Cosmology NAME:

Homework 17 All: Bayesian Analysis

1. The Bayesian analysis iteration formula for iteration ℓ is

$$P(T_i|K_\ell) = \frac{P(D_\ell|T_iK_{\ell-1})P(T_i|K_{\ell-1})}{\sum_j P(D_\ell|T_jK_{\ell-1})P(T_j|K_{\ell-1})} ,$$

where $\{T_i\}$ is an exhaustive set of possible theories about some aspect of reality, K_ℓ is background knowledge after iteration ℓ, D_ℓ is data acquired in iteration ℓ, $P(T_i|K_\ell)$ is the posterior probabiltity of theory T_i to your knowledge for iteration ℓ, $P(D_\ell|T_iK_{\ell-1})$ is the probability of D_ℓ given theory T_i and background knowledge $K_{\ell-1}$, and $P(T_i|K_{\ell-1})$ is the prior probabiltity of theory T_i to your knowledge for iteration ℓ. That the iteration formula exists in principle is vital since it proves that the ideal Bayesian analysis leads to true theories. That the ideal Bayesian analysis can be approached in practice is also vital since that means it is a useful path to true theories. In toy cases, one can actually do ideal Bayesian analysis. But in toy cases, you know the true theory is included in the set of the set of possible theories which is exhaustive by definition.

However, in practice, you usually only do iteraion 1 formally. Initial background knowledge K_0 implicitly contains vague Bayesian analysis iterations going back to vaguely negative infinity. Also, you usually do not have and are not interested in having an exhaustive set of theories $\{T_i\}$. You usually just have interest in a set of interesting theories $\{T_i\}$: i.e., a set of theories that seem likely a priori. You usually just assign the theories equal priors following the principle of indifference, unless you has some other guidance. Evaluating the denominator of the iteration formula is useless in this practical Bayesian analysis, and so is seldom done explicitly. What you do do is evaluate the Bayesian odds ratio for any two of theories to compare them. The Bayesian odds ratio for theories T_i and T_j is

$$\frac{P(T_i|K_\ell)}{P(T_j|K_\ell)} = \frac{P(D_\ell|T_iK_{\ell-1})}{P(D_\ell|T_jK_{\ell-1})}\frac{P(T_i|K_{\ell-1})}{P(T_j|K_{\ell-1})} = k_{\rm B}\frac{P(T_i|K_{\ell-1})}{P(T_j|K_{\ell-1})} ,$$

where

$$k_{\rm B} = \frac{P(D_\ell|T_iK_{\ell-1})}{P(D_\ell|T_jK_{\ell-1})}$$

is the Bayesian k factor or Bayesian evidence. If you have made used the principle of indifference, all you have is the Bayesian evidence to compare the theories by. But most theories have free parameters. How are they accounted for? You expand $P(D_\ell|T_iK_{\ell-1})$ in the terms of the free parameters: i.e.,

$$P(D_\ell|T_iK_{\ell-1}) = \int P(D_\ell|T_i(\theta)K_{\ell-1})\rho(\theta)\,d\theta ,$$

where θ stands symbollically for all free parameters, $\rho(\theta)$ is the probability density for all free parameters, the integration is over all free parameter space, and $P(D_\ell|T_i(\theta)K_{\ell-1})$ is, in fact, the likelihood or likelihood function. The hard part of Bayesian analysis is usually choosing $\rho(\theta)$ which is really the hard prior to evaluate. Usually, you just assign a flat prior $\rho(\theta)$: i.e.,

$$\rho(\theta) = \begin{cases} \dfrac{1}{\Delta\theta_{\text{range}}} & \text{for } \theta \text{ in the range } \Delta\theta_{\text{range}}; \\ 0 & \text{for } \theta \text{ not in the range } \Delta\theta_{\text{range}}. \end{cases}$$

The hard part is thus reduced to determining $\Delta\theta_{\text{range}}$. Independent Bayesian analyses can find very different Bayesian evidences depending on how the researchers choose $\Delta\theta_{\text{range}}$. This is why Bayesian evidence is usually not considered decisive if k_{B} is of order a few or even of order 10. If k_{B} is order 100 or 1000, then that may be decisive depending on who is judging.

