THE UNIVERSITY OF MELBOURNE

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 $X_1, \cdots, X_n$ 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?

$$\ell(\tau) = \sum_{i=1}^{n} \left[\frac{1}{2}\log\left(\frac{\tau}{2\pi}\right) - \frac{\tau(x_i-1)^2}{2} \right]$$

$$= \frac{n}{2}\log(\tau) + \frac{1}{2}\log\left(\frac{1}{2\pi}\right) - \frac{\tau}{2}\sum_{i=1}^{n}(x_i-1)^2$$

(b) What is the Fisher information for this example?

$$U(\tau) = \frac{\partial \ell}{\partial \tau} = \frac{n}{2\tau} - \frac{1}{2}\sum_{i=1}^{n}(x_i-1)^2$$

$$J(\tau) = -\frac{\partial^2 \ell}{\partial \tau^2} = -\left[(-1)\frac{n}{2\tau^2}\right] = \frac{n}{2\tau^2}$$

$$I(\tau) = E[J(\tau)] = \frac{n}{2\tau^2}$$

(c) Find the MLE of τ and its asymptotic distribution.

Let $U(\tau) = \frac{n}{2\tau} - \frac{1}{2}\sum_{i=1}^{n}(x_i - 1)^2 = 0$

$$\Rightarrow \tau = \frac{n}{\sum_{i=1}^{n}(x_i - 1)^2}$$

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       3Q      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 ***
sex          0.163442   0.064454   2.536   0.0112 *
age          0.296311   0.186496   1.589   0.1122
income      -0.195493   0.098511  -1.984   0.0473 *
levyplus     0.143743   0.082153   1.750   0.0802 .
freepoor    -0.404611   0.206938  -1.955   0.0506 .
freerepa     0.118603   0.105656   1.123   0.2617
illness      0.211644   0.019482  10.864   <2e-16 ***
actdays      0.133576   0.005264  25.377   <2e-16 ***
```

```
---
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      F    Pr(>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:
    Min       1Q   Median       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 ***
sex          0.21565    0.05589   3.859 0.000114 ***
age          1.23798    0.13013   9.514  < 2e-16 ***
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)
```

```
    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 5434.9  on 5186  degrees of freedom
AIC: 7774.4

Number of Fisher Scoring iterations: 6

> (phiC <- sum(residuals(modelC, type="pearson")^2/modelC$df.residual))
[1] 1.328231
> (phiD <- sum(residuals(modelD, type="pearson")^2/modelD$df.residual))
[1] 2.027811
```

(a) Do you prefer `modelA` or `modelB`, and why?

```
> 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      F    Pr(>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
```

p-value < 0.05 => reject H_0.

We prefer larger model which is model A.

(b) Show how the F-statistic has been calculated in the analysis of deviance. What are the degrees of freedom for the F statistic?

$$F = \frac{(D^B - D^A)/5}{\hat{\phi}_A} = \frac{(5434.9 - 4394.3)/5}{1.328231} = 156.689612.$$

F has df of 5 and 5181.

$\hat{\phi}_A$:

```
(Dispersion parameter for quasipoisson family taken to be 1.328231)
```

in summary of model A.

(c) Give the estimator for the coefficient of `illness` and its standard error for `modelC`.

Model C is Poisson, so A and C have the same estimate of parameter, but A has larger standard error.

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

↑ in summary of model A

=) Model C estimates coefficient of illness to be 0.211644 with se $0.019482/\sqrt{\phi} = 0.019482/\sqrt{1.328231} = 0.0169$.

(d) For `modelD`, what is the log-likelihood of the fitted model, and the log-likelihood of the full (saturated) model?

