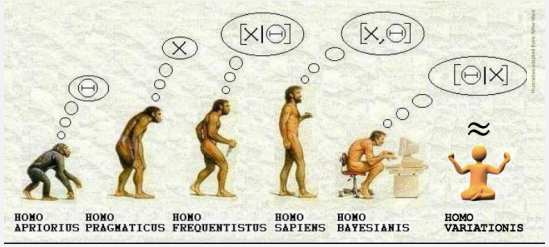
Foundations of Statistical Inference

Julien Berestycki

Department of Statistics
University of Oxford

MT 2016

Lecture 14 : Variational Bayes



"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."
John W. Tukey, 1915 -2000

Laplace approximation

The Laplace approximation provides a way of approximating a density whose normalisation constant we cannot evaluate, by fitting a Gaussian distribution to its mode.

Pierre-Simon Laplace
(1749 - 1827)

$$\underbrace{p(z)}_{\text{proba. density}} = \underbrace{\frac{1}{Z}}_{\text{Unknown constant}} \times \underbrace{f(z)}_{\text{Main part of the density (easy to evaluate)}}$$

Observe this is exactly the situation we face in Bayesian inference

$$\underbrace{p(\theta|y)}_{\text{posterior density}} = \underbrace{\frac{1}{p(y)}}_{\text{marginal dist.}} \times \underbrace{p(\theta, y)}_{\text{joint proba (likelihood x prior)}}$$

Deriving Laplace approximation

Idea: 2nd order Taylor approximation to $\ell(\theta) = \log p(y, \theta)$ around mode θ^*.

$$\begin{aligned} \ell(\theta) &\approx \ell(\theta^*) + \underbrace{\ell'(\theta^*)}_{=0}(\theta - \theta^*) + \frac{1}{2}\ell''(\theta^*)(\theta - \theta^*)^2 \\ &= \ell(\theta^*) + \frac{1}{2}\ell''(\theta^*)(\theta - \theta^*)^2 \end{aligned}$$

Recognize Gaussian density

$$\log \mathcal{N}(\theta|\mu, \sigma^2) = -\log \sigma - \frac{1}{2}\log 2\pi - \frac{1}{2}\sigma^{-2}(\theta - \mu)^2$$

So approximate posterior by:

$$q(\theta) = \mathcal{N}(\theta|\mu, \sigma^2) \qquad \text{with } \mu = \theta^* \qquad \text{(mode of log-posterior)}$$

and $\sigma^{-2} = -\ell''(\theta^*)$ (negative curvature at the mode)

Computing integrals

More generally, assume $f(x)$ has a unique global maximum at x_0.

$$f(x) \simeq f(x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2$$

so

$$\int_a^b e^{Nf(x)}dx \simeq e^{Nf(x_0)} \int_a^b e^{-N|f''(x_0)|(x-x_0)^2/2}dx$$

To obtain

Lemma

$$\int_a^b e^{Nf(x)}x \sim \sqrt{\frac{2\pi}{N|f''(x_0)|}}e^{Nf(x_0)} \text{ as } N \to \infty.$$

Laplace approximations becomes better as N grows.

In dimension $d > 1$

If $x \in \mathbb{R}^d$ then the Taylor expansion becomes

$$f(x0 = f(x_0) + (x - x_0)^T H(x - x_0)$$

where H is the hessian matrix of second derivatives of f. In that case it can be shown that

Lemma

$$\int e^{Nf(x)}x \sim \left(\frac{2\pi}{N}\right)^{d/2} | - H(x_0)|^{-1/2}e^{Nf(x_0)} \text{ as } N \to \infty.$$

