# Variational Bayes Updates for Normal Linear Regression

Santiago Montoya-Blandón

ECON 5120: Bayesian Data Analysis
Adam Smith Business School, University of Glasgow

## Setup

Recall that the Bayesian linear regression framework requires a likelihood and a set of priors. We have an outcome $\boldsymbol{y} = (y_1, \ldots, y_n)$ that we aim to predict using a set of covariates or features $\boldsymbol{X}$ arranged in an $n \times p$ matrix such that

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}.$$

We do not make restrictions on the size of $p$ (the amount of features). Indeed, variational techniques were introduced into learning procedures as a way to obtain reliable and scalable inference for models with large $p$, potentially with $p > n$. As the likelihood for this model we will use a classical normal linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},\ \boldsymbol{u} \sim \mathcal{N}_n(\boldsymbol{0}_n, \sigma^2 \boldsymbol{I}_n) \implies \boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2; \boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)\,, \tag{1}$$

where $\boldsymbol{0}_n$ is a $n$-dimensional vector of zeros and $I_n$ is the $n \times n$ identity matrix. We will use standard independent priors for $\boldsymbol{\beta}$ and $\sigma^2$, such that the joint prior is $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2)$ with

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\underline{\boldsymbol{\beta}}, \underline{\boldsymbol{B}}) \quad \text{and} \quad \sigma^2 \sim \text{Inv-Gamma}(\underline{\alpha}/2, \underline{\delta}/2) \tag{2}$$

Using Bayes' rule (posterior proportional to likelihood times prior), we can write the joint posterior $p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2; \boldsymbol{X})p(\boldsymbol{\beta}, \sigma^2)$ without proportionality constants as

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\} \\ \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})^\top \underline{\boldsymbol{B}}^{-1}(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})\right\} \\ \times (\sigma^2)^{-\underline{\alpha}/2-1} \exp\left\{-\frac{\underline{\delta}}{2\sigma^2}\right\} \end{aligned} \tag{3}$$

We will leave this expression as is, and use the terms we need to build the respective conditionals that we require for the optimal variational approximations. Before proceeding,

I want to remind you of a very useful trick known as "completing the square". For any $p$-dimensional vectors $\boldsymbol{x}$, $\boldsymbol{b}$, and $p \times p$ matrix $\boldsymbol{B}$, we have

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{b})^\top \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{b}) &= \boldsymbol{x}^\top \boldsymbol{B}^{-1}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{B}^{-1}\boldsymbol{x} - \boldsymbol{x}^\top \boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{b}^\top \boldsymbol{B}^{-1}\boldsymbol{b} \\ &= \boldsymbol{x}^\top \boldsymbol{B}^{-1}\boldsymbol{x} - 2\boldsymbol{b}^\top \boldsymbol{B}^{-1}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{B}^{-1}\boldsymbol{b}\end{aligned}$$

This means we can complete a quadratic form as follows:

$$\begin{aligned}Q &:= \boldsymbol{x}^\top \boldsymbol{B}^{-1}\boldsymbol{x} - 2\boldsymbol{b}^\top \boldsymbol{B}^{-1}\boldsymbol{x} \\ \implies Q &= (\boldsymbol{x}-\boldsymbol{b})^\top \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{b}) - \boldsymbol{b}^\top \boldsymbol{B}^{-1}\boldsymbol{b}\end{aligned} \tag{4}$$

# Variational Approximations

## General case with mean-field approximations

Recall that to implement variational Bayes we require a class of approximations $\mathcal{Q}$. We usually use *mean-field approximations*, which for this problem means we can define

$$\mathcal{Q} := \left\{ q(\boldsymbol{\beta}, \sigma^2) : q(\boldsymbol{\beta}, \sigma^2) = q(\boldsymbol{\beta}) \cdot q(\sigma^2) \right\}$$

