THE UNIVERSITY OF MELBOURNE

Student number

Semester 2 Assessment, 2021

School of Mathematics and Statistics

MAST30027 Modern Applied Statistics

Reading time: 30 minutes — Writing time: 2 hours — Upload time: 30 minutes

This exam consists of 12 pages (including this page) with 5 questions and 50 total marks

Permitted Materials

- This exam and/or an offline electronic PDF reader, one or more copies of the masked exam template made available earlier, blank loose-leaf paper and a Casio FX-82 calculator.
- One double sided A4 page of notes (handwritten only).
- No headphones or earphones are permitted.

Instructions to Students

- Wave your hand right in front of your webcam if you wish to communicate with the supervisor at any time (before, during or after the exam).
- You must not be out of webcam view at any time without supervisor permission.
- You must not write your answers on an iPad or other electronic device.
- Off-line PDF readers (i) must have the screen visible in Zoom; (ii) must only be used to read exam questions (do not access other software or files); (iii) must be set in flight mode or have both internet and Bluetooth disabled as soon as the exam paper is downloaded.

Writing

- If you are writing answers on the exam or masked exam and need more space, use blank paper. Note this in the answer box, so the marker knows.
- If you are only writing on blank A4 paper, the first page must contain only your student number, subject code and subject name. Write on one side of each sheet only. Start each question on a new page and include the question number at the top of each page.

Scanning and Submitting

- **You must not leave Zoom supervision to scan your exam.** Put the pages in number order and the correct way up. Add any extra pages to the end. Use a scanning app to scan all pages to PDF. Scan directly from above. Crop pages to A4.
- Submit your scanned exam as a single PDF file and carefully review the submission in Gradescope. Scan again and resubmit if necessary. Do not leave Zoom supervision until you have confirmed orally with the supervisor that you have received the Gradescope confirmation email.
- **You must not submit or resubmit after having left Zoom supervision.**

Page 1 of 12 — add any extra pages after page 12 — Page 1 of 12

Page 2 of 12 — add any extra pages after page 12 — Page 2 of 12

Question 1 (9 marks)

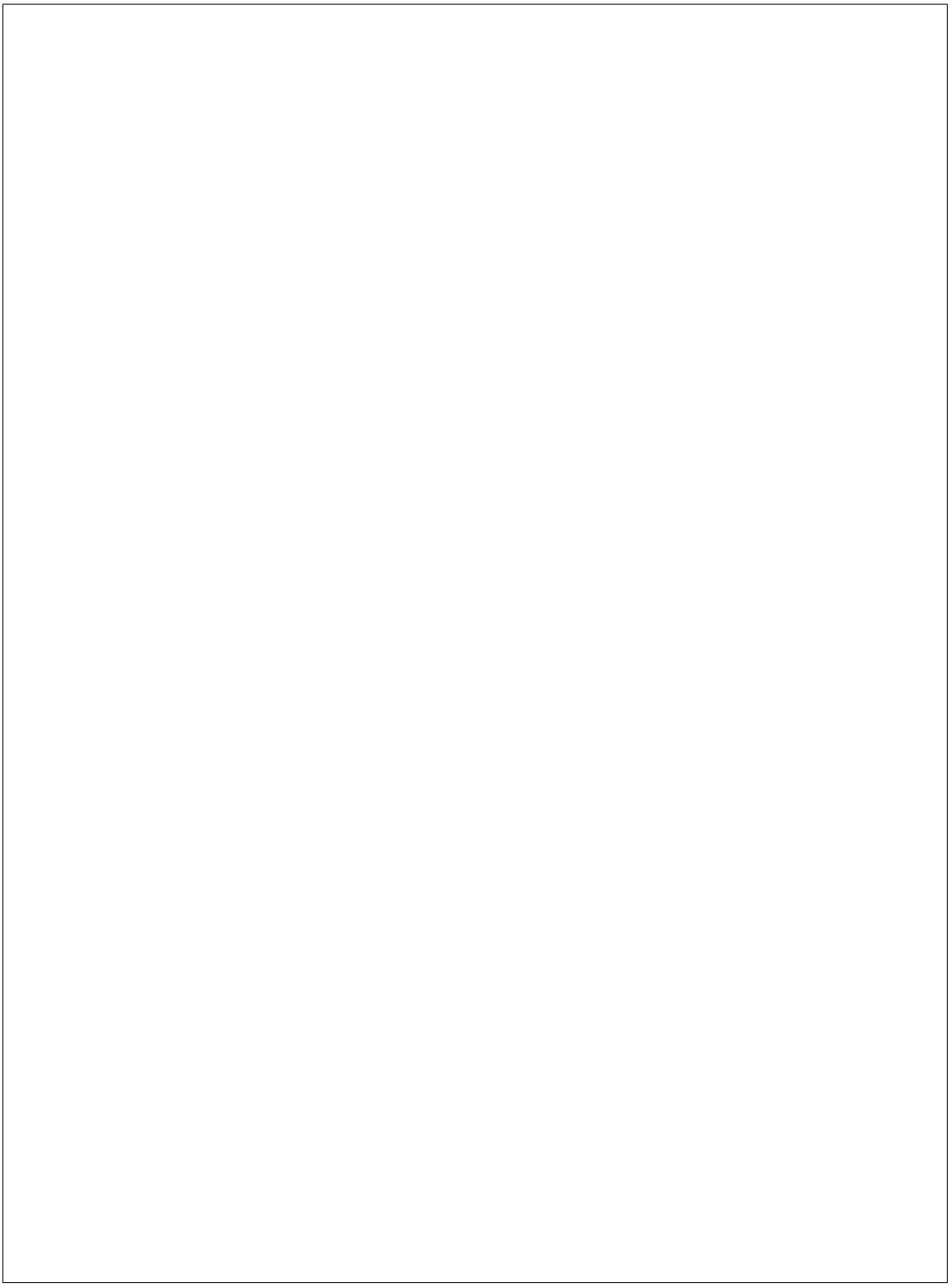
The exponential distribution with $\lambda > 0$ has the probability density function