Using Laplace approximation

Given model with $\theta = (\theta_1, \ldots, \theta_p)$

Step 1

Find mode of log-joint (=MAP) estimate of θ):

$$\theta^* = \text{argmax}_\theta \log p(\theta, y)$$

Step 2

Evaluate curvature of the log-joint at the mode

$$H = D^2 \log p(\theta^*, y)$$

is the Hessian matrix

Step 3

Obtain Gaussian approximation

$$\mathcal{N}(\theta|\mu, \Sigma), \quad \mu = \theta^*, \Sigma = H^{-1}.$$

Example

Suppose the y_i are iid $N(\mu, \sigma^2)$ with a flat prior on μ and on $\log \sigma$.
The posterior is

$$p(\mu, \sigma^2|y) \propto (\sigma^2)^{-\frac{n}{2}-1}e^{-\frac{(n-1)s^2+n(\bar{y}-\mu)^2}{2\sigma^2}}$$

where $\bar{y} = \frac{1}{n}\sum y_i$ and $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$.
Writig $\nu = \log \sigma$ we get

$$p(\mu, \nu|y) \propto f(\mu, \nu) = e^{-n\nu - \frac{(n-1)s^2+n(\bar{y}-\mu)^2}{2e^{2\nu}}}$$

Example

It is easy to check that

$$(\hat{\mu}, \hat{\nu}) = \text{mode}(\mu, \nu | y) = \left(\bar{y}, \frac{1}{2} \log \left(\frac{n-1}{n} s^2 \right) \right)$$

Second order derivatives are
$\frac{\partial^2}{\partial \mu^2} \log f = -ne^{-2\nu}$, $\frac{\partial^2}{\partial \mu \partial \nu} \log f = -2n(\bar{y} - \nu)e^{-2\nu}$ and

$$\frac{\partial^2}{\partial \nu^2} \log f = -2 \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{e^{2\nu}}$$

So that

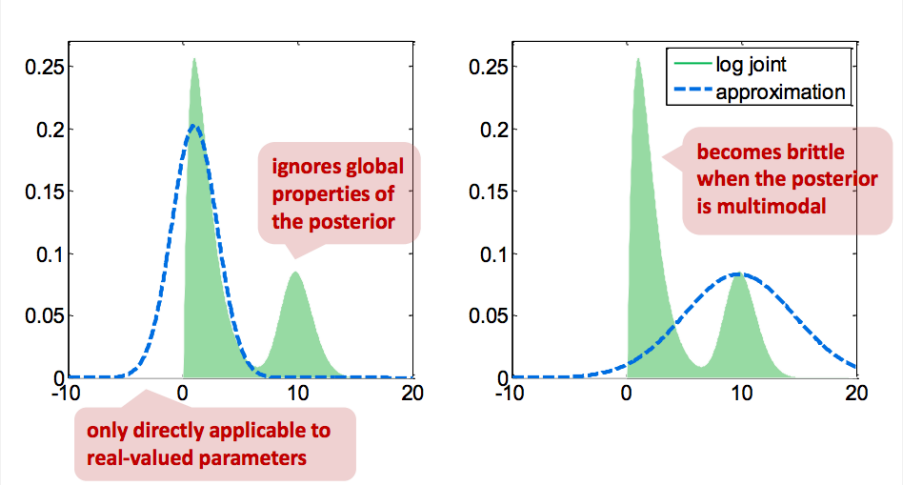
$$H(x_0) = \begin{pmatrix} \frac{n^2}{(n-1)s^2} & 0 \\ 0 & 2n \end{pmatrix}$$

and we have

$$\mu, \nu \sim N \left(\begin{pmatrix} \bar{y} \\ \frac{1}{2} \log \left(\frac{n-1}{n} s^2 \right) \end{pmatrix}, \begin{pmatrix} \frac{(n-1)s^2}{n^2} & 0 \\ 0 & \frac{1}{2n} \end{pmatrix} \right)$$

Limitations of Laplace method

The Laplace approximation is often too strong a simplification.



