixture components.
- $\mu_{i=1\dots K}$ and $\mathbf{\Lambda}_{i=1\dots K}$ specify the parameters (mean and precision) associated with each mixture component.



Bayesian Gaussian mixture model using plate notation. Smaller squares indicate fixed parameters; larger circles indicate random variables. Filled-in shapes indicate known values. The indication [K] means a vector of size *K*; [*D*,*D*] means a matrix of size *D*×*D*; *K* alone means a categorical variable with *K* outcomes. The squiggly line coming from *z* ending in a crossbar indicates a *switch* — the value of this variable selects, for the other incoming variables, which value to use out of the size-*K* array of possible values.

The joint probability of all variables can be rewritten as

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \mathbf{\Lambda}) = p(\mathbf{X} \mid \mathbf{Z}, \mu, \mathbf{\Lambda}) p(\mathbf{Z} \mid \pi) p(\pi) p(\mu \mid \mathbf{\Lambda}) p(\mathbf{\Lambda})$$

where the individual factors are

$$\begin{aligned}
p(\mathbf{X} \mid \mathbf{Z}, \mu, \mathbf{\Lambda}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Lambda}_k^{-1})^{z_{nk}} \\
p(\mathbf{Z} \mid \pi) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \\
p(\pi) &= \frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \\
p(\mu \mid \mathbf{\Lambda}) &= \prod_{k=1}^{K} \mathcal{N}(\mu_k \mid \mu_0, (\beta_0 \mathbf{\Lambda}_k)^{-1}) \\
p(\mathbf{\Lambda}) &= \prod_{k=1}^{K} \mathcal{W}(\mathbf{\Lambda}_k \mid \mathbf{W}_0, \nu_0)
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{N}(\mathbf{x} \mid \mu, \mathbf{\Sigma}) &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \mu)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \right\} \\
\mathcal{W}(\mathbf{\Lambda} \mid \mathbf{W}, \nu) &= B(\mathbf{W}, \nu) |\mathbf{\Lambda}|^{(\nu - D - 1)/2} \exp\left(-\frac{1}{2} \operatorname{Tr}(\mathbf{W}^{-1} \mathbf{\Lambda}) \right) \\
B(\mathbf{W}, \nu) &= |\mathbf{W}|^{-\nu/2} \left\{ 2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu + 1 - i}{2} \right) \right\}^{-1} \\
D &= \text{dimensionality of each data point}
\end{aligned}$$

Assume that $q(\mathbf{Z}, \pi, \mu, \mathbf{\Lambda}) = q(\mathbf{Z}) q(\pi, \mu, \mathbf{\Lambda})$.

Then[3]

$$\begin{aligned}
\ln q^*(\mathbf{Z}) &= \mathrm{E}_{\pi, \mu, \mathbf{\Lambda}}[\ln p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \mathbf{\Lambda})] + \text{constant} \\
&= \mathrm{E}_{\pi}[\ln p(\mathbf{Z} \mid \pi)] + \mathrm{E}_{\mu, \mathbf{\Lambda}}[\ln p(\mathbf{X} \mid \mathbf{Z}, \mu, \mathbf{\Lambda})] + \text{constant} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{constant}
\end{aligned}$$

where we have defined

$$\ln \rho_{nk} = \mathrm{E}[\ln \pi_k] + \frac{1}{2} \mathrm{E}[\ln |\mathbf{\Lambda}_k|] - \frac{D}{2} \ln(2\pi) - \frac{1}{2} \mathrm{E}_{\mu_k, \mathbf{\Lambda}_k}[(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \mathbf{\Lambda}_k (\mathbf{x}_n - \mu_k)]$$

Exponentiating both sides of the formula for $\ln q^*(\mathbf{Z})$ yields

$$q^*(\mathbf{Z}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

Requiring that this be normalized ends up requiring that the ρ_{nk} sum to 1 over all values of k, yielding

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

where

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}$$

In other words, $q^*(\mathbf{Z})$ is a product of single-observation multinomial distributions, and factors over each individual $\mathbf{z}_n$, which is distributed as a single-observation multinomial distribution with parameters r_{nk} for $k = 1 \dots K$.

Furthermore, we note that

$$\mathbf{E}[z_{nk}] = r_{nk}$$

which is a standard result for categorical distributions.

Now, considering the factor $q(\pi, \mu, \mathbf{\Lambda})$, note that it automatically factors into $q(\pi)\prod_{k=1}^{K} q(\mu_k, \mathbf{\Lambda}_k)$ due to the structure of the graphical model defining our Gaussian mixture model, which is specified above.