$$f(x; \lambda) = \lambda e^{-\lambda x}$$

for $x \geq 0$.

(a) Show that the exponential distribution is an exponential family and obtain its variance function.

(b) Let $X_1, \cdots, X_n$ be independent samples from the exponential distribution. Find the MLE of λ and its asymptotic distribution.



Page 3 of 12 — add any extra pages after page 12 — Page 3 of 12

Page 4 of 12 — add any extra pages after page 12 — Page 4 of 12

Question 2 (4 marks)

The `dvisits` data in the `faraway` package comes from the Australian Health Survey of 1977–78 and consists of 5190 observations on single adults, where young and old have been oversampled. We consider `doctorco` as a response and `sex, age, income, levyplus, freepoor, freerepa, illness, actdays` as predictor variables. The description of each variable is as follows.

- `doctorco`: number of consultations with a doctor or specialist in the past 2 weeks
- `sex`: 1 if female, 0 if male
- `age`: age in years divided by 100
- `income`: annual income in Australian dollars divided by 1000
- `levyplus`: 1 if covered by a private health insurance fund for private patients in a public hospital (with doctor of choice), 0 otherwise
- `freepoor`: 1 if covered by government because of low income, recent immigrant, or unemployed, 0 otherwise
- `freerepa`: 1 if covered by government because of old-age or disability pension, or because of invalid veteran or family of deceased veteran, 0 otherwise
- `illness`: number of illnesses in past 2 weeks, with 5 or more coded as 5
- `actdays`: number of days of reduced activity in past two weeks due to illness or injury

We analysed the data using a Poisson regression and obtained the following model after a model selection procedure.

```
> modelA <- glm(doctorco ~ sex + age + income + freepoor,
+               family=poisson, data=dvisits)
> summary(modelA)

Call:
glm(formula = doctorco ~ sex + age + income + freepoor, family = poisson,
    data = dvisits)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0408  -0.7989  -0.6778  -0.5897   6.3412

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.62728    0.09514 -17.104  < 2e-16 ***
sex          0.20525    0.05597   3.667 0.000245 ***
age          1.15055    0.13277   8.666  < 2e-16 ***
income      -0.32698    0.08162  -4.006 6.17e-05 ***
freepoor    -0.50254    0.17550  -2.863 0.004191 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 5425.4  on 5185  degrees of freedom
AIC: 7766.9

Number of Fisher Scoring iterations: 6
```

Using this model, we concluded that `sex, age, income, freepoor` are important factors to predict `doctorco`. However, it turned out that the data are overdispersed for the Poisson regression. By ignoring this overdispersion issue in your analysis, what kind of misleading results can you get, and why?

Page 5 of 12 — add any extra pages after page 12 — Page 5 of 12

Page 6 of 12 — add any extra pages after page 12 — Page 6 of 12

Question 3 (12 marks)

We assume that $X_1, \ldots, X_n$ are independent to each other, and each X_i follows a mixture of two normal distributions. Specifically, for $i = 1, \ldots, n$,

$$Z_i \sim \text{categorical } (p, 1-p),$$

$$X_i | Z_i = 1 \sim \text{Normal}(0, 1^2) \text{ and } X_i | Z_i = 2 \sim \text{Normal}(0, \sigma^2),$$

where the Normal distribution with mean μ and variance σ^2 has the probability density function

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We aim to obtain the MLE of the parameters $\theta = (p, \sigma^2)$ using the EM algorithm.

Note: If your final answers contain $P(Z_i = 1 | X_i, \theta^0)$ or $P(Z_i = 2 | X_i, \theta^0)$, you do not need to provide detailed formulae for them, i.e., your final answers can contain these expressions.