Laplace method for computing the marginal

$$\begin{aligned} P(x) &= \int P(x|\theta)\pi(\theta)d\theta \\ &= \int \exp\left\{ -N(-\frac{1}{N}\log P(x|\theta) - \frac{1}{N}\log \pi(\theta)) \right\} d\theta \end{aligned}$$

Define $h(\theta) = -\frac{1}{N}\log P(x|\theta) - \frac{1}{N}\log \pi(\theta)$ so that the integral we want to compute is of the form $\int \exp\{-Nh(\theta)\}\, d\theta$.

$$h(\theta) \approx h(\theta^*) - \tfrac{1}{2}|h''(\theta^*)|(\theta - \theta^*)^2$$

and we can approximate the integral as

$$\int e^{-Nh(\theta))} d\theta \approx e^{-Nh(\theta^*)} \int \exp\left\{ -\frac{N}{2}|h''(\theta^*)|(\theta - \theta^*)^2 \right\} d\theta$$

Comparing to a normal pdf we have

$$\int e^{-Nh(\theta)} dx \approx e^{-Nh(\theta^*)} (2\pi)^{\frac{1}{2}} |Nh''(\theta^*)|^{-\frac{1}{2}} = p(x|\theta^*)\pi(\theta^*)(2\pi)^{\frac{1}{2}}|Nh''(\theta^*)|^{-\frac{1}{2}}$$

Laplace's method

For a d-dimensional function the analogue of this result is

$$\int e^{Nf(x)} dx \approx e^{Nf(x_0)} (2\pi)^{\frac{d}{2}} N^{-\frac{d}{2}} |f''(x_0)|^{-\frac{1}{2}}$$

where $|f''(x_0)|$ is the determinant of the Hessian of the function evaluated at x_0.

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) takes the approximation one step further, essentially by minimizing the impact of the prior.

Firstly, the MAP estimate $\tilde{\theta}$ is replaced by the MLE $\hat{\theta}$, which is reasonable if the prior has a small effect.

Secondly, BIC only retains the terms that vary in N, since asymptotically the terms that are constant in N do not matter.

Dropping the constant terms we get,

$$\log P(\theta|\boldsymbol{X}) \approx \log P(\boldsymbol{X}|\hat{\theta}) - \frac{d}{2}\log N$$

Bayesian Information Criterion (BIC) - extra details

Why can we ignore the term $\frac{1}{2}\log|f''(\tilde{\theta})|^{-1}$?

Assume

- (as above) that we can ignore the prior i.e. $P(\theta) = 1$
- data points $X_1, \ldots, X_N$ are iid

Then

$$\begin{aligned} f(\tilde{\theta}) &= \frac{1}{N}\log P(\boldsymbol{X}|\theta)|_{\theta=\tilde{\theta}} \\ &= \frac{1}{N}\sum_{i=1}^{N}\log P(X_i|\theta)|_{\theta=\tilde{\theta}} \end{aligned}$$

The thing to notice about this term is that it is now the average log-likelihood.

Bayesian Information Criterion (BIC) - extra details

Now consider random variables $X_i = \log P(X_i|\theta)$ and apply WLLN

$$f(\tilde{\theta}) \quad \to \quad \mathbb{E}[\log P(X_i|\theta)]|_{\theta=\tilde{\theta}}$$

So the (m, n)th element of $f''(\tilde{\theta})$ is

$$\left.\frac{\partial^2 \mathbb{E}[\log P(X_i|\theta)]}{\partial\theta_m \partial\theta_n}\right|_{\theta=\tilde{\theta}}$$

and these are constants i.e expected log-likelihoods for a single data point, so $|f''(\tilde{\theta})|$ is constant, and can be ignored in the BIC approximation.

Variational Bayes

The idea of VB is to find an approximation $Q(\theta)$ to a given posterior distribution $P(\theta|\boldsymbol{X})$. That is

$$Q(\theta) \approx P(\theta|X)$$

where θ is the vector of parameters.

We then use $Q(\theta)$ to approximate the marginal likelihood. In fact, what we do is find a lower bound for the marginal likelihood.

Question How to find a good approximate posterior $Q(\theta)$?

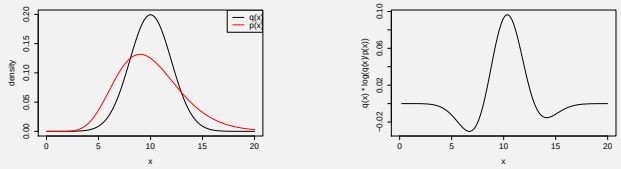
Kullback-Liebler (KL) divergence

The strategy we take is to find a distribution $Q(\theta)$ that minimizes a measure of 'distance' between $Q(\theta)$ and the posterior $P(\theta|\boldsymbol{X})$.