Then,

$$\begin{aligned}\ln q^*(\pi) &= \ln p(\pi) + \mathbf{E}_{\mathbf{Z}}[\ln p(\mathbf{Z} \mid \pi)] + \text{constant} \\ &= (\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \pi_k + \text{constant}\end{aligned}$$

Taking the exponential of both sides, we recognize $q^*(\pi)$ as a Dirichlet distribution

$$q^*(\pi) \sim \mathbf{Dir}(\alpha)$$

where

$$\alpha_k = \alpha_0 + N_k$$

where

$$N_k = \sum_{n=1}^{N} r_{nk}$$

Finally

$$\ln q^*(\mu_k, \mathbf{\Lambda}_k) = \ln p(\mu_k, \mathbf{\Lambda}_k) + \sum_{n=1}^{N} \mathbf{E}[z_{nk}] \ln \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Lambda}_k^{-1}) + \text{constant}$$

Grouping and reading off terms involving μ_k and $\mathbf{\Lambda}_k$, the result is a Gaussian-Wishart distribution given by

$$q^*(\mu_k, \mathbf{\Lambda}_k) = \mathcal{N}(\mu_k \mid \mathbf{m}_k, (\beta_k \mathbf{\Lambda}_k)^{-1})\mathcal{W}(\mathbf{\Lambda}_k \mid \mathbf{W}_k, \nu_k)$$

given the definitions

$$\begin{aligned}\beta_k &= \beta_0 + N_k \\ \mathbf{m}_k &= \frac{1}{\beta_k}(\beta_0 \mu_0 + N_k \bar{\mathbf{x}}_k) \\ \mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\mathbf{x}}_k - \mu_0)(\bar{\mathbf{x}}_k - \mu_0)^{\mathrm{T}} \\ \nu_k &= \nu_0 + N_k \\ N_k &= \sum_{n=1}^{N} r_{nk} \\ \bar{\mathbf{x}}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n \\ \mathbf{S}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^{\mathrm{T}}\end{aligned}$$

Finally, notice that these functions require the values of r_{nk}, which make use of ρ_{nk}, which is defined in turn based on $\mathbf{E}[\ln \pi_k]$, $\mathbf{E}[\ln |\mathbf{\Lambda}_k|]$, and $\mathbf{E}_{\mu_k, \mathbf{\Lambda}_k}[(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \mathbf{\Lambda}_k (\mathbf{x}_n - \mu_k)]$. Now that we have determined the distributions over which these expectations are taken, we can derive formulas for them:

$$\begin{aligned}\mathbf{E}_{\mu_k, \mathbf{\Lambda}_k}[(\mathbf{x}_n - \mu_k)^{\mathrm{T}} \mathbf{\Lambda}_k (\mathbf{x}_n - \mu_k)] &= D\beta_k^{-1} + \nu_k (\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}} \mathbf{W}_k (\mathbf{x}_n - \mathbf{m}_k) \\ \ln \tilde{\Lambda}_k \equiv \mathbf{E}[\ln |\mathbf{\Lambda}_k|] &= \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \ln 2 + \ln |\mathbf{W}_k| \\ \ln \tilde{\pi}_k \equiv \mathbf{E}[\ln |\pi_k|] &= \psi(\alpha_k) - \psi\left(\sum_{i=1}^{K} \alpha_i\right)\end{aligned}$$

These results lead to

$$r_{nk} \propto \tilde{\pi}_k \widetilde{\Lambda}_k^{1/2} \exp\left\{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}} \mathbf{W}_k (\mathbf{x}_n - \mathbf{m}_k)\right\}$$

These can be converted from proportional to absolute values by normalizing over k so that the corresponding values sum to 1.

Note that:

1. The update equations for the parameters β_k, $\mathbf{m}_k$, $\mathbf{W}_k$ and ν_k of the variables μ_k and $\mathbf{\Lambda}_k$ depend on the statistics N_k, $\bar{\mathbf{x}}_k$, and $\mathbf{S}_k$, and these statistics in turn depend on r_{nk}.
2. The update equations for the parameters $\alpha_{1\ldots K}$ of the variable π depend on the statistic N_k, which depends in turn on r_{nk}.
3. The update equation for r_{nk} has a direct circular dependence on β_k, $\mathbf{m}_k$, $\mathbf{W}_k$ and ν_k as well as an indirect circular dependence on $\mathbf{W}_k$, ν_k and $\alpha_{1\ldots K}$ through $\tilde{\pi}_k$ and $\widetilde{\Lambda}_k$.

This suggests an iterative procedure that alternates between two steps:

1. An E-step that computes the value of r_{nk} using the current values of all the other parameters.
2. An M-step that uses the new value of r_{nk} to compute new values of all the other parameters.

Note that these steps correspond closely with the standard EM algorithm to derive a maximum likelihood or maximum a posteriori (MAP) solution for the parameters of a Gaussian mixture model. The responsibilities r_{nk} in the E step correspond closely to the posterior probabilities of the latent variables given the data, i.e. $p(\mathbf{Z} \mid \mathbf{X})$; the computation of the statistics N_k, $\bar{\mathbf{x}}_k$, and $\mathbf{S}_k$ corresponds closely to the computation of corresponding "soft-count" statistics over the data; and the use of those statistics to compute new values of the parameters corresponds closely to the use of soft counts to compute new parameter values in normal EM over a Gaussian mixture model.