(a) Let $X = (X_1, \ldots, X_n)$ and $Z = (Z_1, \ldots, Z_n)$. Derive the expectation of the complete log-likelihood, $Q(\theta, \theta^0) = E_{Z|X,\theta^0}[\log(P(X, Z|\theta))]$.

(b) Derive the M-step of the EM algorithm.

Page 7 of 12 — add any extra pages after page 12 — Page 7 of 12

Page 8 of 12 — add any extra pages after page 12 — Page 8 of 12

Question 4 (14 marks)

Model: Consider a normal model:

$$x \sim \text{Normal}(\mu, \sigma^2).$$

We observe $x = 0$.

(a) **Prior 1:** We impose the following prior for mean and variance parameters:

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

Metropolis-Hastings algorithm: we wish to perform posterior inference using the Metropolis-Hastings algorithm. For the current values of the parameters (μ_c, σ_c^2), we propose new values (μ_n, σ_n^2) as follows: $\sigma_n^2 \sim$ Gamma(shape $= 5\sigma_c^2$, rate $= 5$) and $\mu_n \sim$ Normal(mean $= \mu_c$, variance $= \sigma_n^2$). Compute the acceptance probability when $(\mu_c, \sigma_c^2) = (-1, 2^2), (\mu_n, \sigma_n^2) = (1, 2^2)$. Show your work.

Note: The p.d.f. for the normal distribution is given in Question 3, and the gamma distribution with shape $\nu > 0$ and rate $\lambda > 0$ has the p.d.f.

$$f(x; \nu, \lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x} \quad \text{for } x > 0.$$

Page 9 of 12 — add any extra pages after page 12 — Page 9 of 12

(b) **Prior 2:** We impose the following prior for mean and variance parameters:

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2),$$
$$\mu|\sigma^2 \sim \text{Normal}(0, \sigma^2), \quad \sigma^2 \sim \text{Inverse-Gamma}(1, 1),$$

where the Inverse-Gamma(a, b) has the p.d.f.

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

Variational Inference (VI): we wish to apply VI with the mean-field variational family where $q(\mu, \sigma^2) = q_\mu(\mu)q_{\sigma^2}(\sigma^2)$, and use the CAVI algorithm for optimisation. The CAVI iteratively optimises each factor as follows while holding the other factors fixed:

$$\begin{aligned} q^*_\mu(\mu) &\propto \exp\{\mathbb{E}_{\sigma^2}[\log p(\mu, \sigma^2, x)]\}, \\ q^*_{\sigma^2}(\sigma^2) &\propto \exp\{\mathbb{E}_\mu[\log p(\mu, \sigma^2, x)]\}, \end{aligned}$$

where the expectations $\mathbb{E}_{\sigma^2}$ and $\mathbb{E}_\mu$ are taken with respect to $q^*_{\sigma^2}(\sigma^2)$ and $q^*_\mu(\mu)$, respectively. Derive $q^*_{\sigma^2}(\sigma^2)$ and identify the corresponding distribution name and its parameters. The parameters may contain $\mathbb{E}_\mu[\mu]$ or $\mathbb{V}_\mu[\mu]$, which are the mean and variance of $q^*_\mu(\mu)$, respectively.

There is extra answer space for Question 4 on the next page

Additional answer space only for Question 4—submit this page even if not used

Page 10 of 12 — add any extra pages after page 12 — Page 10 of 12

Page 11 of 12 — add any extra pages after page 12 — Page 11 of 12

Question 5 (11 marks)

Let $X \sim \mathrm{N}(\mu, 1^2)$. The p.d.f. for the normal distribution is given in Question 3. We impose a prior on μ as follows:

$$\begin{cases} \mu \sim \mathrm{N}(0, 1^2) & \text{if } \gamma = 1, \\ \mu = 0 & \text{if } \gamma = 0, \end{cases}$$

and

$$\begin{cases} \gamma = 1 & \text{with probability of } 0.8, \\ \gamma = 0 & \text{otherwise.} \end{cases}$$

We observe $X = 2$. Write down an algorithm to simulate 100 samples from the posterior distribution $p(\mu|X = 2)$. You are not allowed to use a Metropolis-Hastings sampler. You can assume you have access to random number generators for normal, gamma, binomial (including Bernoulli), Poisson, uniform, exponential, beta, and t distributions. If your algorithm simulates samples from a certain distribution, you must specify the exact values for the parameters, and you must provide all derivations of distributions in full.
[Hint: $p(\mu|X = 2) = p(\mu, \gamma = 1|X = 2) + p(\mu, \gamma = 0|X = 2)$.]

There is extra answer space for Question 5 on the next page

Additional answer space only for Question 5—submit this page even if not used

Page 12 of 12 — add any extra pages after page 12 — Page 12 of 12

End of Exam — Total Available Marks = 50