Note maximizing the likelihood gives you the best set of free parameters assuming a theory is true. Using a theory with maximum likelihood parameters biases in favor of the theory in Bayesian analysis since the theory is not assumed to be true or usually even more likely than other interesting theories. In fact, eliminating the free parameters by the integration above implements Occam's razor: *"Numquam ponenda est pluralitas sine necessitate"* ("Plurality must never be posited without necessity"). You eliminate unnecessary and misleading hypotheses about the free parameters. This elimination process is called:

a) Occamization. b) dithering. c) marginalization. d) buffering.
e) obscuration.

2. You are in Las Vegas, right? So you know dice (singular die). Let's see if we can predict the odds for a throw of two dice.

   a) Let's start being general, but not too general. You have two identical dice. They each have I faces with dot count running $i = 1, 2, \ldots, I$. The probability of any face (i.e., any face landing facing up) is P_i. What is the probability for a dice throw yielding faces i and i? What is the probability for a throw yielding first face i and then face j where $i \neq j$. What is the probability for a dice throw yielding faces i and j where $i \neq j$ and you do not distinguish the order or, in other words, you sum over the probabilities for the different orders.

   b) Let the sum of the face dots yielded by a throw be $k = i + j$. What is the run of possible k values (i.e., the ordered sequence of possible k values) and how many values are there? Is there always a middle value? Why? What is the middle value and how many values are above and below it?

   c) Now what we really want to know is what is the probability P_k of the summation of face dots being $k = i + j$: i.e., the probability distribution for a throw of two dice which is our random variable. Determine the two summation formulae needed and the number of terms in each summation. **Hint** The two formulae can be adjusted to look the same, except for their limits. The real hard part is determining limits. Draw an outcome square for the throw results with row index i and column index j. The squares to include in the summation are on the diagonals with $i + j = k$ with k constant.

   d) Specialize the P_k formulae to the case of equal face probability: i.e., all $P_i = 1/I$. Conflate the two formulae into one with transformation $k = k' + (I+1)$, and show that $P_{k'}$ is an even function of k', find the limits on k', and find the maximum P'_k value.

e) Specialize the P_k and $P_{k'}$ formulae to the case of ordinary dice with $I = 6$ and $P_i = 1/6$. Tabulate the probability distributions P_k and $P_{k'}$ for the random variables k and k'.

3. The multinomial theorem (from which the multinomial probability distribution is derived) is generated by the generating function (using that expression loosely)

$$F_N = F_1^N = \left(\sum_{i=1}^{I} P_i\right)^N = \sum_{i,j,\ldots} P_i P_j \ldots \ ,$$

where N is the number of factors in a sequence of factors, F_1 is the multinomial theorem for sequences of length 1, I is the number of variables and the order of the multinomial theorem (e.g., $I = 2$ for the binomial theorem), P_i is factor i (which for the multinomial probability distribution becomes probability of event i), the sequence of factors $P_i P_j \ldots$ are the terms in the multinomial expansion resulting from a straightforward branching multiplication before collecting terms into multinomial terms, $\sum_{i,j,\ldots}$ is the sum over the sequences (i.e., uncollected terms), and the total number of sequences is I^N. Note that all possible sequences of factors P_i must occur uniquely in the $\sum_{i,j,\ldots}$ since the branching pattern of all possibilities is exhaustive and there can be no duplications since obviously the first factor in each sequence is different.

a) There are, as aforesaid, I^N sequences of factors. But what is the count of sequences for each combination: i.e., for each set of sequences have the same sets of factors P_i without distinguishing order. Such a count of sequences is called a multinomial coefficient. Note that sequences differing by undistinguishable factors are the same sequence in the branching multiplication that creates the whole set of sequences.

Let the multinomial coefficient for each combination be $C(N, \{n_i\})$, where $\{n_i\}$ stands for the set of factors P_i in the sequences. To be explicit, every distinct combination has a unique set $\{n_i\}$ otherwise it would not be a distinct combination. Note $\sum_{i=1}^{I} = N$, of course. Derive the formula for $C(N, \{n_i\})$ in terms of N and $\{n_i\}$. **Hint:** You will need factorials. Also, note the odd fact that you have consider permutations of the same factor P_i in a sequence even though these permutations just give the same sequence as it would occur in actually creating the sequences by the branching multiplication.

b) The individual distinct sequences are usually not of interest. What one usually wants is collect all the sequences corresponding to each unique combination since they all have the same numerical value, and so in the probability distribution all have the same probability. The collections are the multinomial terms for the multinomial theorem. Using the result of part (a), derive the formula for a multinomial term

$$\tilde{P}(N, \{n_i\}) \ ,$$

and the formula for multinomial theorem itself in terms of multinomial terms. Just use $\sum_{\{n_i\}}$ for the summation of the multinomial terms since there is no simple way in general to explicitly order them in a summation.

