Prof. David Draper
Department of Statistics
Baskin School of Engineering
University of California, Santa Cruz
Winter 2023

STAT 206 (Applied Bayesian Statistics)

Take-Home Test 3: Part 1 (Required: 70 Total Points)

Absolute due date:
Uploaded to `canvas.ucsc.edu` by 11.59pm on Fri 22 Mar 2024

Name: Devanathan Nallur Gandamani (2086936)

Here are the ground rules: this test is open-book and open-notes, and has three parts, all of which are optional (extra credit). Part 1 consists of 7 true/false questions, each worth 10 points, and is ***Required***.

Students who wish to gain full mastery of all of the material presented this quarter are strongly encouraged to participate in office hour sessions from now through Sun 24 Mar 2024.

Some advice on style as you write up your solutions: pretend that you're sitting next to the grader, having a conversation about problem (x) part (y). You say, "The answer is z," and the grader says, "Why?" You then give your explanation, as succinctly as possible to get your idea across. 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, the graders and 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.

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me or our TAs. 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, 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.

> ***Under UCSC policies, submission of your solutions constitutes acceptance of the ethical rules stated above.***

In what follows I've not left blank spaces for your solutions. Those of You who are using `LaTeX` or some other word-processing environment to prepare Your solutions can stick quote blocks below each question, into which You can type Your answers (I suggest that You use bold or italic font to distinguish Your solutions from the questions). If You're submitting Your answers in longhand, which is perfectly acceptable, You can just write them out on separate sheets of paper, making sure that the grader can easily figure out which chunk of text is the solution to which part of which problem.

Part 1: True/False

[70 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; 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.

In answering these questions you may find it helpful to consult Gelman et al. (2014) Chapter 11.

(A) If You can figure out how to do IID sampling from the posterior distribution of interest to You, this will often be more Monte-Carlo efficient than MCMC sampling from the same posterior, especially when the number k of unknown quantities is small.
Solution: True
This is True, Independent and Identically Distributed (IID) sampling typically requires less computational effort compared to Markov Chain Monte Carlo (MCMC) sampling. However, this efficiency advantage is contingent upon the dimensionality of the dataset. As the dimensionality escalates, the computational demands of IID sampling also rise, potentially diminishing its speed advantage over MCMC sampling. Therefore, for the initial assertion to hold, it is critical that the dataset's dimensionality remains within a manageable range, ensuring that IID sampling retains its computational efficiency relative to MCMC sampling.

(B) A (first-order) Markov chain is a particularly simple stochastic process: to simulate where the chain goes next, You only need to know (i) where it is now and (ii) where it was one iteration ago.
Solution: False
This statement is incorrect. The fundamental principle of a first-order Markov chain is that its future state is determined exclusively by its current state, without the need to consider its past states. Thus, knowledge of the chain's immediate preceding state is sufficient to predict its next state, negating the requirement to trace its history beyond one iteration ago.

(C) The bootstrap is a frequentist simulation-based computational method, which is also a Bayesian nonparametric procedure, that can be used to create approximate confidence intervals for population summaries even when the population distribution/sampling model $[SM]$ of the outcome variable y of interest is not known; for example, if (by exchangeability, implied by the problem

context) all You know is that Your observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ are IID from *some* $[SM]$ with finite mean μ and finite SD σ, You can use the bootstrap to build an approximate confidence interval for μ even though You don't know what the population $[SM]$ is.

Solution: True
This assertion is true, given that the bootstrap method is predicated on resampling from the observed data. This approach enables the derivation of inferential statistics, such as confidence intervals, from the resampled data. Critically, this is achieved without necessitating prior knowledge of the underlying population distribution.

(D) In MCMC sampling from a posterior distribution, You have to be really careful to use a monitoring period of just the right length, because if the monitoring goes on for too long the Markov chain may drift out of equilibrium.

Solution: False
The statement is incorrect. When a Markov chain attains equilibrium, its state distribution will oscillate around this equilibrium distribution, and the primary expenditure of allowing the chain to continue for an extended period is merely electrical energy. The inherent behavior of a Markov chain is to stabilize at its equilibrium distribution once reached, endeavoring to maintain this state thereafter.

(E) Simulation-based computational methods are needed in Bayesian data science (inference, prediction and decision-making) because conjugate priors don't always exist and high-dimensional probability distributions are difficult to summarize algebraically.

Solution: True
This statement is accurate; managing probability distributions in high-dimensional spaces is notably challenging. The computational resources and algorithms introduced in this course are instrumental in facilitating the practice of Bayesian data science, rendering it feasible.

(F) In MCMC sampling from a posterior distribution, You have to be really careful to use a burn-in period of just the right length, because if the burn-in goes on for too long the Markov chain will have missed its chance to find the equilibrium distribution.

Solution: False
The assertion is incorrect. Once a Markov chain has reached its equilibrium, it tends to persist in this state of equilibrium. Therefore, the primary concern should be ensuring that the initial 'burn-in' period is of adequate length to allow the chain to stabilize at this equilibrium, rather than worrying about the chain deviating significantly from it thereafter.

(G) You're MCMC sampling from a posterior distribution for a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, in which $k \geq 1$ is a finite integer. During the monitoring period, the column in the MCMC data set for a component of θ (θ_j, say) behaves like an autoregressive time series of order 1 ($AR_1(\rho_1)$) with estimated first-order autocorrelation $\hat{\rho}_j = 0.992$. As usual, You'll use the sample mean $\bar{\theta}_j^*$ of the monitored draws θ_{ij}^* as Your Monte Carlo estimate of the posterior mean of θ_j. To achieve the same estimated Monte Carlo standard error for $\bar{\theta}_j^*$ that You would have been able to attain if You could have done IID sampling, Your MCMC monitoring sample size would have to be about 250 times bigger than the length of the IID monitoring run.

Solution: True
This is true, the Markov chain has an equilibrium distribution $p(\theta \mid D(SM)(PM)\beta)$
IID MC draws:

$$\text{MCSE}(\theta_j) = \frac{s_j}{\sqrt{M_{\text{IID}}}}$$

MCMC draws:

$$\text{MCSE}(\theta_j^*) = \frac{s_j}{\sqrt{M_{\text{mcmc}}}} \cdot \sqrt{\frac{1+\hat{\rho}_j}{1-\hat{\rho}_j}}$$

Thus for equal accuracy,

$$\frac{s_j}{\sqrt{M_{\text{IID}}}} = \frac{s_j}{\sqrt{M_{\text{mcmc}}}} \cdot \sqrt{\frac{1+\hat{\rho}_j}{1-\hat{\rho}_j}}$$

This simplifies to

$$\frac{M_{\text{mcmc}}}{M_{\text{IID}}} = \frac{1+\hat{\rho}_j}{1-\hat{\rho}_j}$$

which is our variance inflation factor.
Solving for this,

$$\frac{1.992}{1-0.992} = 249$$

which is indeed about equal to 250 making the statement true.