Definition

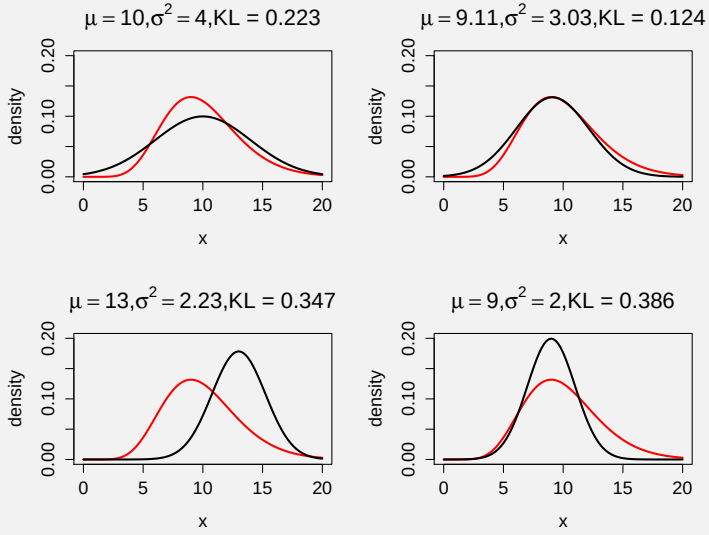
The Kullback-Leibler divergence $KL(q|p)$ between two distributions $q(x)$ and $p(x)$ is

$$KL(q|p) = \int \log\left[\frac{q(x)}{p(x)}\right] q(x)dx$$



Exercise $KL(q|p) \geq 0$ and $KL(q|p) = 0$ iff $q = p$.

$N(\mu, \sigma^2)$ approximations to a Gamma(10,1)



We consider the KL divergence betweem $Q(\theta)$ and $P(\theta|\boldsymbol{X})$

$$\begin{aligned} KL(Q(\theta)||P(\theta|\boldsymbol{X})) &= \int \log\left[\frac{Q(\theta)}{P(\theta|\boldsymbol{X})}\right] Q(\theta)d\theta \\ &= \int \log\left[\frac{Q(\theta)P(\boldsymbol{X})}{P(\theta,\boldsymbol{X})}\right] Q(\theta)d\theta \\ &= \log P(\boldsymbol{X}) - \int \log\left[\frac{p(\theta,\boldsymbol{X})}{Q(\theta)}\right] Q(\theta)d\theta \end{aligned}$$

The log marginal likelihood can then be written as

$$\log P(\boldsymbol{X}) = F(Q(\theta)) + KL(Q(\theta)||P(\theta|\boldsymbol{X})) \quad (1)$$

where $F(Q(\theta)) = \int \log\left[\frac{P(\theta,\boldsymbol{D}|M)}{Q(\theta)}\right] Q(\theta)d\theta$.

Note Since $KL(q|p) \geq 0$ we have that

$$\log P(\boldsymbol{X}) \geq F(Q(\theta))$$

so that $F(Q(\theta))$ is a lower bound on the log-marginal likelihood.

The mean field approximation

We now need to ask what form that $Q(\theta)$ should take?

The most widely used approximation is known as the **mean field approximation** and assumes only that the approximate posterior has a factorized form

$$Q(\theta) = \prod_i Q(\theta_i)$$

The VB algorithm iteratively maximises $F(Q(\theta))$ with respect to the free distributions, $Q(\theta_i)$, which is coordinate ascent in the function space of variational distributions.

We refer to each $Q(\theta_i)$ as a VB component.

We update each component $Q(\theta_i)$ in turn keeping $Q(\theta_j)$ $j \neq i$ fixed.