c) If the factors P_i are identified as probabilities of events i, then we require

$$\sum_{i=1}^{I} P_i = 1 \ .$$

What is value of F_N in this case and what does this value mean? What is the probability of obtaining the combination of events $\{n_i\}$?

d) The multinomial term $\tilde{P}(N, \{n_i\})$ is the multinomial probability distribution itself. We can easily obtain some ancillary formulae about the multinomial probability distribution. For example, the mean number of events j for the multinomial probability distribution is

$$\mu_j = \langle n_j \rangle = \sum_{\{n_i\}} n_j \tilde{P}(N, \{n_i, P_i\}) \ ,$$

where j is just a representative index. Derive the explicit formula for μ_j for the multinomial probability distribution. **Hint:** The trick is treat the P_i as variables in the multinomial theorem in both the forms

$$F_N = \sum_{\{n_i\}} \tilde{P}(N, \{n_i, P_i\})$$

and

$$F_N = F_1^N = \left(\sum_{i=1}^{I} P_i \right)^N \ .$$

You then apply operator $P_j(\partial/\partial P_j)$ to both of forms and afterward impose the constraint that the constraint $F_1 = \sum_i P_i = 1$.

e) The variance/covariance of a multinomial probability distribution is given by

$$\sigma_{jk}^2 = \langle (n_j - \mu_j)(n_k - \mu_k) \rangle = \langle n_j n_k \rangle - \mu_j \mu_k \ .$$

Derive the explicit formula for σ_{jk}^2 for the multinomial probability distribution. Explain the striking feature of covariance case (i.e., the case when $j \neq k$). **Hint:** The trick is used in part (d) still works *mutatis mutandis*.

e) Specialize the results of parts (a), (b), and (c) of the binomial theorem: i.e., the case where $I = 2$. For best understanding, let $n_1 = k$ and $n_2 = N - k$, where $k \in [0, N]$ is a usual parameter for specify all the sets $\{n_i\}$.

4. The Poisson (probability) distribution is

$$P = \frac{\mu^x}{x!} e^{-\mu} \ ,$$

where μ is the mean of integer random variable x, $\sigma = \sqrt{\mu}$, and there is no upper limit on x.

The Poisson distribution is appropriate for analyzing two kinds counting observations which not completely distinct. The first kind of counting observation is where the events occur randomly in time (or some similar variable), but there is a mean number of events per unit time μ and the time of each event is zero or approximately that. In this case, the Poisson distribution is exact if the time of an event is zero. An obvious example of this kind of observation is counting the radioactive decays from a long-lived radioactive sample.

The second kind of counting observation is where the random variable x (the count of events) obeys an extreme binomial distribution where p the probability of x on an individual trial is very small (i.e., $p << 1$) and consequently $\mu << n$, where n is the number of trials. If you actually do know n and p, you could just use the binomial distribution itself, but the Poisson distribution may be an adequate approximation. Note for the binomial distribution $\mu = np$ and $\sigma = \sqrt{np(1-p)} \approx \sqrt{np} = sqrt\mu$.

In both cases, you may often just have one count x, and not know μ nor σ. However, you can estimate $\mu \approx x$, and thus $\sigma \approx \sqrt{x}$ and this is often done.

a) Derive a cute formal general formula for the moments

$$\langle x^\ell \rangle = e^{-\mu} \sum_{x=0}^{\infty} x^\ell \frac{\mu^x}{x!}$$

of the Poisson distribution, where ℓ runs $0, 1, 2, \ldots$. Use the formula to solve moments for $\ell = 0, 1, 2$ and for the formulae for μ and σ. **Hint:** Operating with the operator $[\mu(\partial/\partial\mu)]^\ell$ is the trick.

b) The derivation of the Poisson distribution for the first kind of counting observation mention in the preamble is straightforward. Say τ is the average rate of random events. The probability of observing no events in time t (starting from time $t = 0$) obeys the differential equation

$$dP(x=0,t) = -P(x=0,t)\frac{dt}{\tau} \ ,$$

where $P(x = 0, t)$ is the probability of having no events to t and dt/τ is the differential probability of an event in dt. The solution for $P(x = 0, t)$ is clearly

$$P(x=0,t) = e^{-t/\tau} \ .$$

The differential formula for x events in t is

$$dP(x,t) = e^{-t/\tau} \prod_{i=1}^{x} \frac{dt_i}{\tau} \ ,$$

where we assume the events are instantaneous. Simple integration of all dt_i gives the Poisson distribution plus accounting for overcounting with events pass each other on the time line. Complete the proof of the Poisson distribution. Give the explicit μ and σ formulae for this case.