Exponential-family distributions

Note that in the previous example, once the distribution over unobserved variables was assumed to factorize into distributions over the "parameters" and distributions over the "latent data", the derived "best" distribution for each variable was in the same family as the corresponding prior distribution over the variable. This is a general result that holds true for all prior distributions derived from the exponential family.

See also

- Variational message passing: a modular algorithm for variational Bayesian inference.
- Variational autoencoder: an artificial neural network belonging to the families of probabilistic graphical models and Variational Bayesian methods.
- Expectation–maximization algorithm: a related approach which corresponds to a special case of variational Bayesian inference.
- Generalized filtering: a variational filtering scheme for nonlinear state space models.
- Calculus of variations: the field of mathematical analysis that deals with maximizing or minimizing functionals.
- Maximum entropy discrimination: This is a variational inference framework that allows for introducing and accounting for additional large-margin constraints[7]

References

1. Tran, Viet Hung (2018). "Copula Variational Bayes inference via information geometry". arXiv:1803.10998 (https://arxiv.org/abs/1803.10998) [cs.IT (https://arxiv.org/archive/cs.IT)].
2. Adamčík, Martin (2014). "The Information Geometry of Bregman Divergences and Some Applications in Multi-Expert Reasoning" (https://doi.org/10.3390%2Fe16126338). *Entropy*. **16** (12): 6338–6381. Bibcode:2014Entrp..16.6338A (https://ui.adsabs.harvard.edu/abs/2014Entrp..16.6338A). doi:10.3390/e16126338 (https://doi.org/10.3390%2Fe16126338).
3. Nguyen, Duy (15 August 2023). "AN IN DEPTH INTRODUCTION TO VARIATIONAL BAYES NOTE" (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4541076). doi:10.2139/ssrn.4541076 (https://doi.org/10.2139%2Fssrn.4541076). SSRN 4541076 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4541076). Retrieved 15 August 2023.
4. Lee, Se Yoon (2021). "Gibbs sampler and coordinate ascent variational inference: A set-theoretical review". *Communications in Statistics - Theory and Methods*. **51** (6): 1–21. arXiv:2008.01006 (https://arxiv.org/abs/2008.01006). doi:10.1080/03610926.2021.1921214 (https://doi.org/10.1080%2F03610926.2021.1921214). S2CID 220935477 (https://api.semanticscholar.org/CorpusID:220935477).
5. Boyd, Stephen P.; Vandenberghe, Lieven (2004). *Convex Optimization* (https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf) (PDF). Cambridge University Press. ISBN 978-0-521-83378-3. Retrieved October 15, 2011.
6. Bishop, Christopher M. (2006). "Chapter 10". *Pattern Recognition and Machine Learning*. Springer. ISBN 978-0-387-31073-2.

7. Sotirios P. Chatzis, “Infinite Markov-Switching Maximum Entropy Discrimination Machines (http://proceedings.mlr.press/v28/chatzis13.pdf),” Proc. 30th International Conference on Machine Learning (ICML). Journal of Machine Learning Research: Workshop and Conference Proceedings, vol. 28, no. 3, pp. 729–737, June 2013.

External links

- The on-line textbook: Information Theory, Inference, and Learning Algorithms (https://www.inference.phy.cam.ac.uk/mackay/itila/), by David J.C. MacKay provides an introduction to variational methods (p. 422).
- A Tutorial on Variational Bayes (https://www.robots.ox.ac.uk/~sjrob/Pubs/fox_vbtut.pdf). Fox, C. and Roberts, S. 2012. Artificial Intelligence Review, doi:10.1007/s10462-011-9236-8 (https://doi.org/10.1007%2Fs10462-011-9236-8).
- Variational-Bayes Repository (https://www.gatsby.ucl.ac.uk/vbayes/) A repository of research papers, software, and links related to the use of variational methods for approximate Bayesian learning up to 2003.
- Variational Algorithms for Approximate Bayesian Inference (https://www.cse.buffalo.edu/faculty/mbeal/thesis/index.html), by M. J. Beal includes comparisons of EM to Variational Bayesian EM and derivations of several models including Variational Bayesian HMMs.
- High-Level Explanation of Variational Inference (https://www.cs.jhu.edu/~jason/tutorials/variational.html) by Jason Eisner may be worth reading before a more mathematically detailed treatment.
- Copula Variational Bayes inference via information geometry (pdf) (https://arxiv.org/abs/1803.10998) by Tran, V.H. 2018. This paper is primarily written for students. Via Bregman divergence, the paper shows that Variational Bayes is simply a generalized Pythagorean projection of true model onto an arbitrarily correlated (copula) distributional space, of which the independent space is merely a special case.
- An in depth introduction to Variational Bayes note (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4541076). Nguyen, D. 2023