MATH 80601A *Bayesian Modelling*

Practice final examination

Exam booklet

Examiner: Léo Belzile

Instructions: The time allotted for the examination is 180 minutes. You may answer in either English or French. No written material may be brought into the examination, but a simple (non-programmable) calculator may be used.

There are a total of 45 marks available in the exam paper, the distribution of which can be found in the right margin.

You must hand back the **exam booklet** at the end of the examination.

Question:	1	2	3	4	Total
Points:	10	9	10	16	45
Score:					

Crib sheet

1. binom(n,p) with n fixed and $p \in [0,1]$, has mass function $\binom{n}{p} p^y (1-p)^{n-y}$ for $y \in \{0,\ldots,n\}$. The expectation is np and the variance $np(1-p)$.
2. beta(a,b) random variable on $[0,1]$ has density $f(x) = \Gamma(a+b)/\{\Gamma(a)\Gamma(b)\} x^{a-1}(1-x)^{b-1}$ and expectation $a/(a+b)$.
3. Poisson(λ) with expectation $\lambda > 0$ has mass function $f(x) = \lambda^x / x! \exp(-\lambda), \quad x \in 0, 1, \ldots$.
4. Gauss(μ, σ^2) with expectation $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \qquad x \in \mathbb{R}.$$

5. Gauss$_p(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ with expectation $\boldsymbol{\mu} \in \mathbb{R}^p$ and precision (reciprocal variance) $\boldsymbol{Q} \succeq 0$, has density

$$f(x) = (2\pi)^{-p/2} |\boldsymbol{Q}|^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \qquad \boldsymbol{x} \in \mathbb{R}^p.$$

6. Student(μ, σ, ν) with location $\mu \in \mathbb{R}$, scale $\sigma > 0$ and ν degrees of freedom, has density

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1 + \frac{(x-\mu)^2}{\sigma^2\nu}\right)^{-\frac{\nu+1}{2}}, \qquad x \in \mathbb{R}.$$

7. gamma(α, β) with shape $\alpha > 0$ and rate $\beta > 0$, has expectation α/β, variance α/β^2 and density

$$f(x;\alpha,\beta) = \beta^\alpha / \Gamma(\alpha) x^{\alpha-1} \exp(-\beta x), \quad x > 0,$$

   where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) \mathrm{d}x$ is the gamma function, and $x\Gamma(x) = \Gamma(x+1)$ for $x > 0$. If $\alpha = 1$, we recover expo(β).
8. inv.gamma(α, β) with shape $\alpha > 0$ and rate $\beta > 0$, has expectation $\beta/(\alpha-1)$ for $\alpha > 1$ and density

$$f(x;\alpha,\beta) = \beta^\alpha / \Gamma(\alpha) x^{-\alpha-1} \exp(-\beta/x), \quad x > 0.$$

9. Laplace(μ, σ) with mean $\mu \in \mathbb{R}$ and scale $\sigma > 0$ with density $f(x) = (2\sigma)^{-1} \exp(-|x-\mu|/\sigma)$ for $x \in \mathbb{R}$. It's variance is $2\sigma^2$.

Jeffrey's prior: $p(\theta) \propto |\imath|^{1/2}$, where $\imath$ is the unit Fisher (expected) information.
Change of variable formula: consider an injective (one-to-one) differentiable function $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$, with inverse $\boldsymbol{g}^{-1}$. Then, if $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$, the density of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left\{\boldsymbol{g}^{-1}(\boldsymbol{y})\right\}\left|\mathbf{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{y})\right| = f_{\boldsymbol{X}}(\boldsymbol{x})\left|\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{x})\right|^{-1},$$

where $\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{x})$ is the Jacobian matrix with (i, j)th element $\partial[\boldsymbol{g}(\boldsymbol{x})]_i / \partial x_j$.
Metropolis–Hastings algorithm: consider $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{t-1})$, the proposal density evaluated at $\boldsymbol{\theta}$. The acceptance ratio for proposal $\boldsymbol{\theta}_t^\star$ given the current value $\boldsymbol{\theta}_{t-1}$ is $\min\{1, R\}$, where

$$R = \frac{p(\boldsymbol{\theta}_t^\star \mid \boldsymbol{y})}{p(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{y})} \frac{q(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}_t^\star)}{q(\boldsymbol{\theta}_t^\star \mid \boldsymbol{\theta}_{t-1})}$$