Note that we are not constraining the forms of $q(\boldsymbol{\beta})$ or $q(\sigma^2)$ at this stage, we are only proposing that they will be independent. This means that rather than assuming particular functional forms for these distributions, the distributional forms for $q(\boldsymbol{\beta})$ and $q(\sigma^2)$ will be obtained as a consequence of the optimal approximation. Recall the optimal approximation in the general case is given as

$$\begin{aligned}q_j^*(\theta_j) &\propto \exp\left\{ \mathrm{E}_{-j}[\log p(\theta_j \mid \boldsymbol{\theta}_{-j}; \boldsymbol{y})] \right\} \\ &\exp\left\{ \int \log p(\theta_j \mid \boldsymbol{\theta}_{-j}; \boldsymbol{y}) \prod_{\ell \neq j} q_\ell^*(\theta_\ell) d\boldsymbol{\theta}_{-j} \right\}\end{aligned} \tag{5}$$

where $p(\theta_j \mid \boldsymbol{\theta}_{-j}; \boldsymbol{y})$ is the conditional posterior of $\theta_j$ conditional on all other parameters excluding $\theta_j$, denoted as $\boldsymbol{\theta}_{-j}$, across $j = 1, \ldots, p$. Note that the expectation in (5) is obtained with respect to the optimal approximations of the other parameters $q_1^*(\theta_1), \ldots, q_{j-1}^*(\theta_{j-1})$, $q_{j+1}^*(\theta_{j+1}), \ldots, q_p^*(\theta_p)$. In addition, notice that if we can write $\mathrm{E}_{-j}[\log p(\theta_j \mid \boldsymbol{\theta}_{-j}; \boldsymbol{y})]$ as a function of $\theta_j$, say $f(\theta_j)$, and a constant with respecto to $\theta_j$, then (5) implies we would have

$$q_j^*(\theta_j) \propto \exp\left\{ f(\theta_j) + \text{const.} \right\} \propto \exp\{ f(\theta_j) \}$$

This means we can ignore all constant terms that appear in the $\mathrm{E}_{-j}[\log p(\theta_j \mid \boldsymbol{\theta}_{-j}; \boldsymbol{y})]$ terms.

## Variational approximation for Bayesian linear regression with independent priors

In our normal linear regression case, the optimal approximations take the shape

$$\begin{aligned}q^*(\boldsymbol{\beta}) &\propto \exp\left\{ \mathrm{E}_{q^*(\sigma^2)}[\log p(\boldsymbol{\beta} \mid \sigma^2; \boldsymbol{y}, \boldsymbol{X})] \right\} \\ q^*(\sigma^2) &\propto \exp\left\{ \mathrm{E}_{q^*(\beta)}[\log p(\sigma^2 \mid \boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})] \right\}\end{aligned} \tag{6}$$

We will show the independence assumption of the mean-field approximation is all we need to obtain the specific distributional forms that solve (6). Start with $q^*(\boldsymbol{\beta})$, which requires the conditional posterior $p(\boldsymbol{\beta} \mid \sigma^2; \boldsymbol{y}, \boldsymbol{X})$. To obtain this posterior, we can simply keep all terms from (3) that contain $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}) \propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})^\top \underline{\boldsymbol{B}}^{-1}(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})\right]\right\} \tag{7}$$

Recall the notation $f(x) \propto g(x)$ means $f(x) = c \cdot g(x)$ for any functions $f$ and $g$ and some scalar proportionality constant $c$ that is independent of $x$. This means $\log f(x) = \log(c) + \log g(x) = \log g(x) + \text{const.}$ follows as a consequence. Taking logarithms and expectations with respect to $q^*(\sigma^2)$ yields

