University of California, Santa Cruz
Department of Applied Mathematics and Statistics
Baskin School of Engineering

AMS 206 (Applied Bayesian Statistics)

Take-Home Test 2 (with 2 typos corrected and software help)

Target due date **6 Mar 2018**; drop-dead due date **13 Mar 2018**

Here are the ground rules: this test is open-book and open-notes, and consists of two problems (true/false and calculation); **each of the 6 true/false questions is worth 10 points, and the calculation problem is worth 126 total points, for a total of 186 points**.

The right answer with no reasoning to support it, or incorrect reasoning, will get **half credit**, so try to make a serious effort on each part of each problem (this will ensure you at least half credit). In an AMS graduate class I taught in 2012, on a take-home test like this one there were 15 true/false questions, worth a total of 150 points; one student got a score of 92 out of 150 (61%, a D−, in a graduate class where B− is the lowest passing grade) on that part of the test, for repeatedly answering just "true" or "false" with no explanation. Don't let that happen to you.

On non-extra-credit problems, I mentally start everybody out at −0 (i.e., with a perfect score), and then you accumulate negative points for incorrect answers and/or reasoning, or parts of problems left blank. On extra-credit problems, the usual outcome is that you go forward (in the sense that your overall score goes up) or you at least stay level, but please note that it's also possible to go backwards on such problems (e.g., if you accumulate +3 for part of an extra-credit problem but −4 for the rest of it, for saying or doing something egregiously wrong).

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me or our TA (Daniel Kirsner). The intent is that the course lecture notes and readings should be sufficient to provide you with all the guidance you need to solve the problems posed below, but you may use other written materials (e.g., the web, journal articles, and books other than those already mentioned in the readings), **provided that you cite your sources thoroughly and accurately**; you will lose (substantial) credit for, e.g., lifting blocks of text directly from `wikipedia` and inserting them into your solutions without full attribution.

If it's clear that (for example) two people have worked together on a part of a problem that's worth 20 points, and each answer would have earned 16 points if it had not arisen from a collaboration, then each person will receive 8 of the 16 points collectively earned (for a total score of 8 out of 20), and I reserve the right to impose additional penalties at my discretion. If you solve a problem on your own and then share your solution with anyone else (because people from your cultural background routinely do this, or out of pity, or kindness, or whatever motive you may believe you have; it doesn't matter), you're just as guilty of illegal collaboration as the person who took your solution from you, and both of you will receive the same penalty. This sort of thing is necessary on behalf of the many people who do not cheat, to ensure that their scores are meaningfully earned. In the AMS graduate class in 2012 mentioned above, five people failed the class because of illegal collaboration; don't let that happen to you.

In class I've demonstrated numerical work in `R`; you can (of course) make the calculations and plots requested in the problems below in any environment you prefer (e.g., `Matlab`, ...).

Please collect {all of the code you used in answering the questions below} into an Appendix at the end of your document, so that (if you do something wrong) the grader can better give you part credit. To avoid plagiarism, if you end up using any of the code I post on the course web page or generate during office hours, at the beginning of your Appendix you can say something like the following:

> *I used some of Professor Draper's* `R` *code in this assignment, adapting it as needed.*

Last ground rule: proof by `Maple` or some other symbolic computing package is not acceptable; when I ask You to show something, please do so by hand (You can check Your results with (e.g.) `Maple`, but You need to do the work Yourself).

1 True/False

[60 total points: 10 points each] For each statement below, say whether it's true or false; if true without further assumptions, briefly explain why it's true (and — *extra credit* — what its implications are for statistical inference); if it's sometimes true, give the extra conditions necessary to make it true; if it's false, briefly explain how to change it so that it's true and/or give an example of why it's false. If the statement consists of two or more sub-statements and two or more of them are false, you need to explicitly address all of the false sub-statements in your answer.