Effective sample size: for autocorrelation at lag t of ρ_t for a Markov chain of length B, $\mathsf{ESS} = B / \left\{1 + 2\sum_{t=1}^{\infty} \rho_t\right\}$.
Law of iterated mean and variance:

$$\mathsf{E}_{\boldsymbol{Y}}(\boldsymbol{Y}) = \mathsf{E}_{\boldsymbol{Z}}\left\{\mathsf{E}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\}, \qquad \mathsf{Va}_{\boldsymbol{Y}}(\boldsymbol{Y}) = \mathsf{E}_{\boldsymbol{Z}}\left\{\mathsf{Va}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\} + \mathsf{Va}_{\boldsymbol{Z}}\left\{\mathsf{E}_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y})\right\}.$$

Kullback–Leibler divergence:

$$\mathsf{KL}(f \parallel g) = \int \left\{\log f(\boldsymbol{x}) - \log g(\boldsymbol{x})\right\} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Quadratic form completion:

$$(\boldsymbol{x} - \boldsymbol{a})^\top \mathbf{A} (\boldsymbol{x} - \boldsymbol{a}) + (\boldsymbol{x} - \boldsymbol{b})^\top \mathbf{B} (\boldsymbol{x} - \boldsymbol{b}) \overset{\boldsymbol{x}}{\propto} (\boldsymbol{x} - \boldsymbol{c})^\top \mathbf{C} (\boldsymbol{x} - \boldsymbol{c}),$$

where $\mathbf{C} = \mathbf{A} + \mathbf{B}$ and $\boldsymbol{c} = \mathbf{C}^{-1}(\mathbf{A}\boldsymbol{a} + \mathbf{B}\boldsymbol{b})$.
Laplace approximation: Let $h(\boldsymbol{x})$ be a twice-differentiable function which is concave in the vicinity of it's mode $\boldsymbol{x}_0$, with $\mathbf{H}(\boldsymbol{x})$ the Hessian matrix of $-h(\boldsymbol{x})$. If $h(\boldsymbol{x}_0)$ is $\mathrm{O}(n)$, then as $n \to \infty$,

$$I_n = \int_{\mathbb{R}^d} \exp\{h(\boldsymbol{x})\} \mathrm{d}\boldsymbol{x} \approx (2\pi)^{p/2} |\mathbf{H}(\boldsymbol{x}_0)|^{-1/2} \exp\{h(\boldsymbol{x}_0)\}.$$

Evidence lower bound (ELBO): $\mathsf{ELBO}(g) = \mathsf{E}_g\{\log p(\boldsymbol{\theta}, \boldsymbol{y})\} - \mathsf{E}_g\{\log g(\boldsymbol{\theta}\}$.

Question 1. True or false ***10***

Explain whether each of the following statement is true, false or uncertain. To get marks, you must briefly justify your reasoning by proving the statement or providing a counterexample if the statement is false. **Answers without justifications are worth zero mark.**

1.1 A Gaussian approximation g that minimizes the forward Kullback–Leibler divergence and a Laplace approximation to a univariate density p will have the same mean. [2]

> **Solution:** False; the Laplace approximation will be centered at the mode, while the other model will be centered at the mean of the true distribution. They coincide for symmetric unimodal distributions.

1.2 We can readily obtain marginal posterior moments from variational inference procedures. [2]

> **Solution:** Uncertain; the optimal decomposition might lead to a multivariate family, which may be unnormalized. Gaussian approximations will of course be marginalizable, and we could resort to Monte Carlo methods if we can simulate.

1.3 The impact of the prior vanishes as the sample size gets larger. [2]

> **Solution:** False; this depends largely on what level the parameter or prior is located; for example, a random effect model with different variance per group would vanish only if the number of groups increases (in addition to the number of observations within each group).
>
> Other counterexample: any prior that restricts the support would not vanish.

1.4 Marginalization in Markov chain Monte Carlo methods is always beneficial (fewer parameters, lower dependence between components). [2]

> **Solution:** False; we have seen examples where parameter expansion improves mixing for the eight school example by adding redundancy. Generally, it complexifies the dependence structure, while reducing the dimension (thinks data augmentation in Gibbs sampling).

