STATISTICAL COMPUTING - STAT771

Class Notes

Week 1

I made some introductory remarks about the scope and format of the class (see course information file).
I discussed the class textbooks, by Lange and by Wickham, and our use of R/Rstudio as a device to discuss statistics through computing.
We went over numbers in R, discussing the general idea that in 'standard' R, floating point numbers are represented with finite precision, usually 64 bits per number, which in base 10 corresponds to about 17 significant digits.

```
> options( digits=18 )
```

over-rides the default printing, and shows a fuller internal representation that R uses on a given number. E.g

```
> pi
[1] 3.14159265358979312
```

which is clearly incorrect after the last 3.
This lack of full precision affects computation. For example

```
> factorial(23)
[1] 25852016738884978212864
```

This can't be right since the answer should certainly be a multiple of 1000 (20x10x5), and it is why we need to be careful when doing all sorts of computations. E.g. we are advised not to evaluate choose(A,B) by first computing factorials and then taking ratios; we ought to cancel first, or better work on a log scale, taking differences of `lgamma` functions and then exponentiating to reveal the final ratio.
Even with these finite-precision numbers, R is good at infinities, in the sense that it knows 1/0 = Inf, but 0/0 is not a number.
We discussed the log-sum-exp trick for computing the logarithm of a sum, using functions:

```
f0 <- function(y){
t <- log(sum(y))
return(t)
}

f1 <- function(y){
x <- log(y)
m <- max(x)
t <- log(sum(exp(x-m))) + m
return(t)
}

f2 <- function(y){
i <- which.max(y)
```

```
x <- log(y)
m <- x[i]
t <- log1p(sum(exp(x[-i]-m))) + x[i]
return(t)
}

v <- exp( c(0,-50) )
```

Notice that only `f2(v)` gives the correct answer, as it uses a more accurate approximation of logarithm (log1p) when the argument is near unity.
We went on to an interesting way to use numbers in very high precision, using the package `Rmpfr`.
Try

```
> library("Rmpfr")
> Pi <- Const("pi", prec=260)
> Pi
1 'mpfr' number of precision 260 bits
[1] 3.14159265358979323846264338327950288419716939937510582097494459230781640628620
> factorialMpfr(23)
1 'mpfr' number of precision 75 bits
[1] 25852016738884976640000
```

Further, the `Rmpfr` package contains numerous facilities for computing at high precision, which may be interesting to explore further.

Week 2

Class 2

Today we discuss some methods to calculate frequently occurring probabilities.

- **Normal tail areas**:
  We recall the standard normal density $\phi(x)$ and cumulative distribution function $\Phi(x)$, and we discuss several approaches, beyond just looking things up in a book!

  1. Taylor expand $\exp(y) = 1 + y + y^2/2 + \ldots$, getting $\exp\{-z^2/2\} = \sum_{n=0}^{\infty}(-1)^n \frac{z^{2n}}{n!2^n}$; then integrate term by term over the interval 0 to x, noting that is sufficient, for $x > 0$, since the remaining integral equals 1/2. Thus

  $$\Phi(x) - \frac{1}{2} = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n+1}}{n!(2n+1)2^n}$$

  Thus an approximate form is had by truncating the sum at some big value N. The main problem with this is the loss of significant digits associated with taking differences between finite-precision numbers of similar magnitude. Numerical analysts try to avoid such alternating sums for this reason.

2. A positive term expansion may be derived from $g(x) = \frac{Phi(x)-1/2}{\phi(x)}$, which is an odd function ($g(-x) = -g(x)$), and hence has an expansion around $x = 0$ in odd powers of x:

$$g(x) = \sum_{n=0}^{\infty} c_n x^{2n+1}$$

A close inspection of g allows us to evaluate the coefficients c_n; e.g. by working out that $g'(x) = xg(x) + 1$, and hence... $c_1 = 1$ and $c_n = \frac{c_{n-1}}{2n+1}$, from which we deduce that $c_n = \frac{1}{(2n+1)!!}$, where the denominator is the product of odd factors down from $(2n+1)$ to 1, and clearly $c_n > 0$ for all n. We can approximate $g()$ by evaluating the sum to N terms, and then we can use the result to get Φ.

It turns out that `pnorm()` and other calculators do not use either of these series approximations. To see how they do it, recall first a famous number, the golden ratio (again, we call it ϕ, not the normal density), ϕ satisfies $\phi = 1 + 1/\phi$ (sketch rectangle, where the ratio of the longer side to the shorter is the same as the ratio of the longer + shorter over the longer). We know $\phi = \frac{1+\sqrt{5}}{2}$, which is approximately 1.6, but we also get the interesting result

$$\phi = 1 + 1/[1 + 1/[1 + 1/[1 + 1/[1 + 1/[1 + 1/...]]]]],$$

which is a special continued fraction. These objects are difficult to deal with mathematically, but they are very conducive to computation, since they suggest a recursive algorithm, where in you start deep within the fraction, approximate the final ratio, invert sum, invert sum, etc up to the finish.
In 1812, Laplace presented the following continued fraction expression for Mills ratio $R(t) = \frac{1-\Phi(t)}{\phi(t)}$, from which an approximation to $\Phi(t)$ is readily derived:

$$R(t) = 1/[t + 1/[t + 2/[t + 3/[t + 4/[t + 5/[t + 6/[t + ...]]]]]]]$$

R source in norm.R (see attached) compares the two series approximations with two continued fraction approximations to $\Phi(t)$, revealing that Laplace's method is very accurate.