(A) Consider the sampling model $(Y_i \,|\, \boldsymbol{\theta}\, \mathcal{B}) \stackrel{\text{IID}}{\sim} p(y_i \,|\, \boldsymbol{\theta}\, \mathcal{B})$ for $i = 1, \ldots, n$, where the Y_i are univariate real values, $\boldsymbol{\theta}$ is a parameter vector of length $1 \leq k < \infty$ and $\mathcal{B}$ summarizes Your background information; a Bayesian analysis with the same sampling model would add a prior distribution layer of the form $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ to the hierarchy. The Bernstein-von Mises theorem says that maximum-likelihood (ML) and Bayesian inferential conclusions about $\boldsymbol{\theta}$ will be similar in this setting if (a) n is large and (b) $p(\boldsymbol{\theta})$ is diffuse (low information content), but the theorem does not provide guidance on how large n needs to be for its conclusion to hold in any specific sampling model.

(B) In the basic diagram that illustrates the frequentist inferential paradigm — with the population, sample and repeated-sampling data sets, each containing N, n, and M elements, respectively (see page 2 of the document camera notes from 25 Jan 2018) — when the population parameter of main interest is the mean θ and the estimator is the sample mean $\bar{Y}$, You will always get a Gaussian long-run distribution for $\bar{Y}$ (in the repeated-sampling data set) as long as any one of (N, n, M) goes to infinity.

(C) Being able to express Your sampling distribution as a member of the Exponential Family is helpful, because

- You can then readily identify a set of sufficient statistics, and
- a conjugate prior always then exists and can be identified,

in both cases just by looking at the form of the Exponential Family.

(D) When the sampling model is a regular parametric family $p(\boldsymbol{Y} \,|\, \boldsymbol{\theta}\, \mathcal{B})$, where $\boldsymbol{\theta}$ is a vector of length $1 < k < \infty$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, then for large n the repeated-sampling distribution of the (vector) MLE $\hat{\boldsymbol{\theta}}_{MLE}$ is approximately k–variate normal with mean vector $\boldsymbol{\theta}$ and covariance matrix $\hat{I}^{-1}$ (the inverse of the observed information matrix), and the bias of $\hat{\boldsymbol{\theta}}_{MLE}$ as an estimate of $\boldsymbol{\theta}$ in large samples is $O\big(\frac{1}{n^2}\big)$.

(E) It's easier to reason from the part (or the particular, or the sample) to the whole (or the general, or the population), and that's why statistical inference (inductive reasoning) is easier than probability (deductive reasoning).

(F) When Your sampling model has n observations and a single parameter θ (so that $k = 1$), if the sampling model is regular (i.e., if the range of possible data values doesn't depend on θ), in large samples the observed information $\hat{I}\Big(\hat{\theta}_{MLE}\Big)$ is $O(n)$, meaning that

- information in $\hat{\theta}_{MLE}$ about θ increases linearly with n, and
- the repeated-sampling variance $\hat{V}_{RS}\Big(\hat{\theta}_{MLE}\Big)$ is $O\big(\frac{1}{n}\big)$.

2 Calculation

(A) *[78 total points, plus a total of 25 possible extra-credit points]* (Based on a problem in Gelman et al. (2014)) In late October 1988, a survey was conducted on behalf of *CBS News* of $n = 1{,}447$ adults aged 18+ in the United States, to ask about their preferences in the upcoming presidential election. Out of the 1,447 people in the sample, $n_1 = 727$ supported George H.W. Bush, $n_2 = 583$ supported Michael Dukakis, and $n_3 = 137$ supported other candidates or expressed no opinion. The polling organization used a sampling method called *stratified random sampling* that's more complicated than the two sampling methods we know about in this class — IID sampling (at random with replacement) and simple random sampling (SRS: at random without replacement) — but here let's pretend that they used SRS from the population $\mathcal{P} = \{$all American people of voting age in the U.S. in October 1988$\}$. There were about 245 million Americans in 1988, of whom about 74% were 18 or older, so $\mathcal{P}$ had about 181 million people in it; the total sample size of $n = 1{,}447$ is so small in relation to the population size that we can regard the sampling as effectively IID.

