03-512/02-512/03-712/02-712	Computational Methods for Biological Modeling and Simulation
Wean Hall 5310	Tuesdays and Thursdays 3:00pm-4:20pm
Lecturer: Russell Schwartz	

Problem Set 6

Assigned 11/13/14, Due 11/25/14

(NOTE: THIS IS DUE TWO DAYS EARLY BECAUSE OF THE THANKSGIVING HOLIDAY)

Problem 1 (15 points): Our goal in this problem will be to pose some formal parameter inference problems based on informal problem descriptions. In each case, you will be given a generic description of the problem and data available. You should identify a set of variables and provide an objective or likelihood function you could use to solve the problem. You should also identify a class of method that would be appropriate to solve the problem. You do not need to give details on the method, just a one-sentence statement of the general type of algorithm you would use. (Note that there may be many valid answers to the parts of this question.)

a. Suppose we are interested in learning transmission and recovery rates for a CTMM SIR disease model. We assume we have data providing populations of susceptible, infected, and recovered people over time and want to know rates with which infected people spread the infection to susceptible ones and with which susceptible people recover.

b. Suppose we are trying to optimize growth rate of a strain of bacterium. We believe growth rate depends quadratically on concentrations of two nutrients and can measure growth rate as a function of each nutrient. We measure growth rate at various nutrient concentrations. We would like to learn a model of the dependence so we can find its optimum.

c. Suppose we are interested in learning the frequencies of eye color alleles in a population. We assume there is a single gene responsible with two alleles, B and b, and that there is a dominant brown color and a recessive blue color such that a person with a BB or Bb genotype has brown eyes but bb has blue eyes. We would like to know how common the B and b alleles are based on observing eye colors of random set of people in the population.

Problem 2 (20 points): Suppose we are trying to treat cancer patients and we have a drug that we know will extend the patient's life but will eventually become ineffective due to development of drug resistance. We would like to be able to predict when resistance will develop so we know when we are likely to need to switch the patient to a different therapy. We believe that we can predict resistance based on the concentrations of a set of biomarkers (e.g., gene expression levels) $x_1, \ldots, x_m$ and would like to learn a model of how resistance depends on biomarker levels by observing time-to-resistance as a function of biomarker levels for a set of n patients. Our input data can then be thought of as a matrix M where m_{ij} is the level of biomarker i in patient j and a vector $\vec{r}$ where r_j is the time to resistance of patient j. In this problem, we will consider a few possible alternatives for solving that model-fitting problem.

a. We will first assume that we can solve this as a Gaussian linear regression problem, i.e., assuming that each observed resistance time is a Gaussian random variable whose mean is a linear function of biomarker levels and whose variance is some known value σ^2. Provide a mathematical formula for a likelihood function for this variant of the problem.

b. Suppose instead we pose the problem assuming that the resistance is described by a Gaussian quadratic model of biomarker levels, i.e., the same thing as part a except that the mean of the Gaussian is a quadratic function of biomarker levels. Provide a mathematical formula for a likelihood function for this variant of the problem.

c. We could solve a maximum likelihood version of either of the proceeding variants of the problem by solving a linear system to fit least-squares optimal coefficients to the appropriate regression model. What factors should we consider to decide whether it would be preferable to use the linear model from part a versus the quadratic model from part b?

d. We could alternatively pose the inference of either model as a sampling problem. Suppose we decide to consider the model of part a and want to create a Metropolis sampler over parameter values. We will assume that a potential move in the model corresponds to perturbing a single parameter by $\pm\Delta$ for some discrete step size Δ. Provide pseudocode for one step of a Metropolis sampler to sample from the likelihood of the data given the linear model.

e. What would be an advantage and a disadvantage of sampling over the likelihood versus solving for the maximum likelihood?

Problem 3 (20 points): Suppose we have created a structural model of a channel protein. We would like to simulate a molecule passing through the channel and see how the channel's structure responds to the molecule's presence. We need a high resolution model of the movement of the molecule over time and decide to represent it by identifying a set of discrete time/space points we want it to pass through and then interpolating between them. We will assume that the molecule's position is represented by a single dimension, corresponding to distance through the channel, and that we want it to pass through the following points:

t (ns)	0	2	5	8	10
x (nm)	0	5	6	10	20

For each of the following interpolation methods, derive a function or set of piecewise functions interpolating x over the range $t = [0, 10]$.

a. Piecewise linear

b. Quartic (fourth order) polynomial

c. Quadratic spline matching function values at each data point and first derivatives of consecutive interpolants at shared data points. You can further assume that we want to require $\frac{dx}{dt}(0) = 0$, i.e., that the derivative is zero at the beginning of the observed interval.