$$\begin{aligned}
\mathrm{E}_{q^*(\sigma^2)}[\log p(\boldsymbol{\beta} \mid \sigma^2; \boldsymbol{y}, \boldsymbol{X})] &= -\frac{1}{2}\,\mathrm{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right. \\
&\qquad \left. + (\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})^\top \underline{\boldsymbol{B}}^{-1}(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})\right] + \text{const.} \\
&= -\frac{1}{2}\,\mathrm{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\left(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}\right) \\
&\quad - \frac{1}{2}\left(\boldsymbol{\beta}^\top\underline{\boldsymbol{B}}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\underline{\boldsymbol{B}}^{-1}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{\beta}}^\top\underline{\boldsymbol{B}}^{-1}\underline{\boldsymbol{\beta}}\right) + \text{const.} \\
&= -\frac{1}{2}\left[\boldsymbol{\beta}^\top\left(\mathrm{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\boldsymbol{X}^\top\boldsymbol{X} + \underline{\boldsymbol{B}}^{-1}\right)\boldsymbol{\beta}\right. \\
&\qquad \left. - 2\boldsymbol{\beta}^\top\left(\mathrm{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\boldsymbol{X}^\top\boldsymbol{y} + \underline{\boldsymbol{B}}^{-1}\underline{\boldsymbol{\beta}}\right)\right] + \text{const.} \\
&= -\frac{1}{2}\left(\boldsymbol{\beta}^\top\bar{\boldsymbol{B}}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\bar{\boldsymbol{B}}^{-1}\bar{\boldsymbol{\beta}}\right) + \text{const.}
\end{aligned} \tag{8}$$

where we defined

$$\bar{\boldsymbol{B}} \coloneqq \left(\mathrm{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\boldsymbol{X}^\top\boldsymbol{X} + \underline{\boldsymbol{B}}^{-1}\right)^{-1} \quad \text{and} \quad \bar{\boldsymbol{\beta}} \coloneqq \bar{\boldsymbol{B}}\left(\mathrm{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\boldsymbol{X}^\top\boldsymbol{y} + \underline{\boldsymbol{B}}^{-1}\underline{\boldsymbol{\beta}}\right) \tag{9}$$

Note that while we do not know how to solve for the term $\mathrm{E}_{q^*(\sigma^2)}[1/\sigma^2]$ yet, we do know it will not depend on $\boldsymbol{\beta}$. This means that both $\bar{\boldsymbol{B}}$ and $\bar{\boldsymbol{\beta}}$ are constants with respect to $\boldsymbol{\beta}$. We can then complete the square on $\boldsymbol{\beta}$ in equation (8) using the result in (4) to obtain[1]

$$\mathrm{E}_{q^*(\sigma^2)}[\log p(\boldsymbol{\beta} \mid \sigma^2; \boldsymbol{y}, \boldsymbol{X})] = -\frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top\bar{\boldsymbol{B}}^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + \text{const.}$$

As we can ignore all constant factors in these expectations, we can then write

$$q^*(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top\bar{\boldsymbol{B}}^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right\}$$

[1] We set $\boldsymbol{x} = \boldsymbol{\beta}$, $\boldsymbol{b} = \bar{\boldsymbol{\beta}}$, and $\boldsymbol{B} = \bar{\boldsymbol{B}}$ when completing the square, using the fact that $\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{B}}^{-1}\bar{\boldsymbol{B}}\bar{\boldsymbol{\beta}}$.

Even though we do not yet have a closed-form expression for $\mathrm{E}_{q^*(\sigma^2)}[1/\sigma^2]$, we can recognize the previous equation as the kernel (or quasi-density) of a multivariate normal distribution on $\boldsymbol{\beta}$ with mean vector $\bar{\boldsymbol{\beta}}$ and variance-covariance matrix $\bar{\boldsymbol{B}}$. That is, we have found

$$q^*(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{B}}) \implies \boldsymbol{\beta} \mid y, X \sim_{q^*} \mathcal{N}_p(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{B}}) \tag{10}$$

where the notation "$\sim_{q^*}$" emphasizes that this is not the true exact posterior, but rather the posterior under the optimal variational approximation. Let us proceed with finding $q^*(\sigma^2)$ and we will go back and calculate the elements in (9) once this is done. As before, we start from the conditional posterior $p(\sigma^2 \mid \boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})$, found by keeping all terms from (3) that contain $\sigma^2$:

$$p(\sigma^2 \mid \boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^2)^{-(n+\underline{\alpha})/2-1} \times \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \underline{\delta}\right]\right\} \tag{11}$$