Under these conditions it can be shown, via a generalization of de Finetti's Theorem for binary outcomes, that the only appropriate sampling distribution for the data vector $\boldsymbol{N} = (n_1, n_2, n_3)$ is a generalization of the Binomial distribution called the *Multinomial* distribution (You can look back in Your AMS 131 notes, or the AMS 131 book, to renew Your acquaintance with the Multinomial). Suppose that a population of interest contains items of $p \geq 2$ types (in the example here: people who support {Bush, Dukakis, other}, so that in this case $p = 3$) and that the population proportion of items of type j is $0 < \theta_j < 1$. Letting $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, note that there's a restriction on the components of $\boldsymbol{\theta}$, namely $\sum_{j=1}^{p} \theta_j = 1$. Now, as in the *CBS News* example, suppose that someone takes an IID sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ of size n from this population and counts how many elements in the sample are of type 1 (call this count n_1), type 2 (n_2), and so on up to type p (n_p); let $\boldsymbol{N} = (n_1, \ldots, n_p)$ be the (vector) random variable that keeps track of all of the counts. In this situation people say that $\boldsymbol{N}$ follows the Multinomial distribution

Table 1: *The Binomial as a special case of the Multinomial: notational correspondence.*

Binomial	Multinomial ($p = 2$)
n	n
x	n_1
$(n - x)$	n_2
θ	θ_1
$(1 - \theta)$	θ_2

with parameters n and $\boldsymbol{\theta}$, which is defined as follows: $(\boldsymbol{N} \,|\, n\, \boldsymbol{\theta}\, \mathcal{B}) \sim \text{Multinomial}(n, \boldsymbol{\theta})$ iff

$$P(N_1 = n_1, \ldots, N_p = n_p \,|\, n\, \boldsymbol{\theta}\, \mathcal{B}) = \left\{ \begin{array}{cc} \frac{n!}{n_1!\, n_2! \cdots n_p!}\, \theta_1^{n_1}\, \theta_2^{n_2} \cdots \theta_p^{n_p} & \text{if } n_1 + \cdots + n_p = n \\ 0 & \text{otherwise} \end{array} \right\}, \quad (1)$$

with the further restriction that $0 \leq n_j \leq N$ (for all $j = 1, \ldots, p$). The main scientific interest in this problem focuses on $\gamma = (\theta_1 - \theta_2)$, the margin by which Bush was leading Dukakis on the day of the survey.

(a) *[4 total points for this sub-problem]* Show that the Multinomial is indeed a direct generalization of the Binomial, if we're careful in the notational conventions we adopt. Here's what I mean: the Binomial distribution arises when somebody makes n IID success–failure (Bernoulli) trials, each with success probability θ, and records the number X of successes; this yields the sampling distribution

$$(X \,|\, \theta\, \mathcal{B}) \sim \text{Binomial}(n, \theta) \text{ iff } P(X = x \,|\, \theta\, \mathcal{B}) = \left\{ \begin{array}{cc} \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x\, (1 - \theta)^{n-x} & \text{for } x = 0, \ldots, n \\ 0 & \text{otherwise} \end{array} \right\}. \quad (2)$$

Briefly and carefully explain why the correspondence between equation (2) and {a version of equation (1) with $p = 2$} is as in Table 1 **(typos corrected in this table)** *[4 points]*.

(b) *[10 total points for this sub-problem, plus up to 10 possible extra credit points]* Returning now to the general Multinomial setting, briefly explain why the likelihood function for $\boldsymbol{\theta}$ given $\boldsymbol{N}$ and $\mathcal{B}$ is

$$\ell(\boldsymbol{\theta} \,|\, \boldsymbol{N}\, \mathcal{B}) = c \prod_{j=1}^{p} \theta_j^{n_j}\,, \quad (3)$$

leading to the log-likelihood function (ignoring the irrelevant constant)

$$\ell\ell(\boldsymbol{\theta} \,|\, \boldsymbol{N}\, \mathcal{B}) = \sum_{j=1}^{p} n_j\, \log \theta_j\,. \quad (4)$$

[4 points]. In finding the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, if You simply try, as usual, to set all of the first partial derivatives of $\ell\ell(\boldsymbol{\theta} \,|\, \boldsymbol{N}\, \mathcal{B})$ with respect to the θ_j equal to 0, You'll get a system of equations that has no solution (try it). This is because in so doing we forgot that we need to do a *constrained optimization*, in which the constraint is $\sum_{j=1}^{p} \theta_j = 1$. There are thus two ways forward to compute the MLE:

(i) Solve the constrained optimization problem directly with *Lagrange multipliers (Extra credit [5 points]: do this)*, or

