MAST30027: Modern Applied Statistics

Assignment 4, 2023.

Due: 5pm Sunday Oct 22nd

- This assignment is worth 17% of your total mark.
- To get full marks, show your working including 1) R commands and outputs you use, 2) mathematics derivation, and 3) rigorous explanation why you reach conclusions or answers. If you just provide final answers, you will get zero mark.
- The assignment you hand in must be typed (except for math formulas), and be submitted using LMS as a single PDF document only (no other formats allowed). For math formulas, you can take a picture of them. Your answers must be clearly numbered and in the same order as the assignment questions.
- The LMS will not accept late submissions. It is your responsibility to ensure that your assignments are submitted correctly and on time, and problems with online submissions are not a valid excuse for submitting a late or incorrect version of an assignment.
- We will mark a selected set of problems. We will select problems worth $\geq 50\%$ of the full marks listed.
- If you need an extension, please contact the lecturer before the due date with appropriate justification and supporting documents. Late assignments will only be accepted if you have obtained an extension from the lecturer before the due date. To ensure that the lecturer responds to your extension request email before the due date, please contact 24h before the due date. Under no circumstances an assignment will be marked if solutions for it have been released.
- Also, please read the "Assessments" section in "Subject Overview" page of the LMS.

---

1. The file `assignment4_prob1_2023.txt` contains final exam scores of 100 students in Modern Applied Statistics. We can read the scores as follows.

```
> X = scan(file="assignment4_prob1_2023.txt", what=double())
Read 100 items
> length(X)
[1] 100
> mean(X)
[1] 75.726
```

Suppose that the 100 scores are independent to each other and they follow Normal distribution with mean = 75 and unknown precision τ. Specifically, let $x_1, \ldots, x_{100}$ be the final exam scores, and

$$x_i \sim \mathrm{N}(75, \frac{1}{\tau}) \quad \text{for} \quad i = 1, \ldots, 100.$$

Suppose that the precision τ has a Gamma(2, 1) prior distribution, where Gamma(α, β) has the pdf

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\left(-\beta x\right).$$

(a) **(5 marks)** Derive the posterior distribution of the precision τ conditioned on the final exam scores of 100 students, $p(\tau|x_1, \ldots, x_{100})$. Evaluate parameters in the posterior distribution using the data from `assignment4_prob1_2023.txt`.

(b) **(7 marks)** Derive the posterior predictive distribution for a new score $\tilde{x}$, $p(\tilde{x}|x_1, \ldots, x_{100})$. Evaluate parameters in the posterior predictive distribution using the data from `assignment4_prob1_2023.txt`.

[Hint for (b)] A three-parameter version of a t distribution (Jackman, S. (2009)), denoted by $t(\nu, a, b)$, has the pdf

$$p(x|\nu, a, b) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu b}} \left(1 + \frac{1}{\nu}\frac{(x-a)^2}{b}\right)^{-\frac{\nu+1}{2}}.$$

2. **Data:** The files `assignment4_prob2_x_2023.txt` and `assignment4_prob2_y_2023.txt` contain 100 observations $(x_1, \ldots, x_{100})$ and 150 observations $(y_1, \ldots, y_{150})$, respectively. You can read the data as follows.

```
> x = scan(file="assignment4_prob2_x_2023.txt", what=double())
> y = scan(file="assignment4_prob2_y_2023.txt", what=double())
> length(x)
[1] 100
> length(y)
[1] 150
> mean(x)
[1] 3.196441
> mean(y)
[1] -1.979781
```

Model: We assume that $x_1, \ldots, x_{100}$ and $y_1, \ldots, y_{150}$ are independent and follow normal distributions:

$$x_i \sim \mathrm{N}(\mu_1, 1^2) \text{ for } i = 1, \ldots, 100,$$
$$y_j \sim \mathrm{N}(\mu_2, \left(\frac{1}{\sqrt{2}}\right)^2) \text{ for } j = 1, \ldots, 150.$$

Prior: We impose the following bivariate normal prior for the mean parameters:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \frac{3}{5} & -\frac{2}{5} \\ -\frac{2}{5} & \frac{3}{5} \end{pmatrix}.$$

[Recall that the joint density of $\boldsymbol{x}$ which follows a bivariate normal distribution, i.e., $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{pmatrix} \mu_1^0 \\ \mu_2^0 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ is

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).]$$

Posterior inference using Gibbs sampling

(a) **(10 marks)** Derive the following conditional distributions.

$$p(\mu_1|\mu_2, x_1, \ldots, x_{100}, y_1, \ldots, y_{150}) \quad \text{and} \quad p(\mu_2|\mu_1, x_1, \ldots, x_{100}, y_1, \ldots, y_{150}).$$

If they are known distributions, write distribution names and their parameters. [For example, gamma distribution with shape $= \sum_i^{100} x_i$ and scale $= \sum_i^{100} x_i^2$].

(b) **(5 marks)** Write a code that uses the Gibbs sampling to simulate samples from $p(\mu_1, \mu_2 | x_1, \ldots, x_{100}, y_1, \ldots, y_{150})$. Run two Gibbs sampling chains with the following two initial values.