c) Prove the Poisson distribution by taking the limit of the binomial distribution

$$P(x,n,p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

where $n \to \infty$, $p \to 0$, $np \to \mu$ (which is a finite nonzero value), and x is fixed. **Hint:** You will need to expand $(1-p)^n = (1-\mu/n)^n$ in a binomial theorem expression.

5. Bayes' theorem in symmetric form is

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) ,$$

where P is probability, A and B are events, $P(A)$ is the probability of A, $P(B)$ is the probability of B, $P(AB)$ is the probability of A and B, $P(A|B)$ is the conditional probability of A given B, and $P(B|A)$ is the conditional probability of B given A. In unsymmetric form,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{or equivalently} \qquad P(B|A) = \frac{P(A|B)P(B)}{P(A)} .$$

Note that in the notation we are using, AB is not the product of A and B, but the union of A and B: i.e., AB is A and B.

a) Prove the expansion rule

$$P(AB) = P(A|B)P(B)$$

and Bayes' theorem from frequentist definition of probability. Frequentist definition states given population of events N, the probability of sampling events with property A is $P(A) = N_A/N$ where N_A is the number of events in the population with property A.

Yours truly believes that probability only has meaning from the frequentist definition. You can do a lot of probability formalim without the definition, but it seems to have no meaning without the definition. Maybe yours truly is just ignorant. However, the limitation to the frequentist definition isn't really a limitation in yours truly view since frequentist definition always applies even if you can't can't calculate the probabilities with high accuracy from it. Thus, Bayesian analysis can be applied generally.

b) Yours truly is not going to give a general description of Bayesian analysis procedure here, but just a description of an ideal procedure that concerns itself with the theories in order to find the true one. Say we have a system, the exhaustive finite set of all theories of nonzero probability $\{T_i\}$ that apply to the system $\{T_i\}$ and inital knowledge K_0 about the system (which includes the set of theories, of course). Given that the set of theories is exhaustive, their probabilities to our knowledge (i.e., K_0) is normalizable: i.e., we have $\sum_i P(T_i|K_0) = 1$.

Now how is it possible to assign a probability to a theory T_i? Well if we know the theory is true, $P(T_i|K_0) = 1$ and if we know it is false, $P(T_i|K_0) = 0$. What if you don't know whether T_i is true or false? Well there are procedures of assigning numerical probabilities to theories based background knowledge. After all people are always assessing theories as probable, very probable, improbable, or very improbable based on their background knowledge. This assessment must be based on some fuzzy frequentist analysis of the features that make up a theory. Now the procedure of assigning (numerical) probabilities doesn't have to be perfect—and probably rarely is in practice—but the better it is, the faster in all probability the

Bayesian analysis will converge to the true theory. One procedure is the principle of indifference: just assign equal probabilies to the theories. By the principle of indifference, if there are I theories, $P(T_i|K_0) = 1/I$ for all i.

In fact, the completely fuzzy assignments of probability only happens prior to the first iteration of Bayesian analysis when our background knowledge is K_0. The zeroth probabilities $P(T_i|K_0)$ are our zeroth prior probabilities (AKA zeroth priors). After completing Bayesian analysis iteration $(\ell - 1)$ we have posterior probabilities (AKA posteriors) $P(T_i|K_{\ell-1})$ relative to the $(\ell-1)$th iteration; they are the priors for the iteration ℓ.

In iteration ℓ, we acquire new data D_ℓ which gives us updated knowledge $K_\ell = D_\ell K_{\ell-1}$, where $D_\ell K_{\ell-1}$ recall is a union, not a product. To get the posteriors for the ℓth iteration, we apply Bayes' theorem:

$$\begin{aligned} P(T_i|K_\ell) = P(T_i|D_\ell K_{\ell-1}) &= \frac{P(D_\ell|T_i K_{\ell-1})P(T_i K_{\ell-1})}{P(D_\ell K_{\ell-1})} \\ &= \frac{P(D_\ell|T_i K_{\ell-1})P(T_i|K_{\ell-1})P(K_{\ell-1})}{P(D_\ell|K_{\ell-1})P(K_{\ell-1})} = \frac{P(D_\ell|T_i K_{\ell-1})P(T_i|K_{\ell-1})}{P(D_\ell|K_{\ell-1})} \ . \end{aligned}$$

Note that $P(K_{\ell-1})$ has canceled out, and so the result is valid no matter what the value of $P(K_{\ell-1})$ though if we are doing the Bayesian analysis correctly it should be 1.

