

Student number

Semester 2 Assessment, 2020

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 15 pages (including this page)

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).

Instructions to Students

- There are 5 questions with marks as shown. The total number of marks available is 50.
- During writing time you may only interact with the device running the Zoom session with supervisor permission. The screen of any other device must be visible in Zoom from the start of the session.
- If you have a printer, print the exam one-sided. If you cannot print, download the exam to a second device, which must then be disconnected from the internet.
- Write your answers in the boxes provided on the exam that you have printed or the masked exam template that has been previously made available. If you are unable to answer the whole question in the answer space provided then you can append additional handwritten solutions to the end after the 15 numbered pages. If you do this you MUST make a note in the correct answer space or page for the question, warning the marker that you have appended additional remarks at the end.
- If you have been unable to print the exam and do not have the masked template write your answers on A4 paper. The first page should contain only your student number, the subject code and the 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.
- Assemble all exam pages (or masked template pages) in correct page number order and the correct way up. Add any extra pages with additional working at the end. Use a mobile phone scanning application to scan all pages to a single PDF file. Scan from directly above to reduce keystone effects. Check that all pages are clearly readable and cropped to the A4 borders of the original page. Poorly scanned submissions may be impossible to mark.
- Submit your PDF file to the Canvas Assignment corresponding to this exam using the Gradescope window. Before leaving Zoom supervision, confirm with your Zoom supervisor that you have Gradescope confirmation of submission.

Question 1 (7 marks)

Let X_1, \dots, X_n be independent samples from a Normal distribution $\mathrm{N}(1, \frac{1}{\tau})$ with pdf

$$\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau(x-1)^2}{2}}.$$

(a) What is the log-likelihood for this example?

(h)	What	is the	Fisher	information	on for	this	example?

Question 2 (9 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. Here, 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

Examine the R code and output below, and then answer the questions that follow.

```
> library(faraway)
> data(dvisits)
> modelA <- glm(doctorco ~ sex + age + income + levyplus + freepoor
              + freerepa + illness + actdays,
              family=quasipoisson, data=dvisits)
> summary(modelA)
Call:
glm(formula = doctorco ~ sex + age + income + levyplus + freepoor +
    freerepa + illness + actdays, family = quasipoisson, data = dvisits)
Deviance Residuals:
    Min
              1Q
                   Median
                                ЗQ
                                        Max
-2.7696 -0.6865 -0.5773 -0.4906
                                     5.5745
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.055666
                        0.115808 - 17.751
                                            <2e-16 ***
             0.163442
                        0.064454
                                    2.536
                                            0.0112 *
sex
             0.296311
                        0.186496
                                    1.589
                                            0.1122
age
                                            0.0473 *
income
            -0.195493
                        0.098511 - 1.984
             0.143743
                        0.082153
                                    1.750
                                            0.0802
levyplus
freepoor
            -0.404611
                        0.206938 - 1.955
                                            0.0506
freerepa
             0.118603
                        0.105656
                                    1.123
                                            0.2617
             0.211644
                        0.019482 10.864
                                            <2e-16 ***
illness
             0.133576
                        0.005264 25.377
                                            <2e-16 ***
actdays
```

```
Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
(Dispersion parameter for quasipoisson family taken to be 1.328231)
    Null deviance: 5634.8 on 5189 degrees of freedom
Residual deviance: 4394.3 on 5181 degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 6
> modelB <- glm(doctorco ~ sex + age + income,
              family=quasipoisson, data=dvisits)
> anova(modelB, modelA, test="F")
Analysis of Deviance Table
Model 1: doctorco ~ sex + age + income
Model 2: doctorco ~ sex + age + income + levyplus + freepoor + freerepa +
    illness + actdays
 Resid. Df Resid. Dev Df Deviance
                                            Pr(>F)
                                       F
1
       5186
                5434.9
2
       5181
                4394.3 5
                            1040.5 156.68 < 2.2e-16 ***
Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
> modelC <- glm(doctorco ~ sex + age + income + levyplus + freepoor
              + freerepa + illness + actdays,
              family=poisson, data=dvisits)
> modelD <- glm(doctorco ~ sex + age + income,
              family=poisson, data=dvisits)
> summary(modelD)
Call:
glm(formula = doctorco ~ sex + age + income, family = poisson,
    data = dvisits)
Deviance Residuals:
                  Median
    Min
              1Q
                                3Q
                                        Max
-1.0350 -0.8031 -0.6749 -0.6069
                                     6.3695
Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.71473
                       0.09118 -18.805 < 2e-16 ***
                        0.05589
                                  3.859 0.000114 ***
sex
             0.21565
             1.23798
                        0.13013
                                  9.514 < 2e-16 ***
age
income
           -0.27726
                        0.07969 -3.479 0.000502 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
```