(ii) build the constraint directly into the likelihood function: define

$$\ell(\theta_1, \ldots, \theta_{p-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = c \left(\prod_{j=1}^{p-1} \theta_j^{n_j} \right) \left(1 - \sum_{j=1}^{p-1} \theta_j \right)^{n_p} , \tag{5}$$

from which (ignoring the irrelevant constant) **(typo corrected in equation (6))**

$$\ell\ell(\theta_1, \ldots, \theta_{p-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = \sum_{j=1}^{p-1} n_j \, \log \theta_j + n_p \log \left(1 - \sum_{j=1}^{p-1} \theta_j \right) . \tag{6}$$

For $j = 1, \ldots, (p-1)$, show that

$$\frac{\partial}{\partial\, \theta_j} \, \ell\ell(\theta_1, \ldots, \theta_{p-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = \frac{n_j}{\theta_j} - \frac{n_p}{1 - \sum_{i=1}^{p-1} \theta_i} \tag{7}$$

[2 points]. The MLE for $(\theta_1, \ldots, \theta_{p-1})$ may now be found by setting $\frac{\partial}{\partial \theta_j} \, \ell\ell(\theta_1, \ldots, \theta_{p-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = 0$ for $j = 1, \ldots, (p-1)$ and solving the resulting system of $(p-1)$ equations in $(p-1)$ unknowns *(Extra credit [5 points]: do this for general p)*, but that gets quite messy; let's just do it for $p = 3$, which is all we need for the CBS survey anyway. Solve the two equations

$$\left\{ \frac{n_1}{\theta_1} - \frac{n_3}{1 - \theta_1 - \theta_2} = 0, \quad \frac{n_2}{\theta_2} - \frac{n_3}{1 - \theta_1 - \theta_2} = 0 \right\} \tag{8}$$

for (θ_1, θ_2) and then use the constraints $\sum_{j=1}^{3} \theta_j = 1$ and $\sum_{j=1}^{3} n_j = n$ to get the MLE for θ_3, thereby demonstrating the (entirely obvious, after the fact) result that

$$\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3 \right) = \left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n} \right) \tag{9}$$

[4 points]. (The result for general p, of course, is that $\hat{\boldsymbol{\theta}} = \frac{1}{n} \boldsymbol{N}$.)

(c) *[16 total points for this sub-problem, plus up to 5 possible extra credit points]* It can be shown *(Extra credit [5 points]: do this for general p, by working out the negative Hessian, evaluated at the MLE, to get the information matrix $\hat{\boldsymbol{I}}$ and then inverting $\hat{\boldsymbol{I}}$)* that in repeated sampling the estimated covariance matrix of the MLE vector $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3 \right)$ is

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \frac{\hat{\theta}_1 (1 - \hat{\theta}_1)}{n} & -\frac{\hat{\theta}_1 \, \hat{\theta}_2}{n} & -\frac{\hat{\theta}_1 \, \hat{\theta}_3}{n} \\ -\frac{\hat{\theta}_1 \, \hat{\theta}_2}{n} & \frac{\hat{\theta}_2 (1 - \hat{\theta}_2)}{n} & -\frac{\hat{\theta}_2 \, \hat{\theta}_3}{n} \\ -\frac{\hat{\theta}_1 \, \hat{\theta}_3}{n} & -\frac{\hat{\theta}_2 \, \hat{\theta}_3}{n} & \frac{\hat{\theta}_3 (1 - \hat{\theta}_3)}{n} \end{pmatrix} . \tag{10}$$

Explain why the form of the diagonal elements of $\hat{\boldsymbol{\Sigma}}$ makes good intuitive sense (by thinking about the corresponding results when there are only $p = 2$ outcome categories); also explain why it makes good sense that the off-diagonal elements of $\hat{\boldsymbol{\Sigma}}$ are negative *[4 points].* Use $\hat{\boldsymbol{\Sigma}}$ to compute approximate large-sample standard errors for the MLEs of the θ_i and of γ; for $\widehat{SE}(\hat{\gamma})$ You can either

(i) work it out directly by thinking about the repeated-sampling variance of the difference of two (correlated) random quantities, or