Problem 4 (25 points): In Lewis Caroll's poem "The Hunting of the Snark" a group of characters seek to hunt a creature called a snark. This is a risky pursuit, though, because a subset of snarks are a particularly dangerous kind of snark called a boojum and a hunter who pursues a boojum will disappear never to be seen again. We decide that it would be a good idea to figure out how common boojums are so we know how risky snark hunting will be. We identify a group of islands where snarks live, count the number of snarks on each, and send a hunter to hunt the snarks on each island. If the hunter does not return, then we know that island had a boojum. We would like to use this experiment to estimate the frequency of boojums in the population. Assume for the remainder of this problem that we have observed n islands with population counts $\vec{s} = (s_1, \ldots, s_n)$

of snarks. We have a set of boolean variables $\vec{b} = (b_1, \ldots, b_n)$ where b_i is 1 if island i contained a boojum and zero if island i did not contain a boojum. Our goal is to learn the frequency f of boojums, assuming each snark is independently a boojum with probability f.

a. Provide a formula for the likelihood of the observed data for a given boojum frequency f.

b. We will propose to solve for f by expectation maximization. To do that, we will propose that the latent variables will be a vector $\vec{y} = (y_1, \ldots, y_n)$ representing the number of boojums on each island. Provide a formula for the M-step of the model, estimating f given $\vec{s}$, $\vec{b}$, and $\vec{y}$.

c. Now provide a formula for the E-step of the model. You should have found that the formula for f is linear in $\vec{y}$ so it is sufficient to estimate expected values for each y_i given $\vec{s}$, $\vec{b}$, and f.

d. Write code implementing your EM algorithm. Your code should take as input an initial guess f_0, a number of iterations r, the number of islands n, and then a set of pairs of s_i b_i values as follows:

f_0
n
s_1 b_1
s_2 b_2
$\vdots$
s_n b_n

It should return as output the final inferred f.

e. Provide the f you infer for the following test case:
0.5
10
5
5 1
2 0
6 1
4 1
1 0

***Problem 5 (20 points):** In this problem, we will explore an approach we will not have time to cover in class for dealing with the problem of model overfitting. Suppose we consider a version of the biomarker issue from problem 2. We will suppose that we are trying to predict development of drug resistance in patients from a set of m biomarkers $x_1, \ldots, x_m$, but we will now assume m may be very large (e.g., expression levels of every gene in the genome). We observe a set of n patients and measure biomarkers and time to drug resistance r for each of them, producing a set of data points $(x_{11}, x_{21}, \ldots, x_{m1}, r_1), \ldots, (x_{1n}, x_{2n}, \ldots, x_{mn}, r_n)$. We would like to build a model of time to resistance as a linear function of the biomarkers. If $m >> n$, though, then the problem is underdetermined and we can find many predictors that will perfectly fit the training data but have no predictive value. To resolve this, we decide that we want to find a least-squares estimator of r really only want to use k of the biomarkers, for $k << m$. We will explore some ways to pose that problem.

a. Provide an expression for the objective function of minimizing least-squares error in terms of the

problem variables and the coefficients of the linear model you will infer. (Hint: It will help to define a set of binary auxiliary variables b_i where $b_i = 1$ if you use biomarker x_i and $b_i = 0$ otherwise.)

b. Now come up with a formal mathematical expression of the problem of finding the least-squares best fit model using only k of the biomarkers.

c. Describe how we could solve for the problem statement you derived for part b. You do not need a detailed algorithm, just a general description of the algorithmic techniques we could use to solve for the problem.

d. Rather than picking a fixed k, it is common to try to balance model complexity against quality of fit. So, for example, we would favor small k if adding more variables does not improve the objective funciton by much but would allow k to increase if it would lead to a big reduction in the sum-of-squares error. Repose the optimization problem on the assumption that we will allow any k but the solution quality will be the least-squares score plus a model complexity penalty λk for some parameter λ. Then revise your algorithmic strategy to solve for this problem variant.

e. A common trick to make this method work more efficiently is to drop k and the b_i's altogether and just penalize by λ times the sum of absolute values of the linear coefficients, which will tend to force most of them to zero and infer a simple model. Repose the problem again for this variant and again describe a general algorithmic strategy you could use to resolve for the resulting system. (Note that you will not be able to solve this variant as an unconstrained continuous optimization problem, since absolute values introduce discontinuities into the derivatives. You will want to introduce some new auxiliary variables to get around that.) Why might this variant of the problem be preferable to the variant in part d despite that complication?

*Recall that the starred problems are only for the 02-712/03-712 students.