**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$(\lambda)$ with expectation $\lambda > 0$ has mass function $f(x) = \lambda^x / x! \exp(-\lambda), \quad x \in 0, 1, \ldots$.
4. Gauss$(\mu, \sigma^2)$ 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$(\mu, \sigma, \nu)$ with location $\mu \in \mathbb{R}$, scale $\sigma > 0$ and $\nu$ 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$(\alpha, \beta)$ with shape $\alpha > 0$ and rate $\beta > 0$, has expectation $\alpha/\beta$, variance $\alpha/\beta^2$ 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$(\beta)$.
8. inv.gamma$(\alpha, \beta)$ 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$(\mu, \sigma)$ 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 $\rho_t$ 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]

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

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

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

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]

**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 $\sigma$ only.
- Use this result to derive the corresponding MDI prior for $(\mu,\sigma)$.
- Is the resulting prior proper? Justify your answer.

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]

2.3 The following **R** code implements a simple Markov chain Monte Carlo to sample from the posterior mean and std. deviation $(\mu, \sigma)$, 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
}
```

**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]

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

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

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

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

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]

**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 $i$th row of the $n \times p$ design matrix. [2]

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

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]

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]

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.

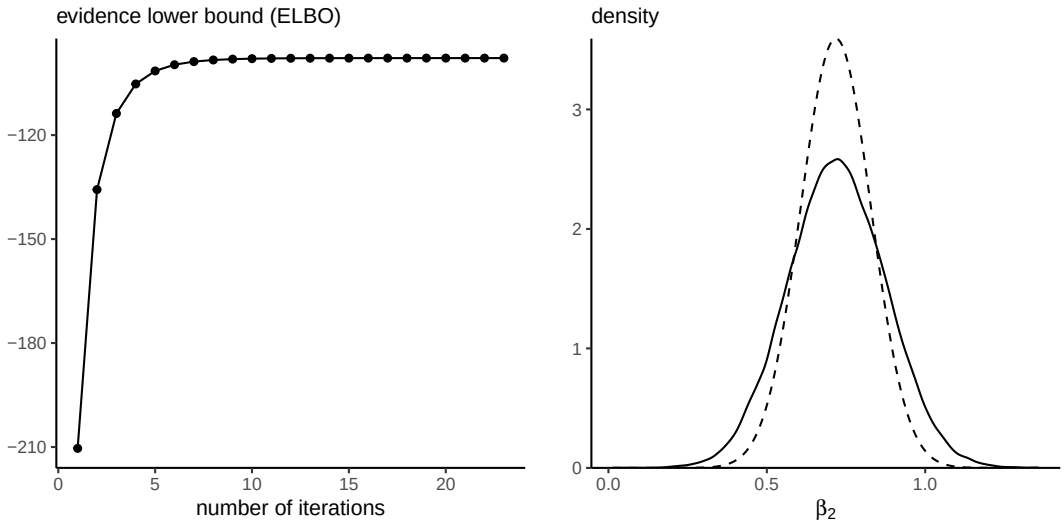


**Figure 1:** Left: evidence lower bound (ELBO) as a function of iteration. Right: marginal density of $\beta_2$ for quantity-integrated binary indicator, based on Monte Carlo samples (full), versus variational approximation (dashed).

*Hint:* if $Y \sim \mathsf{trunc.Gauss}(\mu, \sigma, a, b)$ a truncated Gaussian on $[a, b]$ with location $\mu$ and scale $\sigma$ 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 $\phi$ and $\Phi$ are the density and distribution functions of a standard Gaussian, respectively.

4.5 The right panel of Figure 1 shows the marginal density for the coefficient $\beta_2$. Explain why the two are not identical. [2]

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