	dual deviance: 5434.9 on 5186 degrees of freedom
n10.	7774.4
Numbe	er of Fisher Scoring iterations: 6
> (pl	niC <- sum(residuals(modelC, type="pearson")^2/modelC\$df.residual))
	1.328231 niD <- sum(residuals(modelD, type="pearson")^2/modelD\$df.residual))
_	2.027811
(a)	Do you prefer modelA or modelB, and why?
(**)	
(b)	Show how the F-statistic has been calculated in the analysis of deviance. What ar degrees of freedom for the F statistic?
(b)	Show how the F-statistic has been calculated in the analysis of deviance. What ar degrees of freedom for the F statistic?
(b)	

	e estimator for t				
For moo	elD, what is the	log-likelihood	of the fitted me	odel, and the log	g-likelihood o
full (sa	turated) model?				

Question 3 (10 marks)

We assume that the observed data, $X_1 = 1, X_2 = 2, X_3 = 4$ follow a mixture of two Poisson distributions. Specifically, for i = 1, 2, 3,

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

$$X_i|Z_i=1 \sim \text{Poisson}(\lambda_1=1.2) \text{ and } X_i|Z_i=2 \sim \text{Poisson}(\lambda_2=2.5),$$

where the Poisson distribution has the probability mass function

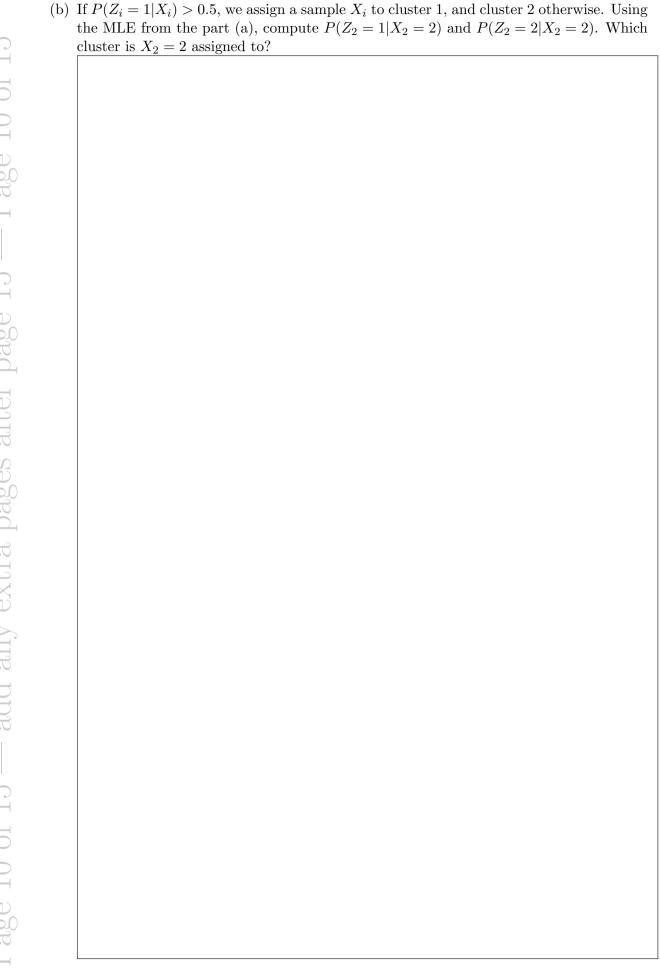
$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Assume that we derived and implemented the EM algorithm to obtain the MLE of the parameter π .