	μ_1	μ_2
1st initial values	0	0
2nd initial values	2	-1

Please run with at least 500 iterations. Make a trace plot for each of parameters and see if samples from different chains are mixed well and behave similarly.

(c) **(7 marks)** Using the simulated sample from one chain, for each parameter 1) make a plot that shows empirical (estimated) marginal posterior distribution, 2) estimate marginal posterior mean, and 3) report a 90% credible interval for the marginal posterior distribution. You can find a 90% credible interval in a number of ways. For this assignment, use 5% in each tail. Discard early iterations as a burn-in. You can decide burn-in period from the trace plots in (b).

Posterior inference using Metropolis-Hastings (MH) algorithm

(d) **(15 marks)** Write a code that uses the MH algorithm to simulate samples from $p(\mu_1, \mu_2 | x_1, \ldots, x_{100}, y_1, \ldots, y_{150})$. For the current values of parameters (μ_1^c, μ_2^c), we propose new values (μ_1^n, μ_2^n) as follows. $\mu_1^n \sim$ Normal(mean $= \mu_1^c$, variance $= 0.1^2$) and $\mu_2^n \sim$ Normal(mean $= \mu_2^c$, variance $= 0.1^2$). Run two MH chains with the following two initial values.

	μ_1	μ_2
1st initial values	0	0
2nd initial values	2	-1

Please run with at least 2000 iterations. Make a trace plot for each of parameters and see if samples from different chains are mixed well and behave similarly.

(e) **(7 marks)** Repeat (c) in the Gibbs sampling.

Posterior inference using Variational Inference (VI)

We will apply VI with the mean-field variational family where $q(\mu_1, \mu_2) = q_{\mu_1}(\mu_1) q_{\mu_2}(\mu_2)$ and use the CAVI algorithm for optimisation. The CAVI iteratively optimises each factor as follows while holding the other factors fixed:

$$\begin{aligned} q^*_{\mu_1}(\mu_1) &\propto \exp\{\mathbb{E}_{\mu_2}[\log p(\mu_1, \mu_2, x_1, \ldots, x_{100}, y_1, \ldots, y_{150})]\}, \\ q^*_{\mu_2}(\mu_2) &\propto \exp\{\mathbb{E}_{\mu_1}[\log p(\mu_1, \mu_2, x_1, \ldots, x_{100}, y_1, \ldots, y_{150})]\}, \end{aligned}$$

where the expectations $\mathbb{E}_{\mu_1}$ and $\mathbb{E}_{\mu_2}$ are taken with respect to $q^*_{\mu_1}(\mu_1)$ and $q^*_{\mu_2}(\mu_2)$, respectively.

(f) **(10 marks)** Derive $q^*_{\mu_1}(\mu_1)$ and $q^*_{\mu_2}(\mu_2)$ and write the corresponding distribution names and their parameters. [For example, for the model and prior we considered in the lecture - see the page 20 of the Variational Inference slides, we derived and presented that $q^*_\mu(\mu)$ is the pdf of $N(\mu^*, \sigma^{2*})$ and $q^*_\tau(\tau)$ is the pdf of Gamma(a^*, b^*), where $\mu^* = \frac{\lambda_0 \mu_0 + n\bar{x}}{\lambda_0 + n}$, $\sigma^{2*} = \frac{1}{(\lambda_0 + n)\mathbb{E}_\tau[\tau]}$, $\mathbb{E}_\tau[\tau] = \frac{a^*}{b^*}$, $a^* = \frac{n+1}{2} + a_0$, $b^* = b_0 + \frac{\sum_i \mathbb{E}_\mu[(x_i - \mu)^2]}{2} + \frac{\lambda_0 \mathbb{E}_\mu[(\mu - \mu_0)^2]}{2}$, $\mathbb{E}_\mu[(x_i - \mu)^2] = \sigma^{2*} + (\mu^*)^2 - 2x_i\mu^* + x_i^2$, and $\mathbb{E}_\mu[(\mu - \mu_0)^2] = \sigma^{2*} + (\mu^*)^2 - 2\mu_0\mu^* + \mu_0^2$.]

(g) **(10 marks)** Derive the ELBO up to constant.

(h) **(10 marks)** Implement the CAVI algorithm and obtain $q^*_{\mu_1}(\mu_1)$ and $q^*_{\mu_2}(\mu_2)$ which minimise the KL divergence by applying the implemented algorithm to $x_1, \ldots, x_{100}, y_1, \ldots, y_{150}$. Set the CAVI algorithms to stop when either the number of iterations reaches 100 (max.iter = 100) or the ELBO has changed by less than 0.00001 ($\epsilon = 0.00001$). Run the CAVI algorithm at least two times using different initial values and report $q^*_{\mu_1}(\mu_1)$ and $q^*_{\mu_2}(\mu_2)$ with the highest ELBO. Provide the initial values which lead the reported $q^*_{\mu_1}(\mu_1)$ and $q^*_{\mu_2}(\mu_2)$ and we will probably run our own implementation using the initial values and see if we can reproduce your answer when marking. For each run, check that the ELBO increase at each iteration by plotting them.