Taking logarithms and expectations with respect to $q^*(\boldsymbol{\beta})$ we obtain

$$\begin{aligned}
\mathrm{E}_{q^*(\beta)}[\log p(\sigma^2 \mid \boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})] &= -\frac{1}{2\sigma^2}\left\{\mathrm{E}_{q^*(\beta)}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right] + \underline{\delta}\right\} \\
&\quad + \left(-\frac{n+\underline{\alpha}}{2} - 1\right)\log(\sigma^2) + \text{const.} \\
&= -\frac{\bar{\delta}}{2\sigma^2} + \left(-\frac{\bar{\alpha}}{2} - 1\right)\log(\sigma^2) + \text{const.}
\end{aligned} \tag{12}$$

where we have defined

$$\bar{\alpha} \coloneqq \underline{\alpha} + n \quad \text{and} \quad \bar{\delta} \coloneqq \underline{\delta} + \mathrm{E}_{q^*(\beta)}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right] \tag{13}$$

As before, we do not have closed-form expressions for these posterior hyperparameters yet, but we do know they are independent of $\sigma^2$. Ignoring constant terms, we now obtain

$$q^*(\sigma^2) \propto (\sigma^2)^{-\bar{\alpha}/2-1} \exp\left\{-\frac{\bar{\delta}}{2\sigma^2}\right\}$$

We recognize this as the kernel of another inverse Gamma distribution, meaning we obtained

$$q^*(\sigma^2) = \text{Inv-Gamma}\left(\sigma^2 \,\middle|\, \frac{\bar{\alpha}}{2}, \frac{\bar{\delta}}{2}\right) \implies \sigma^2 \mid y, X \sim_{q^*} \text{Inv-Gamma}\left(\frac{\bar{\alpha}}{2}, \frac{\bar{\delta}}{2}\right) \tag{14}$$

Combining (10) and (14), we have the full variational approximation

$$q^*(\boldsymbol{\beta}, \sigma^2) = q^*(\boldsymbol{\beta}) \cdot q^*(\sigma^2) \tag{15}$$

In order to make this solution operational, we need to solve for the expectations $\mathrm{E}_{q^*(\sigma^2)}[1/\sigma^2]$ and $\mathrm{E}_{q^*(\beta)}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]$. We now know the form of the individual optimal approximations, so this can be accomplished analytically. For $\mathrm{E}_{q^*(\sigma^2)}[1/\sigma^2]$, note that as $\sigma^2$ is distributed as inverse Gamma under the variational approximation, $1/\sigma^2$ is distributed as

Gamma with shape $\bar{\alpha}/2$ and *rate* $\bar{\delta}/2$. Using known properties of the Gamma distribution, this means

$$\mathrm{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right] = \frac{\bar{\alpha}}{\bar{\delta}} \tag{16}$$

For $\mathrm{E}_{q^*(\beta)}\left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]$, note that $\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X} \sim_{q^*} \mathcal{N}_p(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{B}})$ implies $\boldsymbol{\varepsilon} \mid \boldsymbol{y}, \boldsymbol{X} \sim_{q^*} \mathcal{N}_n(\boldsymbol{y}-\boldsymbol{X}\bar{\boldsymbol{\beta}}, \boldsymbol{X}\bar{\boldsymbol{B}}\boldsymbol{X}^\top)$, where $\boldsymbol{\varepsilon} \coloneqq \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}$. We again invoke known results to find[2]

$$\mathrm{E}_{q^*(\beta)}\left[\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\right] = \mathrm{tr}(\bar{\boldsymbol{B}}\boldsymbol{X}^\top\boldsymbol{X}) + (\boldsymbol{y}-\boldsymbol{X}\bar{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\bar{\boldsymbol{\beta}}) \tag{17}$$

where $\mathrm{tr}(\boldsymbol{A})$ is the trace operator that equals the sum of the elements of the diagonal for any matrix $\boldsymbol{A}$.

