problem set no. 1 — due Tuesday 9/16 at midnight.

question 1.1. (15 point) define the d-dimensional simplex $\mathbb{S}^d$ as the set of points $\vec{u} \in \mathbb{R}^d$ such that the sum $\sum_{i=1}^d u_i < 1$ and individual components $u_i > 0$ for $i = 1 \dots d$. consider $\vec{x} \in \mathbb{R}^d$ with distribution $N_d(\vec{\mu}, \Sigma)$, where $\Sigma_{ii} = \alpha \; \forall i$, $\Sigma_{ij} = -\beta \; \forall ij$, and $\alpha, \beta > 0$. consider the transformation $\vec{u} = g(\vec{x})$ such that

$$u_i = e^{x_i}/(1 + \textstyle\sum_{j=1}^d e^{x_j}) \qquad \text{and} \qquad x_i = \log(\tfrac{u_i}{u_{d+1}})$$

where $u_{d+1} = 1 - \sum_{j=1}^d u_j$. the points $\vec{u}$ live in $\mathbb{S}^d$. compute the density of $\vec{u}$.

note. i wish to see explicit steps in your calculations; this result is known and you can check the paper by Aitchison & Shen (Biometrika, 1980) for correctness—posted online.

general wisdom. maintaining the symmetry of a distribution is often helpful to get through calculations; if you must give symmetry up, you may still want to hold on to it for part of the process. as you are aiming for a p-dimensional distribution, starting from a q-dimensional distribution of basic quantities, $q > p$, is often helpful.

question 1.2. (20 points) derive maximum likelihood estimators for $\vec{\mu}, \alpha, \beta$ of $f_U(\vec{u})$ in *1.1.*. (hint. there is an analytic solution for this MLE, but it may be more efficient in terms of your time to push the analytic solution as far as you can (e.g. getting gradients and the Hessian matrix), then use numerical methods. however, keep in mind that your estimates will be better if you can obtain an analytic solution.)

question 1.3. (5 point) next we will write functions in `R` to work with $f_U(\vec{u})$. please read the section 'what to submit' carefully **before** writing any code. you will need write the functions detailed below:

1. `dlogisticnorm(u,mu,alpha,beta)`: computes the density of a point $\vec{u} \in \mathbb{H}^d$,
2. `logisticnorm.mle(U)`: a function to estimate the parameters $\vec{\mu}, \alpha, \beta$. this function should return a named list of the form:
   `list("mu.hat"=mu.hat,"alpha.hat"=alpha.hat,"beta.hat"=beta.hat)`
   where `*.hat` are the respective MLEs.

(hint. you can simplify your life substantially by building your functions around the corresponding functions for the multivariate normal in the existing `R` package `mvtnorm`.)

question 1.4. (10 points) use the sample of 250 points provided in the file `dataLogisticNorm3D.txt` to compute the estimates for the parameters $\vec{\mu}, \alpha, \beta$. your code should read the file in using the command `read.table("dataLogisticNorm3D.txt",header=TRUE)`. note that for

your code to successfully execute this requires that you execute your script in the folder containing `dataLogisticNorm3D.txt`.

example with derivations. define the d-dimensional hypercube $\mathbb{H}^d \subset \mathbb{R}^d$ as the set of points $\vec{u}$ with individual components $u_i \in [0,1]$, for $i = 1 \dots d$. consider $\vec{x} \in \mathbb{R}^d$ with distribution $N_d(\vec{\mu}, \Sigma)$, where $\Sigma_{ii} = \alpha \ \forall i$, $\Sigma_{ij} = -\beta \ \forall ij$, and $\alpha, \beta > 0$. in the spirit of the example above, consider a transformation $\vec{u} = g(\vec{x})$ such that $g : \mathbb{R}^d \to \mathbb{H}^d$,

$$u_i = \tfrac{1}{1+e^{-x_i}} \quad x_i = \log\left(\tfrac{u_i}{1-u_i}\right).$$

compute the density of $\vec{u}$.