1.5 Consider a prediction $\widetilde{y}$ from a frequentist model with density $f(\widetilde{y};\boldsymbol{\theta})$, where $\widehat{\boldsymbol{\theta}}$ is a point estimator. The posterior predictive distribution $p(\widetilde{y} \mid \boldsymbol{y})$ will be more variable than it's frequentist counterpart $f(\widetilde{y};\widehat{\boldsymbol{\theta}})$. [2]

> **Solution:** True, follows from the law of iterated variance as
>
> $$p(\widetilde{y} \mid \boldsymbol{y}) = \int f(\widetilde{y};\boldsymbol{\theta},\boldsymbol{y})\, p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$

so $\mathsf{Va}_{\widetilde{y}|\boldsymbol{y}}(\widetilde{Y}) = \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\{\mathsf{Va}_{\widetilde{y}|\boldsymbol{\theta}}(\widetilde{Y})\} + \mathsf{Va}_{\boldsymbol{\theta}|\boldsymbol{y}}\{\mathsf{E}_{\widetilde{y}|\boldsymbol{\theta}}(\widetilde{Y})\}$. The frequentist version amounts to taking a point mass at $p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \mathrm{I}(\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}})$, so the second term vanishes as there is no variability.

Question 2. Short questions ***9***

2.1 Zellner (1996) proposed the maximal data information (MDI) prior. The latter is proportional to the exponential of the negative entropy, $p(\boldsymbol{\theta}) \propto \exp\{\mathsf{E}_g(\log g)\}$. [3]

- Show that the entropy of a Gaussian random variable $\mathsf{Gauss}(\mu, \sigma^2)$ is a function of the scale σ only.
- Use this result to derive the corresponding MDI prior for (μ, σ).
- Is the resulting prior proper? Justify your answer.

Solution:

$$-\mathsf{E}_g(\log g) = \frac{1}{2}\log(2\pi) + \log(\sigma) + \frac{1}{2\sigma^2}\mathsf{E}_X\{(X-\mu)^2\} = \frac{1+\log(2\pi)}{2} + \log(\sigma)$$

so the MDI prior is proportional to $1/\sigma$, hence is improper.

2.2 Define the concepts of **burn in**, **warmup**, and **thinning** for Markov chain Monte Carlo. Explain their relevance in the context of a sampling-based Bayesian analysis. [3]

Solution: Burn in refers to discarding the initial transient runs of the Markov chain, to ensure that it has converged to the stationnary distribution. Warmup is used sometimes interchangedly, but also serves for tuning parameters of the algorithm. Both of these initial runs are discarded after sampling is performed, to avoid biasing the results.

Thining refers to the practice of keeping only a fraction of the draws. This is only useful when there is strong autocorrelation to reduce storage costs.

2.3 The following **R** code implements a simple Markov chain Monte Carlo to sample from the posterior mean and std. deviation (μ, σ), where we assume independent and identically observations and a half-Cauchy prior on $[0, \infty)$ for the scale, assumed independent apriori [3]

$$Y_i \mid \mu, \sigma \sim \mathsf{Gauss}(\mu, \sigma^2), i = 1, \ldots, n; \qquad p(\mu) \propto 1; \qquad p(\sigma) \propto \mathsf{Student}_+(1, 0, 1).$$

Find the error in the code and explain why this isn't sampling from the posterior distribution of interest.

```
sd_prop <- 0.05
# Unnormalized log posterior for mu, sigma
logpost <- function(pars, y){
  mu <- pars[1]; sigma <- pars[2]
  if(sigma < 0){ return(-Inf)}
  sum(dnorm(x = y, mean = mu, sd = sigma, log = TRUE)) +  # log likelihood
    log(2) + dt(x = sigma, df = 1, log = TRUE) # log prior for scale
}

B <- 1e4L # number of simulations
mu_s <- sigma_s <- numeric(B) # containers
mu <- mean(y); sigma <- sd(y) # initial values
for(b in 1:B){
  # Gibbs step for mu
  mu <- mu_s[b] <- rnorm(n = 1, mean = mean(y), sd = sigma/sqrt(length(y)))
  # Metropolis random walk on the log scale
  sigma_prop <- exp(rnorm(n = 1, mean = log(sigma), sd = sd_prop))
  # Log of Metropolis acceptance ratio
  logR <- logpost(c(mu, sigma_prop), y = y) - logpost(c(mu, sigma), y = y)
  if(logR > log(runif(1))){
    sigma <- sigma_prop
  }
  sigma_s[b] <- sigma
}
```

Solution: The Jacobian of the transformation is missing; the ratio of proposal densities isn't zero since the lognormal is not symmetric.