## Evidence Lower Bound

Finally, we note how to evaluate the quality of the variational approximation. In the general case, recall we are using the *evidence lower bound* (ELBO) as our objective function to evaluate any approximation $q \in \mathcal{Q}$:

$$\mathrm{elbo}(q) = \mathrm{E}_q\left[\log\left(\frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)\right] \tag{18}$$

For general approximations, if we have a way to sample from the approximated posterior $q(\boldsymbol{\theta})$, we can estimate (18) using these posterior draws. Under the independence of the mean-field approximation, we can further decompose our objective as

$$\mathrm{elbo}(q) = \mathrm{E}_q\left[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})\right] + \mathrm{E}_q\left[\log p(\boldsymbol{\theta})\right] - \sum_{j=1}^{p}\mathrm{E}_{q_j}\left[\log q_j(\theta_j)\right] \tag{19}$$

Replacing the likelihood (1) and prior structure (2) into (19) provides the ELBO formula for Bayesian linear regression. Rather than evaluating all approximations $q \in \mathcal{Q}$, we can focus on computing the ELBO for the *optimal variational approximation* $q^*(\boldsymbol{\beta}, \sigma^2)$:

$$\begin{aligned}\mathrm{elbo}(q^*) = \mathrm{E}_{q^*}\left[\log p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2; \boldsymbol{X})\right] + \mathrm{E}_{q^*(\beta)}\left[\log p(\boldsymbol{\beta})\right] + \mathrm{E}_{q^*(\sigma^2)}\left[\log p(\sigma^2)\right] \\ - \mathrm{E}_{q^*(\beta)}\left[\log q^*(\boldsymbol{\beta})\right] - \mathrm{E}_{q^*(\sigma^2)}\left[\log q^*(\sigma^2)\right]\end{aligned} \tag{20}$$

Now that we know the optimal approximations and that they come from simple families, we can analytically compute every term in (20). These are given in the following expressions,

[2]See `https://en.wikipedia.org/wiki/Quadratic_form_(statistics)#Expectation` for a full derivation.

obtained without proof as they use more known results.[3]

$$
\begin{aligned}
\mathrm{E}_{q^*}\left[\log p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2; \boldsymbol{X})\right] &= \frac{n}{2}\left[\psi\left(\frac{\bar{\alpha}}{2}\right) - \log\left(\frac{\bar{\delta}}{2}\right)\right] - \frac{n}{2}\log(2\pi) \\
&\quad - \frac{\bar{\alpha}}{2\bar{\delta}}\left[\mathrm{tr}(\bar{\boldsymbol{B}}\boldsymbol{X}^\top\boldsymbol{X}) + (\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}})\right] \\
\mathrm{E}_{q^*(\beta)}\left[\log p(\boldsymbol{\beta})\right] &= -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\underline{\boldsymbol{B}}|) \\
&\quad - \frac{1}{2}\left[\mathrm{tr}(\bar{\boldsymbol{B}}\underline{\boldsymbol{B}}^{-1}) + (\bar{\boldsymbol{\beta}} - \underline{\boldsymbol{\beta}})^\top\underline{\boldsymbol{B}}^{-1}(\bar{\boldsymbol{\beta}} - \underline{\boldsymbol{\beta}})\right] \\
\mathrm{E}_{q^*(\sigma^2)}\left[\log p(\sigma^2)\right] &= \frac{\underline{\alpha}}{2}\log\left(\frac{\underline{\delta}}{2}\right) - \log\left[\Gamma\left(\frac{\underline{\alpha}}{2}\right)\right] \\
&\quad + \left(\frac{\underline{\alpha}}{2} + 1\right)\left[\psi\left(\frac{\bar{\alpha}}{2}\right) - \log\left(\frac{\bar{\delta}}{2}\right)\right] - \frac{\underline{\delta}\cdot\bar{\alpha}}{2\bar{\delta}} \\
\mathrm{E}_{q^*(\beta)}\left[\log q^*(\boldsymbol{\beta})\right] &= -\frac{p}{2}[1 + \log(2\pi)] - \frac{1}{2}\log(|\bar{\boldsymbol{B}}|) \\
\mathrm{E}_{q^*(\sigma^2)}\left[\log q^*(\sigma^2)\right] &= \left(\frac{\bar{\alpha}}{2} + 1\right)\psi\left(\frac{\bar{\alpha}}{2}\right) - \frac{\bar{\alpha}}{2} - \log\left[\Gamma\left(\frac{\bar{\alpha}}{2}\right)\right] - \log\left(\frac{\bar{\delta}}{2}\right)
\end{aligned}
\tag{21}
$$

Note that for computing these expectations we do require the proportionality constants, as they ensure the likelihood, prior and variational distributions are all in the same scale. We need this property as we hope to *compare across models* with different posteriors; i.e., different posterior hyperparameters.