(a) Assume we ran the EM algorithm two times with different initial values. The following table shows two estimates returned by different runs of the EM algorithm. Which estimate should we use as the MLE of the parameter π ? Why?

	initial value	estimate for π
run 1	0.1	0.3
$\operatorname{run} 2$	0.9	0.4





Question 4 (18 marks)

Model: We assume that x and y are independent and follow normal distributions:

$$x \sim N(\mu_1, 1^2)$$
 , $y \sim N(\mu_2, 1^2)$.

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

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

Recall that $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, iff \boldsymbol{x} has joint density

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

(a) We wish to perform posterior inference using the Gibbs sampling. Derive the conditional distribution

$$p(\mu_1|\mu_2, x, y).$$

If it is a known distribution, identify the distribution name and its parameters.

$(2, 0), (\mu_1, \mu_2)$	$\binom{n}{2} = (3,1), x = 1$	-, 9		

(c) We wish to perform posterior inference using variational inference with the mean-field variational family where $q(\mu_1, \mu_2) = q_1(\mu_1)q_2(\mu_2)$ and use the CAVI algorithm for optimisation. The CAVI algorithm iteratively optimises each factor as follows while holding the other factor fixed:

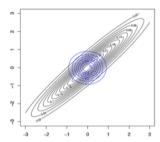
$$q_1^*(\mu_1) \propto \exp\{\mathbb{E}_{\mu_2}[\log p(\mu_1, \mu_2, x, y)]\},$$

 $q_2^*(\mu_2) \propto \exp\{\mathbb{E}_{\mu_1}[\log p(\mu_1, \mu_2, x, y)]\},$

where the expectations \mathbb{E}_{μ_2} and \mathbb{E}_{μ_1} are taken with respect to $q_2^*(\mu_2)$ and $q_1^*(\mu_1)$, respectively. Derive $q_1^*(\mu_1)$ and $q_2^*(\mu_2)$, and identify the corresponding distribution names and their parameters.

Question 5 (6 marks)

Recall that we discussed the problem of approximating the density of the bivariate normal distribution, $p(x_1, x_2)$, using variational inference (VI) with the mean-field variational family (in the tutorial of week 12). We observed that the approximated density, $q(x_1, x_2)$, using VI with the mean-field variational family is very different from the true density when the true correlation between x_1 and x_2 is very strong. The following figure shows the contour plots of $q(x_1, x_2)$ (in blue) and $p(x_1, x_2)$ (in black) where the mean is successfully captured by VI, but the variance of $q(x_1, x_2)$ is controlled by the direction of smallest variance of $p(x_1, x_2)$, and the variance along the orthogonal direction is significantly underestimated. It is a general result that VI tends to underestimate the marginal variance.



It is known that this result is due to the Kullback-Leibler (KL) divergence which VI uses as a measurement of difference between two functions. Recall that VI with the mean-field variational family, $D = \{q(x_1, x_2) | q(x_1, x_2) = q_1(x_1)q_2(x_2)\}$, aims to find $q^*(x_1, x_2) = q_1^*(x_1)q_2^*(x_2)$ which minimises the KL divergence to the exact density $p(x_1, x_2)$,

$$\begin{split} q^*(x_1,x_2) &= \underset{q(x_1,x_2) \in D}{\operatorname{arg\,min}} \quad KL(q(x_1,x_2)||p(x_1,x_2)) \\ &= \underset{q(x_1,x_2) \in D}{\operatorname{arg\,min}} \quad \int_{x_1,x_2} q(x_1,x_2)[\log q(x_1,x_2) - \log p(x_1,x_2)] dx_1 x_2. \end{split}$$

Explain why the KL divergence above results in the underestimation of the marginal variance in the approximated density.