```
    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 5434.9  on 5186  degrees of freedom
AIC: 7774.4
```

Model D has deviance of 5434.9 and AIC of 7774.4, $p = 4$.

Since $AIC = 2p - 2\log L(\hat{\beta})$

$7774.4 = 2(4) - 2\log L(\hat{\beta}) \Rightarrow \log L(\hat{\beta}) = -3883.2$

Since $D = -2[\log L(\hat{\beta}) - \log L(S)]$

$5434.9 = -2[-3883.2 - \log L(S)] \Rightarrow \log L(S) = -1165.75$

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
run 2	0.9	0.4

$P(x_i|\theta)$
$= P(x_i|z_i=1,\theta)P(z_i=1|\theta)$
$+ P(x_i|z_i=2,\theta)P(z_i=2|\theta)$.

Choose the one with higher incomplete log likelihood.

$$\ell(\theta) = \sum_{i=1}^{n} \log P(x_i|\theta) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} f(x_i;\lambda_j)\pi_j\right), \quad n=3, \; k=2.$$

$$= \sum_{i=1}^{n} \log\left[f(x_i;\lambda_1)\pi_1 + f(x_i;\lambda_2)\pi_2\right]$$

$$= \sum_{i=1}^{n} \log\left(\frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!}\pi_1 + \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!}\pi_2\right)$$

$$= \log\left(\frac{\lambda_1^{x_1} e^{-\lambda_1}}{x_1!}\pi_1 + \frac{\lambda_2^{x_1} e^{-\lambda_2}}{x_1!}\pi_2\right)$$

$$+ \log\left(\frac{\lambda_1^{x_2} e^{-\lambda_1}}{x_2!}\pi_1 + \frac{\lambda_2^{x_2} e^{-\lambda_2}}{x_2!}\pi_2\right)$$

$$+ \log\left(\frac{\lambda_1^{x_3} e^{-\lambda_1}}{x_3!}\pi_1 + \frac{\lambda_2^{x_3} e^{-\lambda_2}}{x_3!}\pi_2\right)$$

$$= \log\left(\frac{1.2^1 e^{-1.2}}{1!}\pi_1 + \frac{2.5^1 e^{-2.5}}{1!}\pi_2\right)$$
$$+\log\left(\frac{1.2^2 e^{-1.2}}{2!}\pi_1 + \frac{2.5^2 e^{-2.5}}{2!}\pi_2\right)$$
$$+\log\left(\frac{1.2^4 e^{-1.2}}{4!}\pi_1 + \frac{2.5^4 e^{-2.5}}{4!}\pi_2\right)$$

When $\pi_1 = 0.3$, $\pi_2 = 0.7$,

$$\ell(\theta) = -5.075$$

When $\pi_1 = 0.4$, $\pi_2 = 0.6$,

$$\ell(\theta) = -5.1044$$

Hence $\pi_1 = 0.3$ should be chosen.

(b) If $P(Z_i = 1|X_i) > 0.5$, we assign a sample X_i to cluster 1, and cluster 2 otherwise. Using the MLE from the part (a), compute $P(Z_2 = 1|X_2 = 2)$ and $P(Z_2 = 2|X_2 = 2)$. Which cluster is $X_2 = 2$ assigned to?

$$P(Z_i = j | X_i) = \frac{P(X_i|Z_i = j)P(Z_i = j)}{\sum_{j'=1}^{k} P(X_i|Z_i = j')P(Z_i = j')}$$

$$P(Z_2 = 1|X_2 = 2) = \frac{P(X_2 = 2|Z_2 = 1)P(Z_2 = 1)}{P(X_2 = 2|Z_2 = 1)P(Z_2 = 1) + P(X_2 = 2|Z_2 = 2)P(Z_2 = 2)}$$

$$= \frac{\frac{\lambda_1^{x_2} e^{-\lambda_1}}{x_2!}\pi_1}{\frac{\lambda_1^{x_2} e^{-\lambda_1}}{x_2!}\pi_1 + \frac{\lambda_2^{x_2} e^{-\lambda_2}}{x_2!}\pi_2}$$

$$= \frac{\frac{1.2^2 e^{-1.2}}{2!}0.3}{\frac{1.2^2 e^{-1.2}}{2!}0.3 + \frac{2.5^2 e^{-2.5}}{2!}0.7} = 0.266$$

$$P(Z_2 = 2|X_2 = 2) = 1 - P(Z_2 = 1|X_2 = 2) = 0.734$$

$\Rightarrow$ Cluster 2.

Question 4 (18 marks)

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

$$x \sim \mathrm{N}(\mu_1, 1^2) \quad , \quad y \sim \mathrm{N}(\mu_2, 1^2).$$

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.

Joint: $p(\mu_1, \mu_2, x, y) = f_{\mu,\Sigma}(\mu_1, \mu_2) f(x|\mu_1) f(y|\mu_2)$

$= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mu-0)^T\Sigma^{-1}(\mu-0)\right\} \frac{1}{\sqrt{2\pi(1)^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2(1)^2}\right\}$