Now if we actually have data D_ℓ, then $P(D_\ell) = 1$. But $P(D_\ell)$ is not what is in the denominator of the result. We have $P(D_\ell|K_{\ell-1})$ which the probability of getting data D_ℓ given that we know $K_{\ell-1}$ which recall includes the knowledge that the set $\{T_i\}$ exists. We can, in fact, expand $P(D_\ell|K_{\ell-1})$ in the set $\{T_i\}$:

$$P(D_\ell|K_{\ell-1}) = \sum_j P(D_\ell|T_j K_{\ell-1})P(T_j|K_{\ell-1}) \ ,$$

where the summation is over all the set $\{T_i\}$. Now we have the Bayesian analysis iteration formula

$$P(T_i|K_\ell) = \frac{P(D_\ell|T_i K_{\ell-1})P(T_i|K_{\ell-1})}{\sum_j P(D_\ell|T_j K_{\ell-1})P(T_j|K_{\ell-1})} \ .$$

We note that $P(D_\ell|K_{\ell-1})$ is the weighted mean of the $P(D_\ell|T_j K_{\ell-1})$'s where the $P(T_j|K_{\ell-1})$'s are the weights.

The last equation is in fact the probability update formula. Those theories T_i whose ℓth posteriors are greater/lesser/equal relative to their $(\ell - 1)$th priors gain/lose/conserve credence.

We now assume that there is enough potential knowledge K_L for a decisive determination: i.e.,

$$P(T_i|K_L) = \begin{cases} 1 & \text{if } T_i \text{ is true;} \\ 0 & \text{if } T_i \text{ is false.} \end{cases}$$

This means that the Bayesian analysis converges to truth as $\ell \to L$. Note that convergence happens no matter how imperfect our method of assigning probabilities is provided we keep iterating until we reach K_L where only one viable theory remains. However, the amount of K_L actually varies depending on which data sets D_ℓ we acquire and how accurate are our probability assignments for $P(T_i|K_0)$ and $P(D_\ell|T_iK_{\ell-1})$. Obviously, if we make really good choices for the data sets D_ℓ and for probability assignments, convergence should be fast. If we make really poor choices, we may be iterate to a very large K_L and all the probabilities calculated in the iteration may be wildly in accurate except that we can calculate the $P(T_i|K_L)$'s accurately and end the iteration. In this extreme case, the Bayesian analysis wasn't very useful, except as a tactic to keep going. We just accumulated data until we had exhausted the possibilities and arrived at truth.

There's a relevant aphorism attributed to Ernest Rutherford (1871–1937): "If you need statistics, you are doing the wrong experiment." In fact, all aphorisms are true and false (including this one). Howsoever, the point of Rutherford's aphorism is that you choose data acquisitions as decivively as possible to speed the Bayesian analysis iteration (in a formal or informal sense) to completion.

The Bayesian analysis procedure described above is an ideal one which is probably very seldom fully carried out. Much less ideal procedures are usually used—and for darn good reasons. But it is important that the ideal procedure exists: a procedure which guarantees the arrival at truth. We could not trust Bayesian analysis if there were no ideal procedure to approach. If there were no ideal procedure to approach, Bayesian analysis might fail in some cases no matter how well we did it.

Does the foregoing seem OK to you? If not, why not?

c) From the Bayesian analysis iteration formula given in part (b) prove that the $P(T_i|K_\ell)$'s are normalized even if the the $P(T_i|K_{\ell-1})$'s are not. Why does this normalization inevitably happen?

d) What does it mean if all $P(T_i|K_\ell)$ are zero in Bayesian iteration step?

e) What does it mean if $P(T_i|K_\ell) = 1$, but your set of theories $\{T_i\}$ was not actually exhaustive.

f) What does it mean if $P(T_i|K_\ell) = 1$ and your set of theories $\{T_i\}$ was actually exhaustive.