(ii) use the fact (from AMS 131) that if $\hat{\boldsymbol{\theta}}$ is a random vector with covariance matrix $\hat{\boldsymbol{\Sigma}}$ and $\gamma = \boldsymbol{a}^T \boldsymbol{\theta}$ for some vector $\boldsymbol{a}$ of constants, then in repeated sampling

$$\hat{V}(\hat{\gamma}) = \hat{V}\left(\boldsymbol{a}^T \hat{\boldsymbol{\theta}}\right) = \boldsymbol{a}^T \hat{\boldsymbol{\Sigma}}\, \boldsymbol{a} \tag{11}$$

[4 points]. Finally, use Your estimated SE for $\hat{\gamma}$ to construct an approximate (large-sample) 95% confidence interval for γ *[4 points].* Was Bush ahead of Dukakis at the point when the survey was conducted by an amount that was large in practical terms? Was Bush's lead at that point statistically significant? Explain briefly. *[4 points]*

(d) *[8 total points for this sub-problem]* Looking back at equation (3), if a conjugate prior exists for the Multinomial likelihood it would have to be of the form

> θ_1 to a power times θ_2 to a (possibly different) power times ... times θ_p to a (possibly different) power.

There is such a distribution — it's called the *Dirichlet*($\boldsymbol{\alpha}$) distribution, with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ chosen so that all of the α_i are positive:

$$p(\boldsymbol{\theta} \,|\, \mathcal{B}) = c \prod_{i=1}^{p} \theta_i^{\alpha_i - 1} \,. \tag{12}$$

Briefly explain why this means that the conjugate updating rule is

$$\left\{ \begin{array}{ccc} (\boldsymbol{\theta} \,|\, \mathcal{D}) & \sim & \text{Dirichlet}(\boldsymbol{\alpha}) \\ (\boldsymbol{N} \,|\, \boldsymbol{\theta}\, \mathcal{B}) & \sim & \text{Multinomial}(n, \boldsymbol{\theta}) \end{array} \right\} \longrightarrow (\boldsymbol{\theta} \,|\, \boldsymbol{N}\, \mathcal{D}\, \mathcal{B}) \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{N}) \,, \tag{13}$$

in which $\mathcal{D}$ stands for the Dirichlet prior disribution assumption, which is not part of $\mathcal{B}$ *[4 points].* Given that $\boldsymbol{N} = (n_1, \ldots, n_p)$ and that the n_j represent sample sizes (numbers of observations y_i) in each of the p Multinomial categories, briefly explain why this implies that, if context suggests a low-information-content prior, this would correspond to choosing the α_i all close to 0. *[4 points]*

(e) *[40 total points for this sub-problem, plus up to 10 possible extra credit points]* Briefly explain why, if You have a valid way of sampling from the Dirichlet distribution, it's not necessary in this problem in fitting model (13) to do MCMC sampling: IID Monte Carlo sampling is sufficient *[4 points].* It turns out that the following is a valid way to sample a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ from the Dirichlet($\boldsymbol{\alpha}$) distribution:

- pick any $\beta > 0$ of Your choosing ($\beta = 1$ is a good choice that leads to fast random number generation);
- for ($j = 1, \ldots, p$), make p independent draws g_j with draw j from the $\Gamma(\alpha_j, \beta)$ distribution; and
- then just normalize:

$$g_j \overset{\text{I}}{\sim} \Gamma(\alpha_j, \beta) \quad \text{and} \quad \theta_j = \frac{g_j}{\sum_{k=1}^{p} g_k} \,, \tag{14}$$