VB components

Lemma

The VB components take the form

$$\log Q(\theta_i) \quad = \quad \mathbb{E}_{Q(\theta_{-i})}\Big(\log P(\boldsymbol{X}, \theta)\Big) + const$$

Proof Writing $Q(\theta) = Q(\theta_i)Q(\theta_{-i})$ where $\theta_{-i} = \theta \backslash \theta_i$, the lower-bound can be re-written as

$$\begin{aligned} F(Q(\theta)) &= \int \log\Big[\frac{P(\theta, \boldsymbol{X})}{Q(\theta)}\Big] Q(\theta) d\theta \\ &= \int\int \log\Big[\frac{P(\theta, \boldsymbol{X})}{Q(\theta_i)Q(\theta_{-i})}\Big] Q(\theta_i)Q(\theta_{-i}) d\theta_i d\theta_{-i} \\ &= \int Q(\theta_i)\Big[\int \log P(\boldsymbol{X}, \theta) Q(\theta_{-i}) d\theta_{-i}\Big] d\theta_i \\ &\quad - \int\int Q(\theta_i)Q(\theta_{-i}) \log Q(\theta_i) d\theta_{-i} d\theta_i \\ &\quad - \int\int Q(\theta_i)Q(\theta_{-i}) \log Q(\theta_{-i}) d\theta_{-i} d\theta_i \end{aligned}$$

$$\begin{aligned} F(Q(\theta)) &= \int Q(\theta_i)\Big[\int \log P(\boldsymbol{X}, \theta) Q(\theta_{-i}) d\theta_{-i}\Big] d\theta_i \\ &\quad - \int\int Q(\theta_i)Q(\theta_{-i}) \log Q(\theta_i) d\theta_{-i} d\theta_i - \int\int Q(\theta_i)Q(\theta_{-i}) \log Q(\theta_{-i}) d\theta_{-i} d\theta_i \\ &= \int Q(\theta_i)\Big[\int \log P(\boldsymbol{X}, \theta) Q(\theta_{-i}) d\theta_{-i}\Big] d\theta_i \\ &\quad - \int Q(\theta_i) \log Q(\theta_i) d\theta_i - \sum_{j \neq i} \int Q(\theta_j) \log Q(\theta_j) d\theta_j \end{aligned}$$

If we let $Q^*(\theta_i) = \frac{1}{Z} \exp\Big[\int \log P(\boldsymbol{D}, \theta | M) Q(\theta_{-i}) d\theta_{-i}\Big]$ where Z is a normalising constant and write $H(Q(\theta_j)) = -\int Q(\theta_j) \log Q(\theta_j) d\theta_j$ as the entropy of $Q(\theta_j)$ then

$$\begin{aligned} F(Q(\theta)) &= \int Q(\theta_i) \log \frac{Q^*(\theta_i)}{Q(\theta_i)} d\theta_i + \log Z + \sum_{j \neq i} H(Q(\theta_j)) \\ &= -KL(Q(\theta_i) || Q^*(\theta_i)) + \log Z + \sum_{j \neq i} H(Q(\theta_j)) \end{aligned}$$

$$F(Q(\theta)) \quad = \quad -KL(Q(\theta_i) || Q^*(\theta_i)) + \log Z + \sum_{j \neq i} H(Q(\theta_j))$$

We then see that $F(Q(\theta))$ is maximised when $Q(\theta_i) = Q^*(\theta_i)$ as this choice minimises the Kullback-Liebler divergence term.

Thus the update for $Q(\theta_i)$ is given by

$$Q(\theta_i) \quad \propto \quad \exp\Big[\int \log P(\boldsymbol{X}, \theta) Q(\theta_{-i}) d\theta_{-i}\Big]$$

or

$$\log Q(\theta_i) \quad = \quad \mathbb{E}_{Q(\theta_{-i})}\Big(\log P(\boldsymbol{X}, \theta)\Big) + const$$

VB algorithm

This implies a straightforward algorithm for variational inference:

1. Initialize all approximate posteriors $Q(\theta) = Q(\mu)Q(\tau)$, e.g., by setting them to their priors.
2. Cycle over the parameters, revising each given the current estimates of the others.
3. Loop until convergence.