$\frac{1}{\sqrt{2\pi(1)^2}} \exp\left\{-\frac{(y-\mu_2)^2}{2(1)^2}\right\}.$

where $\Sigma^{-1} = \left(\frac{1}{3}\right)^{-1} \frac{1}{2\times2-(-1)(-1)} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$,

$\mu^T\Sigma^{-1}\mu = [\mu_1 \ \mu_2]\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = 2\mu_1^2 + 2\mu_1\mu_2 + 2\mu_2^2$

$\propto \exp\left\{-\frac{1}{2}(2\mu_1^2 + 2\mu_1\mu_2 + 2\mu_2^2)\right\} \exp\left\{-\frac{1}{2}(x-\mu_1)^2\right\} \exp\left\{-\frac{1}{2}(y-\mu_2)^2\right\}.$

$\propto \exp\left\{-\frac{1}{2}\left(2\mu_1^2 + 2\mu_1\mu_2 + 2\mu_2^2 + x^2 - 2x\mu_1 + \mu_1^2 + y^2 - 2y\mu_2 + \mu_2^2\right)\right\}$

$p(\mu_1|\mu_2, x, y) \propto p(\mu_1, \mu_2, x, y) \propto \exp\left\{-\frac{1}{2}(2\mu_1^2 + 2\mu_1\mu_2 - 2x\mu_1 + \mu_1^2)\right\}$

$\propto \exp\left\{-\frac{1}{2}\left[3\mu_1^2 - 2\mu_1(x-\mu_2)\right]\right\} \propto \exp\left\{-\frac{3}{2}\left(\mu_1 - \frac{x-\mu_2}{3}\right)^2\right\}.$

$\Rightarrow \mu_1|\mu_2, x, y \sim N\left(\frac{x-\mu_2}{3}, \frac{1}{3}\right).$

proposal of μ_1^n does not depend on μ_1^c

(b) We wish to perform posterior inference using the Metropolis-Hastings algorithm. For the current values of the parameters (μ_1^c, μ_2^c), we propose new values (μ_1^n, μ_2^n) as follows: $\mu_1^n \sim \mathrm{N}(0, 1^2)$ and $\mu_2^n \sim \mathrm{N}(\mu_2^c, 1^2)$. Compute the acceptance probability when $(\mu_1^c, \mu_2^c) = (2, 0), (\mu_1^n, \mu_2^n) = (3, 1), x = 1, y = 0$.

$$A = \min\left\{ \frac{\pi(\theta')\, q(\theta', \theta)}{\pi(\theta)\, q(\theta, \theta')}, 1 \right\}$$

$q(\theta, \theta')$ = "density of 3 (μ_1^n) in $N(0, 1^2)$" × "density of 1 (μ_2^n) in $N(0, 1^2)$"

$$= \frac{1}{\sqrt{2\pi(1)^2}} \exp\left\{ -\frac{(3-0)^2}{2(1)^2} \right\} \frac{1}{\sqrt{2\pi(1)^2}} \exp\left\{ -\frac{(1-0)^2}{2(1)^2} \right\}$$

$q(\theta', \theta)$ = "density of 2 (μ_1^c) in $N(0, 1^2)$" × "density of 0 (μ_2^c) in $N(1, 1^2)$"

$$= \frac{1}{\sqrt{2\pi(1)^2}} \exp\left\{ -\frac{(2-0)^2}{2(1)^2} \right\} \frac{1}{\sqrt{2\pi(1)^2}} \exp\left\{ -\frac{(0-1)^2}{2(1)^2} \right\}$$

$\pi(\theta')$ = "density of $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$ from posterior of $\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$"

$\pi(\theta)$ = "density of $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$ from posterior of $\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$"

$$\Rightarrow \frac{\pi(\theta')}{\pi(\theta)} = \frac{p(\mu_1^n, \mu_2^n \mid x, y)}{p(\mu_1^c, \mu_2^c \mid x, y)} = \frac{p(\mu_1^n, \mu_2^n, x, y)}{p(\mu_1^c, \mu_2, x, y)} = \frac{p(3, 1, 1, 0)}{p(2, 0, 1, 0)}$$

$$= \frac{\exp\left\{ -\frac{1}{2}\left(2(3)^2 + 2(3)(1) + 2(1)^2 + (1)^2 - 2(1)(3) + (3)^2 + 0^2 - 2(0)(1) + (1)^2 \right) \right\}}{\exp\left\{ -\frac{1}{2}\left(2(2)^2 + 2(2)(0) + 2(0)^2 + (1)^2 - 2(1)(2) + (2)^2 + 0^2 - 2(0)(0) + 0^2 \right) \right\}}$$

$$\Rightarrow \frac{\pi(\theta')\, q(\theta', \theta)}{\pi(\theta)\, q(\theta, \theta')} = a$$

$$\Rightarrow A = \min\{a, 1\} = \begin{cases} a, & \text{if } a \le 1 \\ 1, & \text{if } a > 1 \end{cases}$$

