Modern Astrostatistics
Exam examples for Part A and B

Student name:

Student university-ID number:

Notation:
Vectors and matrices must be visually distinguishable. Transposed vectors and matrices are to be marked as such, also in scalar products and outer products.

..... Part A: Understanding. Totalling: 15 CP. .....

Since last Monday, I had 27 questions concerning the exam, of which 26 were about MCMC. I understand that MCMC is steadily mentioned at the institute, thereby creating the illusion of singular importance. This seems to create the false thinking that the exam will focus on MCMC. **This is not correct.** If you spend more than 10% of your exam preparation time on MCMC, then you are making a mistake. Here I provide further examples of Part A and B, which will be crucial parts to collect enough points for passing. Due to inventing this short-notice, the examples do not focus on astronomy and the points may not add up. It is however correct that the exam allots 15 points to Part A 'Understanding', 20 to Part B 'Analytics', and 15 to Part C 'Numerics'.
I confirm that the exam is in black-and-white in order to create equal chances for our colourblind participants. I confirm that you are under all circumstances allowed to use np.random.uniform. The packages emcee and corner are *not* allowed. I also confirm that the exam rules were adapted to allow use of private laptops, such that you can use your native keyboards. Kindly be reminded that you must then succeed in copying your files onto the university computers, and submit them for marking via the linux command cp. *The university spam filter will reject codes sent by email.*

I) Tigers in India ..... **6.5 CP**
It is uncertain how many tigers are still living in India. It is estimated that between 3000 to 4000 tigers are still populating India's nature. All numbers between these two limits are equally likely. Tigers come in two variations: white tigers and yellow tigers. Only 3.5 percent of all tigers are white. Let 'T' abbreviate 'tiger', 'W' shall abbreviate 'white' and 'Y' shall abbreviate 'yellow'.

1. Plot the distribution $\mathcal{P}(T)$ and give its name (0.5 CP)
2. What does $\mathcal{P}(W|T)$ mean (in words), and can you give its numerical value? (0.5+0.5 CP)
   P.S: The answer '$\mathcal{P}(W|T)$ is pronounced as probability of W given T' is insufficient. Explain it such that a top-ranking biologist or your best friend from medicine knows that you are talking about.
3. What does $\mathcal{P}(T|Y)$ mean (in words), and can you give its numerical value? (0.5+0.5 CP)

4. Explain what the integral $\int T\mathcal{P}(T)\mathrm{d}T$ means, and give its numerical value, if possible. (1+1 CP)
5. Explain what the integral $\int T\mathcal{P}(T|W)\mathrm{d}T$ means, and give its numerical value, if possible. (1+1 CP)

II) Elements of importance ..... 5 CP
Define (with equations) the following statistical concepts and explain them (in words).

1. Maximum-likelihood estimator. (1 CP)
2. Covariance matrix. (1 CP)
3. Statistical independence of two random variables. (1 CP)
4. conditional distribution. (1 CP)
5. marginal distribution. (1 CP)

III) Definitions and explanations ..... 5 CP
Define (with equations or through examples in equation form), *and* explain in words the following terms

1. A measure (as in 'measure space'). (1 CP)
2. Bayesian Evidence. (1 CP)
3. Bayes' Theorem. (1 CP)
4. Matched filter. (1 CP)
5. Rejection sampling. (1 CP)

IV) Correct interpretation of statistical tests ..... 5 CP

1. Describe the χ^2/degF test: what it is used for and what does it compute? (1 CP)
2. Name one caveat against it, where the problem is caused by the value of χ^2.
   Demonstrate your named caveat by giving/inventing a concrete example. (1+1 CP)
3. Name one caveat against it, where the problem is caused by the degrees of freedom degF.
   Demonstrate your named caveat by giving/inventing a concrete example. (1+1 CP)

V) Understanding ..... 4 CP

1. Explain why Bayesian parameter inference works even if the posterior is not normalized. (1 CP)
2. Explain the 'Curse of Dimension', which makes sampling and inference in high-dimensional spaces difficult. (1 CP)
3. Explain which interpretations Bayesians and frequentists attribute to the term 'probability'. I.e.: define 'probability' from the Bayesian perspective (1 CP) and from the frequentist perspective (1 CP).

VI) Understanding ..... x CP
Filters can be constructed to quickly sift through data and search for hidden signals. They can also be optimized, meaning they are the best solution under certain preconditions. Demonstrate your understanding of the *matched filter*, by answering the following questions.

1. For a signal vector $\vec{s}$ and data vector $\vec{x}$, give the equation for the matched filter. Define yourself symbols for potentially missing quantities, and name them.
2. The matched filter is an optimzed filter, but what exactly does it optimize? State which quantity it maximizes.
3. Explain *how* the matched filter was optimized: which is the crucial step in its derivation that leads to the matched filter being optimal? If you do not recall the decisive equation, state its name.

Part B: Analytics. Totalling: 20 CP.

The following are indicative elements of recurrent calculations which you will need to be able to conduct quickly during the exam.

I) Beta distribution ..... x CP

The beta-distribution of a random variable x is given by

$$x \sim x^{a-1}(1-x)^{b-1}. \tag{1}$$

1. What is the most likely data point to be drawn in a single measurement?
2. How many peaks does the distribution have?
3. Sketch $\mathcal{P}(x|a,b)$ for different integer values of a and b.
4. Adopting a flat prior, and assuming $b = 5$ is known, derive the posterior of unknown parameter a given a measured x.
5. Assume your colleague knows a bit more about a from previous measurements, and he gives you the extra information $\pi(a) \propto a^{-1}$. Derive the new posterior of a given a measured value of x and your colleagues' prior information.
6. Sketch the posterior. Label your axes carefully. Is the most likely value of a the same as the average value for a?

Hint: the solutions to the plots can be seen when querying Wikipedia for 'beta distribution'.

II) Logistic distribution ..... x CP

The logistic distribution is given by

$$x \sim \frac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2}. \tag{2}$$

1. What is the random variable here?
2. How many free parameters does this distribution have?
3. Convert equation (2) into the writing style $\mathcal{P}(x|...)$ or $\mathcal{P}(...|x)$. (Which of the two do you have to pick? Does x need to be before or after the conditional sign? Replace the dots by the missing parameters.)
4. Is this a symmetric distribution? Hint: What happens if you replace x by $-x$?
5. Do mean and peak of the distribution coincide?

6. If you change s, what happens to the shape of the distribution?
7. If you change μ, what happens to the distribution?
8. From 6 and 7, create a plot which indicates likelihoods of s and μ which are *definitely* incompatible with drawn data x. Indicate a drawn data point x in your plot.
9. If x drawn from equation (2), what is the distribution of $y = kx + l$, for constants k, l?
10. Imagine I gave you $x = \pi$, and I asked you for the full posterior of μ, because μ has some physical meaning that we are interested in. Sadly, we do not know anything about s. Which computation would you need to set up?
    - $\mathcal{P}(s, \mu|x = \pi)$?
    - $\int \mathcal{P}(\mu|x = \pi, s)\mathrm{d}s$?
    - $\int \mathcal{P}(\mu, s|x = \pi)\mathrm{d}s$?
    - $\int \mathcal{P}(\mu, s|x)\mathrm{d}x$?
    - $\int \mathcal{P}(\mu, s, x)\mathrm{d}x$?

Hint: The solutions to most of these questions can be deduced from the Wikipedia page of the 'logistic distribution'.