While these equations might look complicated, they are all easy to calculate using a computer. This means that once we have obtained a set of values $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{B}}, \bar{\alpha}, \bar{\delta})$, we can easily compute $\mathrm{elbo}(q^*)$ associated to the current optimal variational approximation given by those parameter values. We then select the best fitting model (hyperparameter values) as the one that maximizes the ELBO.

# Implementation and Comparison to Gibbs Sampling

We now explore how to implement these equations in practice. Note the interdependent nature of (16) and (17), as computing one requires knowledge of the other. Additionally, note that while we started with completely general approximations, the families of the optimal variational approximations $q^*(\boldsymbol{\beta})$ (normal) and $q^*(\sigma^2)$ (inverse Gamma) were found analytically. All that is left to determine are their posterior hyperparameter values.

In practice, this means we choose initial values for either $\bar{\alpha}$ and $\bar{\delta}$ or $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{B}}$, and iteratively update each set of parameters values, alternating between using (9) and (13). This allows us to first compute say $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{B}}$, then compute $\bar{\alpha}$ and $\bar{\delta}$, and repeat until convergence to obtain $(\boldsymbol{\beta}^*, \boldsymbol{B}^*, \alpha^*, \delta^*)$. Finally, we use these pre-optimized values for the

[3]See `https://en.wikipedia.org/wiki/Multivariate_normal_distribution` and `https://en.wikipedia.org/wiki/Inverse-gamma_distribution` for sources and discussions of these properties. Here, $\psi(\cdot)$ represents the digamma function.

posterior hyperparameters to quickly sample many draws from the full variational posterior approximation:

$$q^*(\boldsymbol{\beta}, \sigma^2) = \mathcal{N}_p(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \boldsymbol{B}^*) \cdot \text{Inv-Gamma}\left(\sigma^2 \,\middle|\, \frac{\alpha^*}{2}, \frac{\delta^*}{2}\right) \tag{22}$$

A procedure that implements the coordinate ascent mean-field variational Bayes is given in Algorithm 1. It is instructive to compare this algorithm to Algorithm 2. This procedure implements a Gibbs sampler that can sample from the exact joint posterior (3) using MCMC methods. Indeed, using the same steps as those used for deriving (8) and (12), we could directly obtain the necessary conditional posteriors (**you should verify this!**):

$$\begin{gathered} \boldsymbol{\beta} \mid \sigma^2; \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{B}}) \quad \text{with} \\ \widetilde{\boldsymbol{B}} \coloneqq \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \underline{\boldsymbol{B}}^{-1}\right)^{-1} \quad \text{and} \quad \widetilde{\boldsymbol{\beta}} \coloneqq \widetilde{\boldsymbol{B}}\left(\frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} + \underline{\boldsymbol{B}}^{-1}\underline{\boldsymbol{\beta}}\right) \end{gathered} \tag{23}$$

$$\begin{gathered} \sigma^2 \mid \boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X} \sim \text{Inv-Gamma}\left(\sigma^2 \,\middle|\, \frac{\widetilde{\alpha}}{2}, \frac{\widetilde{\delta}}{2}\right) \quad \text{with} \\ \widetilde{\alpha} = \bar{\alpha} \quad \text{and} \quad \widetilde{\delta} \coloneqq \underline{\delta} + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{gathered} \tag{24}$$

Notice that Gibbs uses the realizations of $\boldsymbol{\beta}$ and $\sigma^2$ to update the posterior hyperparameters within each simulation loop. In contrast, variational Bayes replaces the realized value with the posterior mean of $\boldsymbol{\beta}$ and $\sigma^2$ (under the optimal variational approximation) when computing the posterior hyperparameters through (9) and (13). By moving the optimization of the posterior hyperparameters from the inner loop to the outer loop of the code, we achieve dramatic speed gains in fitting and sampling posteriors from Bayesian approaches.

