We have the model

$$y_t = X_t^{'}\beta + \epsilon_t,$$

where $Y$ is a $T \times 1$ random vector, $X$ is a $k \times T$ matrix of covariates, $\beta$ is a $k \times 1$ matrix of coefficients and $\epsilon_t$ is iid N$(0, \sigma^2)$.

We are interested in the posterior density

$$p(\sigma^2, \beta | y) = \frac{p(y|\sigma^2, \beta)p(\sigma^2, \beta)}{p(y)}$$

Taking $p(\sigma^2, \beta) \propto \frac{1}{\sigma^2}$,

$$\begin{aligned} p(\sigma^2, \beta | y) &\propto \prod_{t=1}^{T} \frac{1}{\sigma} exp\left\{ \frac{-1}{2\sigma^2}(y_t - X_t^{'}\beta)^2 \right\} \frac{1}{\sigma^2} \\ &= \sigma^{-(T+2)} exp\left\{ \frac{-1}{2\sigma^2} \sum_{t=1}^{T} (y_t - X_t^{'}\beta)^2 \right\} \\ &= \sigma^{-(T+2)} exp\left\{ \frac{-1}{2\sigma^2}(T-k)\hat{\sigma}^2 \right\} exp\left\{ \frac{-1}{2\sigma^2}(\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \right\} \quad (1) \end{aligned}$$

Where $\hat{\beta} = (X'X)^{-1}X'Y$, the OLS estimator of $\beta$ and $\hat{\sigma}^2 = (y - X\hat{\beta})'(Y - X\hat{\beta})/(T-k)$ is an unbiases estimator of $\sigma^2$.

We approximate $p(\beta, \sigma^2|y)$ with $Q(\beta, \sigma^2) = Q(\beta)Q(\sigma^2)$, the factorisable distribution that minimises the Kullback - Leibler divergence between $p(\beta, \sigma^2|y)$ and $Q(\beta, \sigma^2)$. It can be shown that the distributions $Q(\beta)$ and $Q(\sigma^2)$ that minimise this are:

$$ln(Q(\beta)) = E_{\sigma^2}\left[ln(y, \sigma^2, \beta))\right] + C \quad (2)$$

and

$$ln(Q(\sigma^2)) = E_{\beta}\left[ln(p(y, \sigma^2, \beta))\right] + C \quad (3)$$

Substituting (1) into (2) and ignoring the terms that do not depend on $\beta$, we get

$$ln(Q(\beta)) = E_{\sigma^2}\left[\frac{-1}{2\sigma^2}(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right] + C$$

.

Recognising the kernel of a normal distribution, we see that

$$Q(\beta) \sim N(\mu = \hat{\beta}, \lambda = (X'X)^{-1}E[\sigma^{-2}]^{-1})$$

Similarly substituting (1) into (3) results in

$$\begin{aligned} ln(Q(\sigma^2)) &= -(T+2)ln(\sigma) - \frac{1}{2\sigma^2}\Big((T-k)\hat{\sigma}^2 + E_\beta[(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]\Big) + C \\ &= -(T+2)ln(\sigma) - \frac{1}{2\sigma^2}\big((T-k)\hat{\sigma}^2 + tr(X'XVar(\beta))\big) + C \end{aligned}$$

Recoginising the kernel of an inverse gamma distribution, we see that

$$Q(\sigma^2) \sim Inv.Gamma(shape = (T+1)/2 = a, scale = \frac{(T-k)\hat{\sigma}^2 + tr(X'X\lambda)}{2} = b)$$

Therefore $1/\sigma^2 \sim Gamma(shape = a, rate = b)$ so $E[\sigma^{-2}] = a/b$, and substituting this into $\lambda$ gives us:

$$\lambda = (X'X)^{-1}b/a$$

and so

$$\begin{aligned} b &= \frac{(T-k)\hat{\sigma}^2 + tr(X'X(X'X)^{-1})b/a}{2} \\ &= \frac{(T-k)\hat{\sigma}^2}{2} + \frac{kb}{T+1} \\ &= \frac{(T-k)\hat{\sigma}^2}{2(1-k/(T+1))} \end{aligned}$$

Figure 1 shows the results of variational bayes compared to the exact posterior for the linear regression:

$$y_i = b_1x_{1i} + b_2x_{2i} + \epsilon_i$$

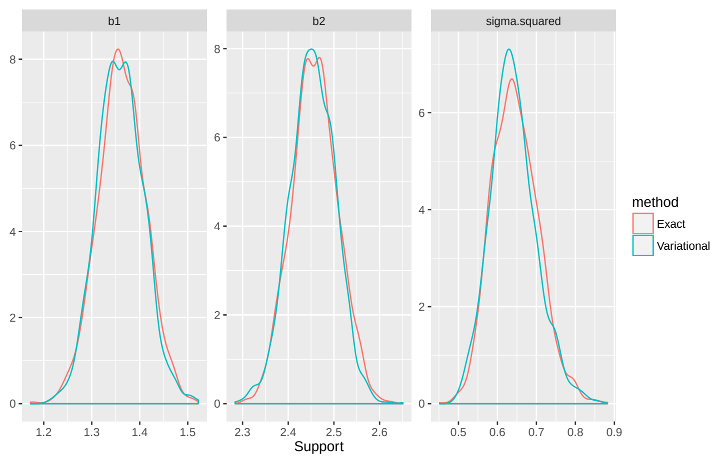


Figure 1: Variational Bayes compared to the exact posterior. It took three iterations for the parameters to converge.