in which $\overset{\text{I}}{\sim}$ means *are independently distributed as.*

(Help with the code added here) I've written an `R` function called `rdirichlet`, posted on the course web page, that implements this algorithm. Use my function (or an equivalent in Your favorite non-`R` environment) to generate M IID draws from the posterior distribution specified by model (13), using the CBS News polling data and a diffuse Dirichlet($\boldsymbol{\alpha}$) prior with $\boldsymbol{\alpha} = (\epsilon, \ldots, \epsilon)$ for some small $\epsilon > 0$ such as 0.01; in addition to monitoring the components of $\boldsymbol{\theta}$, also monitor $\gamma = (\theta_1 - \theta_2)$ *[10 points]*. Choose a value of M large enough so that the Monte Carlo standard errors of the posterior means of γ and the components of $\boldsymbol{\theta}$ are small enough to make all of the posterior mean estimates reliable to at least 3 significant figures, and justify Your choice *[4 points]*. Make graphical and numerical summaries of the posterior distributions for γ and for each of the components of $\boldsymbol{\theta}$, and compare Your posterior distribution for γ with Figure 3.2 (p. 70) from the Gelman et al. (2014) book that's available at the course web site *[10 points]*. How do Your Bayesian answers compare with those from maximum likelihood in this problem? Explain briefly *[4 points]*. Compute a Monte Carlo estimate of $p(\gamma > 0 \,|\, \boldsymbol{N}\,\mathcal{D}\,\mathcal{B})$, which quantifies the current information about whether Bush is leading Dukakis in the population of all adult Americans, and attach a Monte Carlo standard error to Your estimate *[4 points]*. What substantive conclusions do You draw about where the Presidential race stood in late October of 1988, on the basis of Your analysis? Explain briefly. *[4 points] (Extra credit [10 points]: Use* `Maple` *or some equivalent environment (or paper and pen, if You're brave) to see if You can derive a closed-form expression for $p(\gamma > 0 \,|\, \boldsymbol{N}\,\mathcal{D}\,\mathcal{B})$, and compare Your mathematical result with Your simulation-based findings; if no such expression is possible, briefly explain why not.)*

(B) *[48 total points, plus a total of 14 possible extra-credit points]* In this problem I'll guide You through a likelihood analysis of the NB10 data, to illustrate several important aspects of the maximum likelihood story that we haven't looked at carefully yet: working with a vector $\boldsymbol{\theta}$ of unknowns of length $k > 1$, and maximizing a log-likelihood function numerically.

(a) *[12 total points for this sub-problem]* Based on the discussion in class, with $\boldsymbol{y} = (y_1, \ldots, y_n)$ as the observed vector of weighings of NB10, our sampling distribution is $(Y_i \,|\, \mu\,\sigma\,\nu\,\mathcal{T}\,\mathcal{B}) \overset{\text{IID}}{\sim} t_\nu(\mu, \sigma)$, in which $\mathcal{T}$ denotes the scaled t sampling distribution assumption (which is not part of $\mathcal{B}$). From the form of the scaled t density (see Appendix A in Gelman et al. (2014)), and letting $\boldsymbol{\theta} = (\mu, \sigma, \nu)$ (so that $k = 3$), the likelihood function is

$$\ell(\boldsymbol{\theta} \,|\, \boldsymbol{y}\,\mathcal{T}\,\mathcal{B}) = \ell(\mu\,\sigma\,\nu \,|\, \boldsymbol{y}\,\mathcal{T}\,\mathcal{B}) = \prod_{i=1}^{n} \left\{ \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\,\sqrt{\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{1}{\nu} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right]^{-\left(\frac{\nu+1}{2}\right)} \right\} . \qquad (15)$$

When viewed as a joint sampling distribution, is equation (15) a member of the 3–parameter exponential family? Explain briefly *[4 points]*. Show that equation (15) leads to the log-likelihood function

$$\begin{aligned} \ell\ell(\boldsymbol{\theta} \,|\, \boldsymbol{y}\,\mathcal{T}\,\mathcal{B}) \;\; &= \;\; \ell\ell(\mu\,\sigma\,\nu \,|\, \boldsymbol{y}\,\mathcal{T}\,\mathcal{B}) = n\,\log\Gamma\left(\frac{\nu+1}{2}\right) - n\,\log\sigma - n\,\log\Gamma\left(\frac{\nu}{2}\right) \\ &\quad -\frac{n}{2}\log\nu - \left(\frac{\nu+1}{2}\right)\sum_{i=1}^{n}\log\left[1 + \frac{1}{\nu}\left(\frac{y_i - \mu}{\sigma}\right)^2\right] \qquad (16) \end{aligned}$$

[4 points]. Can You find a 3–dimensional set of sufficient statistics with this sampling model? Explain briefly *[4 points]*.

