STAT 215 – Midterm 2 NAME:

In your solutions, please use notation like $N(y|m,S)$ to indicate a normal p.d.f. with moments (m,S) evaluated at y, etc.

1. [10pts] Consider the following harmonic regression model for observations y_i at times t_i:

$$y_i = a\ \cos(\omega t_i) + b\ \sin(\omega t_i) + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0,\sigma^2),$$

$i = 1,\dots,n$. Let $\beta = (a,b)$. We complete the model with a non-informative prior

$$p(\beta,\sigma^2,\omega) \propto 1/\sigma^2\ p(\omega).$$

Let $y = (y_1,\dots,y_n)$ denote the data vector.

In your solution use Z for the $n\times 2$ design matrix with $Z_{i1} = \cos(\omega t_i)$ and $Z_{i2} = \sin(\omega t_i)$.

1a. [2.5] Find $p(\beta|\omega,\sigma^2,y)$.

Rewrite the model as $y = Z\beta + \epsilon$ and recognize it a normal linear regression model with traditional reference prior. Let $\widehat{\beta} = (Z'Z)^{-1}Z'y$ and $V_\beta = (Z'Z)^{-1}$. Then:

$$p(\beta|\omega,\sigma^2,y) = N(\widehat{\beta},\sigma^2 V_\beta)$$

1b. [2.5] Find $p(\beta,\omega,\sigma^2|y)$.

$$p(\beta,\omega,\sigma^2|y) \propto p(\beta,\sigma^2)p(\omega)p(y|\beta,\sigma^2,\omega) = p(\omega)\frac{1}{\sigma^2}N(y|Z\beta,\sigma^2 I).$$

1c. [2.5] Find $p(\omega,\sigma^2|y)$.

$$p(\omega,\sigma^2|y) = \frac{p(\beta,\omega,\sigma^2|y)}{p(\beta|\omega,\sigma^2,y)} = \frac{\frac{1}{\sigma^2}N(y|Z\beta,\sigma^2 I)\ p(\omega)}{N(\beta|\widehat{\beta},\sigma^2 V_\beta)} =$$
$$= N(y|Z\widehat{\beta},\sigma^2 I)|V_\beta|^{1/2}p(\omega) \propto \left(\sigma^2\right)^{-n/2} e^{\frac{1}{2\sigma^2}(y-Z\widehat{\beta})'(y-Z\widehat{\beta})}|V_\beta|^{1/2}p(\omega) \propto \left(\sigma^2\right)^{-n/2}|V_\beta|^{1/2}p(\omega)e^{-\frac{1}{2\sigma^2}(y'y-\widehat{\beta}Z'Z\widehat{\beta})}$$

The first equality is candidate's formula. For the third equality we evaluate candidate's formula for $\beta = \widehat{\beta}$ (the ratio is the same for all values of β). For the last equality use the identity $(y - Z\widehat{\beta})'(y - Z\widehat{\beta}) = \ldots = y'y - \widehat{\beta}Z'Z\widehat{\beta}$.

1d. [2.5] Find the marginal posterior $p(\omega|y)$.

$$p(\omega|y) = \int p(\omega,\sigma^2|y)d\sigma^2 \propto p(\omega)\int \left(\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}(y'y-\widehat{\beta}Z'Z\widehat{\beta})}|V_\beta|^{1/2}d\sigma^2 =$$
$$= \left(y'y - \widehat{\beta}Z'Z\widehat{\beta}\right)^{-(n-2)/2}|V_\beta|^{1/2} \propto |V_\beta|^{1/2}\left[1 - \frac{\widehat{\beta}'Z'Z\widehat{\beta}}{y'y}\right]^{-(n-2)/2} \qquad (1)$$

Note: If $t_i = i$ and $\omega = j/n$ for some $j < n/2$ then (using simple trigonometric identities) $Z'Z = \frac{n}{2}I$ and we find

$$\widehat{a} = \frac{2}{n}\sum_{i=1}^{n} y_i\cos(\omega t_i),\ \widehat{b} = \frac{2}{n}\sum_{i=1}^{n} y_i\sin(\omega t_i),$$

and

$$p(\omega|\sigma^2,y) \propto e^{I(\omega)},$$

where $I(\omega) = \frac{n}{2}[\widehat{a}^2 + \widehat{b}^2]$ is the periodogram evaluated at ω. Thus, the traditional periodogram can be interpreted as a log posterior in a harmonic regression model.