Question 3. Stochastic versus deterministic approximations ***10***

3.1 Define the **marginal likelihood** as a function of the prior $p(\boldsymbol{\theta})$ and the likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$. [2]

Solution: The marginal likelihood is the normalizing constant for the posterior, defined as

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

3.2 Explain in your own words why estimation of the marginal likelihood is difficult. [2]

Solution: The marginal likelihood is often small (think discrete components with probabilty) and decreases with n. We have seen examples where numerical integral will fail because of numerical overflow. Likewise, Monte Carlo does not readily work because sampling from the prior unless the latter aligns with the likelihood model.

3.3 Explain how to obtain a Laplace approximation of the marginal likelihood. Under which circumstances will it be a good approximation? [2]

Solution: Write $\widehat{\boldsymbol{\theta}}$ for the maximum a posteriori (MAP) and $\mathbf{H}(\widehat{\boldsymbol{\theta}})$ for the Hessian of the negative log posterior evaluate at the MAP. Then a straightforward application of Laplace approximation gives

$$p(\boldsymbol{y}) \approx (2\pi)^{p/2} |\mathbf{H}(\widehat{\boldsymbol{\theta}})|^{-1/2} p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}) p(\widehat{\boldsymbol{\theta}})$$

It is a good approximation if the sample size n is large and the posterior is approximately symmetric in the parameterization considered.

3.4 Explain the relevance of the marginal likelihood in the context of [2]

1. calculation of posterior moments of the form $\mathsf{E}_{\boldsymbol{\Theta}|\boldsymbol{Y}}\{g(\boldsymbol{\theta})\}$.
2. model comparison using Bayes factor.

Solution: Given the joint $p(\boldsymbol{y}, \boldsymbol{\theta})$, any expectation is of the form

$$\mathsf{E}_{\boldsymbol{\Theta}|\boldsymbol{Y}}\{g(\boldsymbol{\theta})\} = p(\boldsymbol{y})^{-1} \int g(\boldsymbol{\theta}) p(\boldsymbol{y}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

we need the normalizing constant to compute the moments.

Model comparison in Bayesian is based on the posterior model odds, which are

$$p(\mathscr{M}_i \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \mathscr{M}_i)p(\mathscr{M}_i).$$

For two models $\mathscr{M}_1$ and $\mathscr{M}_2$, the Bayes factor involve the ratio of marginal likelihoods $p(\boldsymbol{y} \mid \mathscr{M}_1)/p(\boldsymbol{y} \mid \mathscr{M}_2)$. These require necessarily proper priors.

3.5 How does Markov chain Monte Carlo (MCMC) methods get around estimation of the marginal likelihood $p(\boldsymbol{y})$? *Hint:* consider the Metropolis–Hastings acceptance ratio. [2]

Solution: We compare current value with the proposal; since the marginal likelihood is a constant, taking ratios of posterior leads to it's cancellation in the Metropolis–Hastings ratio.

Once we have posterior samples, we can evaluate any functional of interest via Monte Carlo, circumventing the evaluation of the normalizing constant. We however cannot use this for model comparison.

Question 4. Probit regression ***16***

Experiment 2 of Duke and Amir (2023) consider the effect of sequential versus integrated decisions on customers decision to buy. Customers in an online experiment where exposed to products and decided whether to buy ($Y_i = 1$) or not ($Y_i = 0$). To model these, we consider a simple probit regression model with response $Y_i \sim \mathsf{binom}(1, p_i)$, where

$$p_i = \Pr(Y_i = 1) = \Phi(\mathbf{x}_i\boldsymbol{\beta}),$$

with $\Phi(\cdot)$ the distribution function of the standard Gaussian distribution. We set $\boldsymbol{\beta} \sim \mathsf{Gauss}_p(\mathbf{0}_p, c\mathbf{I}_p)$ for $c > 0$ a known positive constant.

The model matrix include an intercept, a coefficient for age and a binary indicator equal to 1 if the participant was exposed to quantity-integrated decision, and zero for quantity-sequential (control group).

4.1 Consider the data augmentation scheme where $Y_i = \mathrm{I}(Z_i > 0)$, where $Z_i \sim \mathsf{Gauss}(\mathbf{x}_i\boldsymbol{\beta}, 1)$, with $\mathbf{x}_i$ denoting the ith row of the $n \times p$ design matrix. [2]

Write down the expression for the joint distribution $p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}) = p(\boldsymbol{y} \mid \boldsymbol{z})p(\boldsymbol{z} \mid \boldsymbol{\beta})p(\boldsymbol{\beta})$.