**Algorithm 1** Coordinate Ascent Mean-Field Variational Bayes for Linear Regression with Independent Priors

**Require:** Outcome $\boldsymbol{y}$, covariates or features $\boldsymbol{X}$, prior hyperparameters $(\underline{\boldsymbol{\beta}}, \underline{\boldsymbol{B}}, \underline{\alpha}, \underline{\delta})$, initial values $\alpha_0$ and $\delta_0$, burn-in iterations $S_0$, keep iterations $S_1$, and tolerance $\epsilon$

$S \leftarrow S_0 + S_1$
Initialize $\bar{\alpha} \leftarrow \alpha_0$ and $\bar{\delta} \leftarrow \delta_0$ ▷ Initial value of parameters
Initialize $\text{elbo}_0 \leftarrow -\infty$ and $\text{elbo}_1 \leftarrow 0$
$\tau \leftarrow |\,\text{elbo}_1 - \text{elbo}_0\,|$ ▷ Convergence test quantity
**while** $\tau > \epsilon$ **do** ▷ Fitting loop goes first
    Update $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{B}}$ according to (9)
    Update $\bar{\alpha}$ and $\bar{\delta}$ according to (13)
    $\text{elbo}_1 \leftarrow \text{elbo}(q^*)$ using (20) and (21)
    $\tau \leftarrow |\,\text{elbo}_1 - \text{elbo}_0\,|$ ▷ Update test quantity
    $\text{elbo}_0 \leftarrow \text{elbo}_1$
**end while**
$\boldsymbol{\beta}^* \leftarrow \bar{\boldsymbol{\beta}}$, $\boldsymbol{B}^* \leftarrow \bar{\boldsymbol{B}}$, $\alpha^* \leftarrow \bar{\alpha}$, and $\delta^* \leftarrow \bar{\delta}$ ▷ Save pre-optimized values
**for** $s = 1, \ldots, S$ **do** ▷ Main simulation loop
    Draw $\boldsymbol{\beta}^{(s)}$ from $\mathcal{N}_p(\boldsymbol{\beta}^*, \boldsymbol{B}^*)$
    Draw $\sigma^{2(s)}$ from Inv-Gamma$(\alpha^*/2, \delta^*/2)$
**end for**
**return** $(\boldsymbol{\beta}^{(S_0+1)}, \ldots, \boldsymbol{\beta}^{(S)})$ and $(\sigma^{2(S_0+1)}, \ldots, \sigma^{2(S)})$

**Algorithm 2** Gibbs Sampler for Bayesian Linear Regression with Independent Priors

**Require:** Outcome $\boldsymbol{y}$, covariates or features $\boldsymbol{X}$, prior hyperparameters $(\underline{\boldsymbol{\beta}}, \underline{\boldsymbol{B}}, \underline{\alpha}, \underline{\delta})$, initial value $\sigma_0^2$, burn-in iterations $S_0$, and keep iterations $S_1$

$S \leftarrow S_0 + S_1$
Initialize $\sigma^{2(0)} \leftarrow \sigma_0^2$ ▷ Initial draw
**for** $s = 1, \ldots, S$ **do** ▷ Main simulation loop
    Calculate $\widetilde{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{\beta}}$ using (23) with $\sigma^2 = \sigma^{2(s-1)}$
    Draw $\boldsymbol{\beta}^{(s)}$ from $\mathcal{N}_p(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{B}})$ ▷ Update $\boldsymbol{\beta}$
    Calculate $\widetilde{\alpha}$ and $\widetilde{\delta}$ using (24) with $\boldsymbol{\beta} = \boldsymbol{\beta}^{(s)}$
    Draw $\sigma^{2(s)}$ from Inv-Gamma$(\widetilde{\alpha}/2, \widetilde{\delta}/2)$ ▷ Update $\sigma^2$
**end for**
**return** $(\boldsymbol{\beta}^{(S_0+1)}, \ldots, \boldsymbol{\beta}^{(S)})$ and $(\sigma^{2(S_0+1)}, \ldots, \sigma^{2(S)})$