2. [10pts] Consider two competing models: H_1: a linear; H_2 : a quadratic regression.

Model H_1:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i + \epsilon_i, \quad i = 1, \dots, n \\ & \quad \epsilon_i \sim N(0,1) \\ \beta_1 &\sim N(0,1), \quad \beta_2 \sim N(1,1), \end{aligned}$$

with β_1 and β_2 *a priori* independent.

Model H_2:

$$\begin{aligned} y_i &= \gamma_1 + \gamma_2 x_i + \gamma_3 x_i^2 + \epsilon_i \\ & \quad \epsilon_i \sim N(0,1) \\ \gamma_1 &= N(0,1), \quad \gamma_2 \sim N(1,1), \quad \gamma_3 \sim N(0,1), \end{aligned}$$

with $\gamma_1, \gamma_2, \gamma_3$ *a priori* independent.

2a. [5] Find the marginal distribution $p(y|H_1) = \int p(y|\beta)p(\beta)d\beta$. and $p(y|H_2) = \int p(y|\gamma)p(\gamma)d\gamma$.

Denote with X_1 the $n \times 2$ design matrix with 1 in the first column and x in the second column, and with X_2 the $n \times 3$ design matrix with columns 1,x and x^2. Let $\mu_1 = (0,1)'$, $\mu_2 = (0,1,0)$.
Then (for example, by Lemma 1 on the HW9 solution) $p(y|H_1) = N(y|X_1\mu_1, I + X_1X_1')$ and $p(y|H_2) = N(y|X_2\mu_2, I + X_2X_2')$.

2b. [5] Write down the Bayes factor $B = p(y|H_2)/p(y|H_1)$ for comparing model H_1 vs. model H_2.

$$B = \frac{N(y|X_2\mu_2, I + X_2X_2')}{N(y|X_1\mu_1, I + X_1X_1')}.$$

3. [10pts] Let $x_1, \dots x_n$ be an i.i.d. sample from a Poisson $P(\theta)$ distribution.
Let $q(\theta) = P(x_1 = 0) = e^{-\theta}$. Denote with $x = (x_1, \dots, x_n)$.

Hint: Remember that $\sum x_i \sim P(n\theta)$.

3a. [3] Find an unbiased estimator $S(x)$ for $q(\theta)$.

Use, for example, the indicator $S(x) = I(x_1 = 0)$. $E[S(x)] = E[I(x_1 = 0)] = Pr(x_1 = 0)$.

3b. [3] Show that $T(x) = \sum_{i=1}^n x_i$ is a complete sufficient statistic for θ.

Note that

$$p_\theta(x_i) = \frac{1}{x_i!}e^{-\theta}\theta^{x_i} = \underbrace{1/x_i!}_{f(x_i)}\underbrace{\exp(-\theta)}_{g(\theta)}\exp[\underbrace{\log(\theta)}_{\phi(\theta)}\underbrace{x_i}_{u(x_i)}]$$

is a 1 parameter exponential family with natural sufficient statistic $u(x_i) = x_i$. Thus $T(x) = \sum_{i=1}^n x_i$ is a complete sufficient statistic.

3c. [4] Find the UMVU estimator for $q(\theta)$.

Let $t = \sum x_i$. The UMVUE is $T^*(x) = [(n-1)/n]^t$:

From Lehmann-Scheffé's theorem we know that $T^*(x) = E[S(x)|T(x)]$ is the UMVUE.

$$\begin{aligned} E[S(x)|T = t] &= P(x_1 = 0|\sum_{i=1}^n x_i = t) = \\ &= \frac{P(x_1 = 0, \sum_{i=2}^n x_i = t)}{P(\sum_{i=1}^n x_i = t)} = \frac{e^{-\theta} \times 1/t![(n-1)\theta]^t e^{-(n-1)\theta}}{1/t![n\theta]^t e^{-n\theta}} = [(n-1)/n]^t. \end{aligned}$$

The second last equality follows from the fact that $y = \sum_{i=2}^n x_i \sim P[(n-1)\theta]$ and $z = \sum_{i=1}^n x_i \sim P[n\theta]$.