Convergence is checked by calculating the VB lower bound at each step i.e.

$$F(Q(\theta)) = \int \log\Big[\frac{P(\theta, \boldsymbol{X})}{Q(\theta)}\Big] Q(\theta) d\theta$$

The precise form of this term needs to be derived, and can be quite tricky.

Example 1

Consider applying VB to the hierarchical model

$$X_i \sim N(\mu, \tau^{-1}) \quad i = 1, \ldots, P, \quad \mu \sim N(m, (\tau\beta)^{-1}) \quad \tau \sim \Gamma(a, b)$$

Note We are using a prior of the form $\pi(\tau, \mu) = \pi(\mu|\tau)\pi(\tau)$.

Let $\theta = (\mu, \tau)$ and assume $Q(\theta) = Q(\mu)Q(\tau)$.

We will use notation $\langle \theta_i \rangle = \mathbb{E}_{Q(\theta_{-i})}\theta_i$.

The log joint density is

$$\begin{aligned} \log P(X, \theta) &= \frac{P}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{P}(X_i - \mu)^2 + \frac{1}{2}\log\tau - \frac{\tau\beta}{2}(\mu - m)^2 \\ &\quad +(a-1)\log\tau - b\tau + K \end{aligned}$$

$$\begin{aligned} \log P(X, \theta) &= \frac{P}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{P}(X_i - \mu)^2 + \frac{1}{2}\log\tau - \frac{\tau\beta}{2}(\mu - m)^2 \\ &\quad +(a-1)\log\tau - b\tau + K \end{aligned}$$

We can derive the VB updates one at a time. We start with $Q(\mu)$. **Note** We just need to focus on terms involving μ.

$$\begin{aligned} \log Q(\mu) &= \mathbb{E}_{Q(\tau)}\Big(\log P(X, \theta)\Big) + C \\ &= -\frac{\langle\tau\rangle}{2}\Big(\sum_{i=1}^{P}(X_i - \mu)^2 - \beta(\mu - m)^2\Big) + C \end{aligned}$$

where $\langle\tau\rangle = \mathbb{E}_{Q(\tau)}(\tau)$. We will be able to determine $\langle\tau\rangle$ when we derive the other component of the approximate density, $Q(\tau)$.

We can see this log density has the form of a normal distribution

$$\begin{aligned} \log Q(\mu) &= -\frac{\langle\tau\rangle}{2}\Big(\sum_{i=1}^{P}(D_i - \mu)^2 - \beta(\mu - m)^2\Big) + C \\ &= -\frac{\beta'}{2}(\mu - m')^2 \end{aligned}$$

where

$$\begin{aligned} \beta' &= (\beta + P)\langle\tau\rangle \\ m' &= \beta'^{-1}\Big(\beta m + \sum_{i=1}^{P} X_i\Big) \end{aligned}$$

Thus $Q(\mu) = N(\mu|m', \beta'^{-1})$.

$$\begin{aligned} \log P(X, \theta) &= \frac{P}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{P}(X_i - \mu)^2 + \frac{1}{2}\log\tau - \frac{\tau\beta}{2}(\mu - m)^2 \\ &\quad +(a-1)\log\tau - b\tau + K \end{aligned}$$

The second component of the VB approximation is derived as

$$\begin{aligned} \log Q(\tau) &= \mathbb{E}_{Q(\mu)}\Big(\log P(X, \theta)\Big) + C \\ &= ((P+1)/2 + a - 1)\log\tau - \frac{\tau}{2}\left\langle\sum_{i=1}^{P}(X_i - \mu)^2\right\rangle \\ &\quad -\frac{\tau}{2}\Big\langle\beta(\mu - m)^2\Big\rangle + C \end{aligned}$$