Solution:

$$p(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2c}\boldsymbol{\beta}^\top\boldsymbol{\beta} - \frac{1}{2}(\boldsymbol{z}-\mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{z}-\mathbf{X}\boldsymbol{\beta})\right\} \times \prod_{i=1}^{n} \mathsf{I}(z_i > 0)^{y_i}\mathsf{I}(z_i \leq 0)^{1-y_i}$$

4.2 Derive the conditional distributions $p(\boldsymbol{\beta} \mid \boldsymbol{z})$ and that of $p(z_i \mid y_i, \boldsymbol{\beta})$ for $i = 1, \ldots, n$. [4]

Solution: A simple Gaussian completion of squares gives

$$\boldsymbol{\beta} \mid \boldsymbol{z}, \boldsymbol{y} \sim \mathsf{Gauss}_p\left\{(\mathbf{X}^\top\mathbf{X} + c^{-1}\mathbf{I}_p)^{-1}\mathbf{X}^\top\boldsymbol{z}, (\mathbf{X}^\top\mathbf{X} + c\mathbf{I}_p)^{-1}\right\}.$$

The augmented variables Z_i are conditionally independent and truncated Gaussian with unit variance

$$Z_i \mid y_i, \boldsymbol{\beta} \sim \begin{cases} \mathsf{trunc.Gauss}(\mathbf{x}_i\boldsymbol{\beta}, 1, -\infty, 0) & y_i = 0 \\ \mathsf{trunc.Gauss}(\mathbf{x}_i\boldsymbol{\beta}, 1, 0, \infty) & y_i = 1. \end{cases}$$

4.3 Based on the conditional distributions detail a Gibbs sampling algorithm for $\boldsymbol{\beta}$ and $\boldsymbol{z}$. Explain the benefit of the latter over the marginal posterior $p(\boldsymbol{\beta} \mid \boldsymbol{y})$. [2]

Solution: We obtain conditional conjugacy, which simplifies the likelihood. The algorithm is also automatic since the acceptance rate of Gibbs sampling is one; there are no tuning parameters.

4.4 Suppose that we instead used coordinate-ascent variational inference with a factorization of the posterior $p_{\boldsymbol{Z}}(\boldsymbol{z})p_{\boldsymbol{\beta}}(\boldsymbol{\beta})$. [4]

Write down the optimal form of these distributions and the parameter updates. Explain how you would assess convergence.

Hint: if $Y \sim \mathsf{trunc.Gauss}(\mu, \sigma, a, b)$ a truncated Gaussian on $[a, b]$ with location μ and scale σ has expectation

$$\mathsf{E}(Y) = \mu - \sigma\frac{\phi\{(b-\mu)/\sigma\} - \phi\{(a-\mu)/\sigma\}}{\Phi\{(b-\mu)/\sigma\} - \Phi\{(a-\mu)/\sigma\}},$$

where ϕ and Φ are the density and distribution functions of a standard Gaussian, respectively.

Solution: If we consider a factorization of the form $g_{\boldsymbol{Z}}(\boldsymbol{z})g_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, then we exploit the conditionals in the same way as for Gibbs sampling, but substituting unknown parameter functionals by their expectations. Furthermore, the optimal form of the density further factorizes as $g_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{i=1}^n g_{Z_i}(z_i)$.

The model depends on the mean parameter of $\boldsymbol{Z}$, say $\mu_{\boldsymbol{Z}}$, and that of $\boldsymbol{\beta}$, say $\mu_{\boldsymbol{\beta}}$. To see this, consider the terms in the posterior proportional to Z_i, where

$$p(z_i \mid \boldsymbol{\beta}, y_i) \propto -\frac{z_i^2 - 2z_i\mathbf{x}_i\boldsymbol{\beta}}{2} \times \mathrm{I}(z_i > 0)^{y_i}\mathrm{I}(z_i < 0)^{1-y_i}$$

which is linear in $\boldsymbol{\beta}$. The expectation of a univariate truncated Gaussian $Z \sim \mathsf{trunc.Gauss}(\mu, \sigma^2, l, u)$ is

$$\mathsf{E}(Z) = \mu - \sigma\frac{\phi\{(u-\mu/\sigma)\} - \phi\{(l-\mu/\sigma)\}}{\Phi\{(u-\mu/\sigma)\} - \Phi\{(l-\mu/\sigma)\}}.$$