(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:

$$\begin{aligned} q_1^*(\mu_1) &\propto \exp\{\mathbb{E}_{\mu_2}[\log p(\mu_1, \mu_2, x, y)]\}, \\ q_2^*(\mu_2) &\propto \exp\{\mathbb{E}_{\mu_1}[\log p(\mu_1, \mu_2, x, y)]\}, \end{aligned}$$

where the expectations $\mathbb{E}_{\mu_2}$ and $\mathbb{E}_{\mu_1}$ 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.

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

$= c \cdot \exp\left\{-\frac{1}{2}\left(2\mu_1^2 + 2\mu_1\mu_2 + 2\mu_2^2 + x^2 - 2x\mu_1 + \mu_1^2 + y^2 - 2y\mu_2 + \mu_2^2\right)\right\}$

$\log p(\mu_1, \mu_2, x, y) = -\frac{1}{2}\left(3\mu_1^2 + 3\mu_2^2 + 2\mu_1\mu_2 - 2x\mu_1 - 2y\mu_2 + x^2 + y^2\right) + \log(c)$ — not important

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

$\propto \exp\left\{-\frac{1}{2}\left[3\mu_1^2 + 2\mathbb{E}_{\mu_2}[\mu_2]\mu_1 - 2x\mu_1\right] + c'\right\}$

$\propto \exp\left\{-\frac{3}{2}\left(\mu_1^2 - 2\mu_1 \frac{x - \mathbb{E}_{\mu_2}[\mu_2]}{3}\right)\right\}$

$\propto \exp\left\{-\frac{3}{2}\left(\mu_1 - \frac{x - \mathbb{E}_{\mu_2}[\mu_2]}{3}\right)^2\right\}$

$\Rightarrow q_1^*(\mu_1) \sim \mathcal{N}(\mu_1^*, \sigma_1^{2*})$ where $\mu_1^* = \frac{x - \mathbb{E}_{\mu_2}[\mu_2]}{3}$, $\sigma_1^{2*} = \frac{1}{3}$

Similarly

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

$\propto \exp\left\{-\frac{1}{2}\left[3\mu_2^2 + 2\mathbb{E}_{\mu_1}[\mu_1]\mu_2 - 2y\mu_2\right]\right\}$

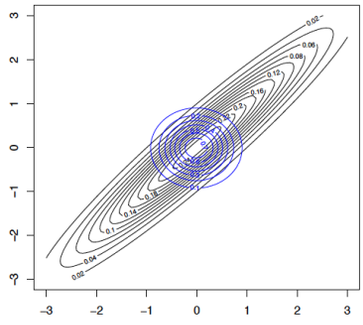
$\propto \exp\left\{-\frac{3}{2}\left(\mu_2^2 - 2\mu_2 \frac{y - \mathbb{E}_{\mu_1}[\mu_1]}{3}\right)\right\}$

$\propto \exp\left\{-\frac{3}{2}\left(\mu_2 - \frac{y - \mathbb{E}_{\mu_1}[\mu_1]}{3}\right)^2\right\}$

$\Rightarrow q_2^*(\mu_2) \sim \mathcal{N}(\mu_2^*, \sigma_2^{2*})$ where $\mu_2^* = \frac{y - \mathbb{E}_{\mu_1}[\mu_1]}{3}$, $\sigma_2^{2*} = \frac{1}{3}$.

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 $p(x_1, x_2)$,

$$\begin{aligned} q^*(x_1, x_2) &= \underset{q(x_1,x_2)\in D}{\arg\min} \quad KL(q(x_1, x_2)||p(x_1, x_2)) \\ &= \underset{q(x_1,x_2)\in D}{\arg\min} \quad \int_{x_1,x_2} q(x_1, x_2)[\log q(x_1, x_2) - \log p(x_1, x_2)]dx_1x_2. \end{aligned}$$

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

When p is small but q is large, KL is very high, hence this region contributes a large portion of total KL. As a result of optimisation, q would concentrate its density to where p is high, not covering low density region of p.

This is also partly due to our choice of mean-field family of distributions which assumes all parameters are mutually exclusive. For example in the above plot, to cover the top right region which

has relatively low density of p, contour of q would have to cover bottom right and top left region as well, which have even lower density of p. This is because in the above example, contour of q from mean-field would form a concentric circle. The algorithm would deem this unworthy as it would create a large amount of KL.
In general, overestimating p when p is high is less risky than overestimating p when p is low according to KL.

End of Exam — Total Available Marks = 50