(b) *[4 total points for this sub-problem, plus up to 4 possible extra credit points]* By taking the first partial derivative of $\ell\ell(\mu\,\sigma\,\nu \,|\, \boldsymbol{y}\,\mathcal{T}\,\mathcal{B})$ with respect to one of the components of $\boldsymbol{\theta}$, briefly explain why our usual differentiation approach doesn't lead to a closed-form solution for the MLE in this problem *[4 points]. (Extra credit [4 points]: Write out explicit expressions for the other two first partial derivatives.)*

(c) *[32 total points for this sub-problem, plus up to 10 possible extra credit points]* In part (b) You've demonstrated that we need a new way to find the MLE, and You can guess what the new method has to be: when analytic solutions aren't available, we need to go numerical, which in this case means numerically maximizing the log-likelihood function $\ell\ell(\boldsymbol{\theta} \,|\, \boldsymbol{y}\,\mathcal{T}\,\mathcal{B})$. Fortunately, `R` has a good general-purpose numerical optimization function called `optim`, which works especially well when the function to be optimized has a single mode (You'll see below that this is true with this model and data set). `optim` has built-in code for six different iterative algorithms; You need to give the function decent starting values $\boldsymbol{\theta}_0$, and it then iterates from these initial values toward the nearest local maximum to $\boldsymbol{\theta}_0$ (which in our problem turns out to be the global max). The file `ams-206-optim-R.txt` on the course web page contains some `R` code that gets You started with `optim` in this problem. (Feel free to work with another numerical optimizer in a different computing environment, but study the `R` code and output to ensure that You produce similar results with Your software.) For initial values I've used $(\mu_0, \sigma_0, \nu_0) \doteq (404.6, 5.3, 6.0)$, based on the following reasoning:

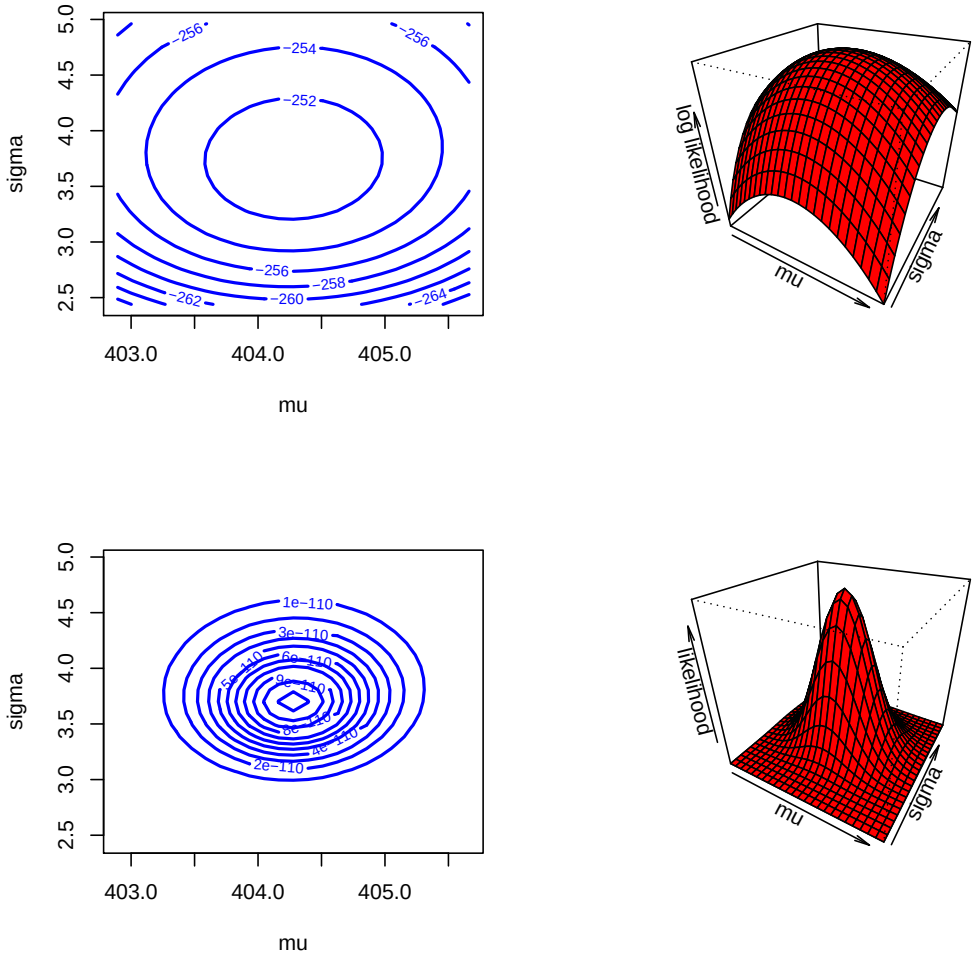
- The MLE for μ in the scaled t model should be reasonably close to the sample mean $\bar{y}$;
- From experience with the scaled t sampling distribution, I chose a value for ν that's rather small, corresponding to reasonably heavy tails; and
- It can be shown (You don't have to show this) that if $Y \sim t_\nu(\mu, \sigma)$ and $\nu > 2$, $V(Y) = \sigma^2 \left(\frac{\nu}{\nu-2}\right)$, from which I can get a decent starting value by calculating $\sigma_0 = s\sqrt{\frac{\nu_0-2}{\nu_0}}$, where s is the sample SD of the y_i values.