If we replace $\mu = \mathbf{x}_i\mu_{\boldsymbol{\beta}}$ in this expression, we get the update

$$\mu_{Z_i}(z_i) = \begin{cases} \mathbf{x}_i\mu_{\boldsymbol{\beta}} - \frac{\phi(\mathbf{x}_i\mu_{\boldsymbol{\beta}})}{1-\Phi(\mathbf{x}_i\mu_{\boldsymbol{\beta}})} & y_i = 0; \\ \mathbf{x}_i\mu_{\boldsymbol{\beta}} + \frac{\phi(\mathbf{x}_i\mu_{\boldsymbol{\beta}})}{\Phi(\mathbf{x}_i\mu_{\boldsymbol{\beta}})} & y_i = 1, \end{cases}$$

since $\phi(x) = \phi(-x)$.

The optimal form for $\boldsymbol{\beta}$ is Gaussian and the only unknown parameter is $\boldsymbol{\mu}_{\boldsymbol{\beta}}$; the log density only involves a linear form for $\boldsymbol{Z}$, so follows from the same principle. We get regrouping the vector of means for the latent variables

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X} + \mathbf{Q_0})^{-1}(\mathbf{X}\boldsymbol{\mu}_{\boldsymbol{Z}} + \mathbf{Q}_0\boldsymbol{\mu}_0)$$

Starting with an initial vector for either parameters, the CAVI algorithm alternatives between updates of $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\boldsymbol{\mu}_{\boldsymbol{Z}}$ until the value of the evidence lower bound stabilizes.

4.5 The right panel of Figure 1 shows the marginal density for the coefficient β_2. Explain why the two are not identical. [2]

Solution: CAVI solves an approximate problem by maximizing the ELBO (or equivalently minimizing the reverse KL divergence), so it needs not coincide. We see that the Gaussian approximation does not fully capture the scale of the marginal.

4.6 What can we conclude from Figure 1 as to what is the most effective method? [2]

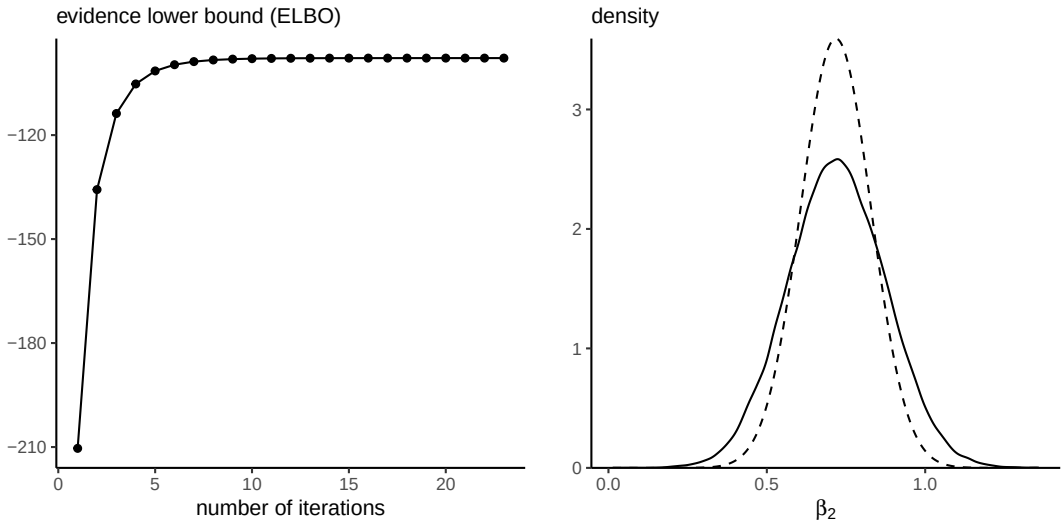

Figure 1: Left: evidence lower bound (ELBO) as a function of iteration. Right: marginal density of β_2 for quantity-integrated binary indicator, based on Monte Carlo samples (full), versus variational approximation (dashed).

Solution: The posterior for the effect of quantity-integrated (versus control for the intercept has marginal posterior probability $\Pr(\beta_2 > 0) \approx 1$. The latter thus indicates that, for the same age, it leads to more sales. This finding is contingent on the model being correct.