derivations for the hypercube example. Define $\mathbb{H}^d = \left\{\vec{u} \in \mathbb{R}^d : 0 \leq \vec{u} \leq 1\right\}$. Let $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$. We will obtain the density of the random variable $\vec{u} = g(\vec{x})$ given by

$$u_i = \frac{1}{1+e^{-x_i}} \qquad \text{and} \qquad x_i = \log(\frac{u_i}{1-u_i})$$

We have the density of $\vec{x}$, given by

$$f_{\vec{X}}(\vec{x}) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(x-\vec{\mu}\right)' \Sigma^{-1} \left(x-\vec{\mu}\right)\right).$$

We can represent the density of $\vec{u}$ as

$$f_{\vec{U}}(\vec{u}) = f_{\vec{X}}(g^{-1}(\vec{u}))||J||$$

where $||J||$ is the absolute value of the determinant of the Jacobian of the function $g^{-1}(\vec{u})$. This Jacobian is clearly diagonal, as, for $i \neq j$,

$$\frac{\partial x_i}{\partial u_j} = 0$$

For the diagonal elements of J, we obtain

$$\frac{\partial x_i}{\partial u_i} = \frac{1}{u_i} + \frac{1}{1-u_i} = \frac{1}{u_i(1-u_i)}$$

As J is diagonal, its determinant is the product of its diagonal entries. Furthermore, as $0 < \vec{u} < 1$, $u_i(1-u_i) > 0$ for all i. Thus,

$$||J|| = \frac{1}{\prod_{j=1}^{d} u_i(1-u_i)}$$

Now, combining our value for $||J||$ with the fact that $g^{-1}(\vec{x}) = \log(\frac{\vec{u}}{1-\vec{u}})$ we obtain the density of $\vec{u}$ as

$$f_{\vec{U}}(\vec{u}) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}} \prod_{j=1}^{d} u_i(1-u_i)} \exp\left(-\frac{1}{2}\left(\log(\frac{\vec{u}}{1-\vec{u}}) - \vec{\mu}\right)' \Sigma^{-1} \left(\log(\frac{\vec{u}}{1-\vec{u}}) - \vec{\mu}\right)\right).$$

problem 2 This problem is based on replicating an analysis from "Malaria risk on the Amazon frontier" by de Castro et. al. (PNAS 2006). They use an unsupervised method known as a "grade of membership" or GoM model; this is a special form of mixture model. Your task will be to replicate a subset of their model-building and computation. This will involve probability modeling and fundamental optimization methods.

The data consists of features for a set of plots in the border of the Amazon inhabited by settlers. The objective of this analysis is to identify features associated with high and low risks of contracting malaria; for example, having high-quality walls for one's home would be expected to correspond to lower risk, whereas living close to water would likely raise one's risk.

We assume that each plot has features that originate from one of two distributions: high or low risk. Our objective is then to estimate both the parameters associated with each risk profile and the probability that each plot is in the high-risk group. Note that the approach used here does not make use of the actual occurance of malaria; instead, the authors assume there are two types of plots and attempt to identify them using only other features.

All features are categorical and differ in the number of levels; for example, there are only two levels for `chainsaw`, but there are five levels for `dist-hosp`. The data generating process underlying this model is:

Given $G_{N\text{x}2}$ and $\Theta_{2\text{x}p}$
for $n = 1, \cdots, N$ **do**
  for $p = 1, \cdots, P$ **do**

$$P(X_{n\text{x}p}|\overrightarrow{g}_n, \overrightarrow{\theta}_{L,p}, \overrightarrow{\theta}_{H,p}) = g_{L,n} \cdot \text{Mult}(\overrightarrow{\theta}_{L,p}, 1) + g_{H,n} \cdot \text{Mult}(\overrightarrow{\theta}_{H,p}, 1)$$

    where $\overrightarrow{g}_n = (g_{L,n}, g_{H,n})$
  end for
end for

Here, $X_{n\,x\,p}$ is a matrix of features for plot n, $\theta_{L,p}$ is the vector of multinomial probabilities for feature p in the low-risk group, $\theta_{H,p}$ is an analogous vector for the high-risk group.