I've written the call to `optim` in such a way that the maximizer computes numerical approximations of both the MLE vector and {the Hessian (matrix of second partial derivatives) evaluated at the MLE}.

(i) Run the `R` code I've given You down through the comment line that says '`end of numerical analysis`', and study the output. Use these results to compute approximate 95% confidence intervals for μ, σ and ν *[6 points]*; in class we'll compare these likelihood intervals with their Bayesian counterparts.

(ii) Visualizing the log-likelihood and likelihood functions in this problem is at least slightly non-trivial, because we have to solve a 4–dimensional plotting problem (3 dimensions of $\boldsymbol{\theta}$ plus the (log) likelihood dimension). It's not particularly easy to make a 4D plot in `R`, but it *is* relatively easy to make 3D contour and perspective plots of the (log) likelihood function(s) plotted against two of the three components of $\boldsymbol{\theta}$ with the third dimension set equal to its MLE, which gives You a good idea of what's going on. Run the rest of the `R` code I've given You; this will produce a (2×2) matrix of contour and perspective plots of the log-likelihood and likelihood functions in this case study, for μ and σ while holding ν constant at its MLE (Your plot should look like Figure 1 below). Do these plots indicate that `optim` has indeed found the global maximum of the (log) likelihood function(s)? Explain briefly *[4 points]*. What do the (μ, σ) contour plots reveal about the correlation between these two parameters, and how (if at all) does this relate to the correlation matrix of $\hat{\boldsymbol{\theta}}_{MLE}$ in the numerical output? Explain

Figure 1: *Top row: contour plot (left) and perspective plot (right) of the log-likelihood function for μ and σ (holding ν at its MLE) in the NB10 case study; bottom row: contour plot (left) and perspective plot (right) of the likelihood function.*



briefly *[4 points]*.

(iii) Locate the comment line in the `R` code that starts with the text '`beginning of code block`'. Modify the code that runs from there to the comment line starting with '`end of code block`' to create the analogue of Figure 1 for (μ, ν); repeat for (σ, ν); and include the resulting two plots in Your solutions *[4 points]*. Briefly discuss the two separate pairwise correlations for each of (μ, ν) and (σ, ν), relating Your graphical findings to the numerical correlation matrix of $\hat{\boldsymbol{\theta}}_{MLE}$ *[8 points]*. Do *these* plots continue to indicate that `optim` has indeed found the global maximum of the (log) likelihood function(s)? Explain briefly *[4 points]*.

(iv) *(Extra credit [10 points]: See if You can figure out how to use a* ***CRAN*** *function in* ***R*** *to make a 4D* ***heat plot****, in which the three components of* $\boldsymbol{\theta}$ *are displayed in a 3D perspective plot and the fourth dimension is the likelihood surface evaluated at the grid of* $\boldsymbol{\theta}$ *values, coded so that largest likelihood values are bright red and smaller values taper down to gray. Please email Your* ***R*** *code to me; I'd like to know how to do this.)*