We can see this log density has the form of a gamma distribution

$$\begin{aligned}\log Q(\tau) &= ((P+1)/2 + a - 1)\log\tau - \frac{\tau}{2}\left\langle \sum_{i=1}^{P}(X_i-\mu)^2 \right\rangle \\ &\quad -\frac{\tau}{2}\left\langle \beta(\mu-m)^2 \right\rangle + C\end{aligned}$$

which is $\Gamma(a', b')$ where

$$\begin{aligned}a' &= a + (P+1)/2 \\ b' &= b + \frac{1}{2}\Big(\sum_{i=1}^{P} X_i^2 - 2\langle\mu\rangle + P\langle\mu^2\rangle\Big) + \frac{\beta}{2}\Big(m^2 - 2\langle\mu\rangle + \langle\mu^2\rangle\Big)\end{aligned}$$

So overall we have

1. $Q(\mu) = N(\mu|m', \beta'^{-1})$ where

$$\beta' = (\beta + P)\langle\tau\rangle \quad (2)$$

$$m' = \beta'^{-1}\Big(\beta m + \sum_{i=1}^{P} D_i\Big) \quad (3)$$

2. $Q(\tau) = \Gamma(\tau|a', b')$ where

$$\begin{aligned}a' &= a + (P+1)/2 \\ b' &= b + \frac{1}{2}\Big(\sum_{i=1}^{P} X_i^2 - 2\langle\mu\rangle + P\langle\mu^2\rangle\Big) + \frac{\beta}{2}\Big(m^2 - 2\langle\mu\rangle + \langle\mu^2\rangle\Big)\end{aligned}$$

To calculate these we need

$$\begin{aligned}\langle\tau\rangle &= \frac{a'}{b'} \\ \langle\mu\rangle &= m' \\ \langle\mu^2\rangle &= \beta'^{-1} + m'^2\end{aligned}$$

Example 1

For this model the exact posterior was calculate in Lecture 6.

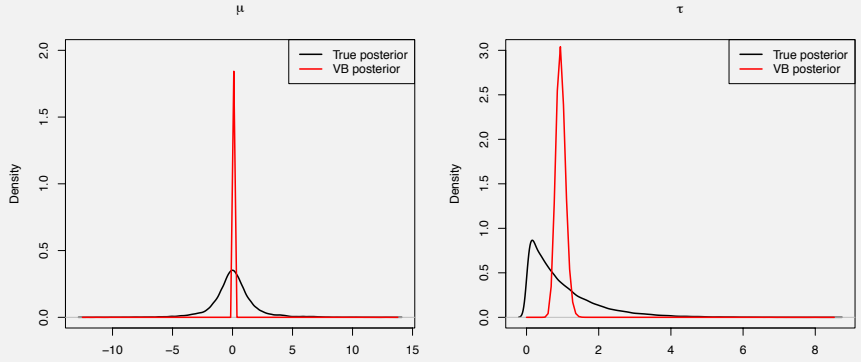
$$\pi(\tau, \mu \mid \boldsymbol{X}) \propto \tau^{\alpha'-1}\exp\left[-\tau\left\{b' + \frac{\beta'}{2}(m'-\mu)^2\right\}\right]$$

where

$$\begin{aligned}a' &= a + \frac{P+1}{2} \\ b' &= b + \frac{1}{2}\sum(X_i - \bar{X})^2 + \frac{1}{2}\cdot\frac{P\beta}{P+\beta}(\bar{X} - m)^2 \\ \beta' &= \beta + P \\ m' &= \beta'^{-1}(\beta m + \sum_{i=1}^{P} X_i)\end{aligned}$$

We note some similarity between the VB updates and the true posterior parameters.

We can compare the true and VB posterior when applied to a real dataset. We see that VB approximations **underestimate** posterior variances.



General comments

- The property of VB underestimating the variance in the posterior is a general feature of the method, when there exists correlation between the θ_i's in the posterior, which is usually the case. This may not be important if the purpose of inference is model comparison i.e. comparing the approximate marginal likelihoods between models.
- VB is often much, much faster to implement than MCMC or other sampling based methods.
- The VB updates and lower bound can be tricky to derive, and sometime further approximation is needed.
- The VB algorithm will find a local mode of the posterior, so care should be taken when the posterior is thought/known to be multi-modal.