Finally, $g_{L,n}$ and $g_{H,n}$ are the probabilities that plot n belongs to the low or high-risk group, respectively ($g_{L,n} = 1 - g_{H,n}$).

The data for this problem can be found in the `dat` folder. The file `data1985_area2.csv` contains the features for this problem, along with a column labelled `id`. This is an identifier for each plot and should **not** be included in your X matrix. The file `theta0list.Rdata` contains an R list with the initial values of $\theta_{H,p}$ and $\theta_{L,p}$ for each feature. If you are not using R, the same value are contained in the set of 49 CSVs found in `malaria_theta0.zip`. We have also included a CSV containing descriptions of each variable and its levels (`variables_theta0_1985.csv`) for your information.

question 2.1. (15 points) Write out the log-likelihood for this model. Be sure to specify all relevant variables and watch your subscripts.

question 2.2. (5 points) Write an function `ll(G, theta, X)` that computes the log-likelihood derived in 2.1.

question 2.3 (25 points) Find the MLE for (G, Θ) using a coordinate ascent approach, maximizing over parameters in the following sequence:

1. g_L, n for $n = 1, \cdots, N$
2. θ_L, j for $j = 1, \cdots, J$
3. θ_H, j for $j = 1, \cdots, J$

This should be implemented as function `gomMLE(X, G0, theta0)` where `G0` and `theta0` are the provided starting values for your iterations. Your function should return a list of the form `list(G.hat=G.hat, theta.hat=theta.hat, maxlik=maxlik)`, where the first two entries are the MLEs and the last is the maximized likelihood. When running this on the provided data, set G0 to 1/2 for all plots, and use the provided values for `theta0`.

Note: Your code can assume that the mixture has only two components (H and L), but it should not assume that the number of covariates, the number of levels per covariate, or the number of observations is fixed. It should be flexible enough to run on similar data without modification. We will test your code on an independent set of data with different number of plots, features and different covariates.

computing note. it is strongly recommended that all students use `R` for the computing tasks required in this course. we will, however, accept `MATLAB` code solutions to all assignments. please note, that **no guarantees of support, troubleshooting or assistance** will be provided for `MATLAB` code. any `MATLAB` code that produces an incorrect answer is likely to receive less partial credit than similar `R` code. please also note that all coding restrictions and guidelines described must also be followed by `MATLAB` users (with appropriate file extension changes: `.R` to `.M`). in short: if you are an experienced `MATLAB` user and are confident you can complete the assignments throughout the course, then you are welcome to use `MATLAB` instead of `R` (acknowledging all warnings).

what to submit. please submit written answers (using latex or word) to all relevant questions via email before midnight on Tuesday 9/16. Pick a username (lastname_firstname is recommended) throughout this class and email a ZIP file named as `username_ps1.ZIP` to **stat221.harvard.fall2014@gmail.com**. Important: Please remember to have an attachment and one of the keywords "stat221", or "pset" or "submission" in your subject. The ZIP file should contain two `R` scripts; one should be named as `username_ps1_prob1.R` with your code for problem 1, and the second should be named as `username_ps1_prob2.R` for problem 2 code. your `R` script file must contain all of the functions detailed above, must read the supplied data file, estimate the parameters, and produce the requested output, assuming all files are in the same folder. the ZIP file should contain a PDF file named as `username_ps1.pdf`, as well as the TEX or DOC file with the solution, named as `username_ps1.tex` or `username_ps1.doc`. the latex source of this pset is provided for your convenience. if you have any questions about code formatting and organization then please ask on Piazza. happy coding!