- **Poisson Binomial**:
Given a series $X_1, X_2, ..., X_n$ of mutually independent Bernoulli trials, the Poisson Binomial distribution is the distribution of $S_n = \sum_{i=1}^{n} X_i$; at issue is the success probability the trials, since equal probabilities would correspond to the well-known binomial distribution, but different probabilities lead to something else. Let the probabilities be $\theta_1, \theta_2, \ldots, \theta_n$, and let bar theta denote the average value, then it is readily shown that

$$\begin{aligned} E(S_n) &= n\bar{\theta} \\ Var(S_n) &= \sum_{i=1}^{n} \theta_i(1-\theta_i) \end{aligned}$$

Somewhat counter-intuitively, it holds that $var(S_n) \leq n\bar{\theta}(1-\bar{\theta})$, with the upper bound (max variance) case being the binomial case.
The Poisson Binomial arises in various applications; in large-scale Bayesian hypothesis testing it arises as the posterior distribution of the number of non-null hypotheses.
At issue is how to compute $P(S_n = j)$ for $j = 0, 1, \ldots, n$. It's tricky, since it seems to depend on which particular j trials are successes. In fact, with $\mathcal{F}_j$ = size j subsets of $\{1, 2, \ldots, n\}$,

the mass is

$$P(S_n = j) = \sum_{A \in \mathcal{F}_j} \left[\prod_{i \in A} \theta_i \right] \left[\prod_{i \in A^c} j(1 - \theta_i) \right]$$

which seems very difficult to evaluate. Letting $q_j(i) = P(S_j = i)$, we can condition on S_{j-1} to get

$$q_j(i) = \sum_{s=0}^{j-1} P(S_j = i, S_{j-1} = s) = \sum_{s=0}^{j-1} P(S_j = i | S_{j-1} = s) q_{j-1}(s),$$

and then $P(S_j = i | S_{j-1} = s) = (1 - \theta_j) 1\{s = i\} + \theta_j 1\{s = i - 1\}$, and so

$$q_j(i) = (1 - \theta_j) q_{j-1}(i) + \theta_j q_{j-1}(i - 1).$$

This simple recursion allows us to build up a matrix full of probabilities, the last row of which corresponds to the distribution of S_n.

Class 3

R package presentations

Weex 3

Class 4

Numerical Linear Algebra: topics relevant to the linear model
Two asides on simple matrix procedures:

1. If the $m \times m$ matrix $L = (l_{i,j})$ has $l_{i,j} = 0$ whenever $j > i$, then we say L is lower triangular. A simple scheme, called forward substitution, allows us to solve the equation $Lx = b$, given L and b, in x, whenever all $l_{i,i} \neq 0$.
   Similarly backward substitution works for upper triangular matrix equations.

2. Sweep on a diagonal entry. Given an $m \times m$ matrix A, and a diagonal $a_{k,k} \neq 0$, we sweep on this entry in creating a new matrix $\tilde{A}$ with entries:

$$\tilde{a}_{k,k} = -1/a_{k,k}$$

$$\tilde{a}_{i,k} = a_{i,k}/a_{k,k}; \quad \tilde{a}_{k,j} = a_{k,j}/a_{k,k}$$

$$\text{and for other } (i,j), \quad \tilde{a}_{i,j} = a_{i,j} - a_{i,k} \cdot a_{k,j}/a_{k,k}$$

   we work a simple 3x3 example to illustrate.

Now on to linear regression:

We recall $Y = X\beta + \epsilon$ for normal vector epsilon, X an $n \times p$ design matrix, and β a $p \times 1$ regression coefficient;
...three kinds of lies...
...there's more to heaven and earth Hortatio...

But in any case we do consider this; I like to start with likelihood, so I write the log likelihood in terms of sum of squares $(Y - X\beta)^T(Y - X\beta)$ and see that ML equals LS for i.i.d. Normal errors.
We use vector differentiation to get the optimal solution, satisfying the normal equations $(X^TX)\hat{\beta} = X^TY$... and , if the inverse exists, $\hat{\beta} = ...$
We discuss sampling properties of $\hat{\beta}$, (mean and variance), and other objects we compute in regression (e.g. fitted values, residual sum of squares).
Like with `pnorm()`, where we discussed three options for algorithmic approaches, we here consider three options

1. sweep: take the case of simple linear regression as an instance, and sweep on the two diagonal entries of X^TX in the augmented matrix (including X^TY), to find that all objects of interest may be neatly computed (plus Lange's general theorem , 7.5.2)
2. Cholesky decomposition of (X^TX): note the existence of a lower triangular L with positive diagonal elements as the Cholesky factor may be proved also from Lange's 7.5.2 (i.e. the sweep theorem).
   Using Cholesky plus backsolve/forwardsolve produces $\hat{\beta}$.
3. using $X = QR$: discussing the standard form with Q $n \times p$ and having orthonormal columns, and R upper triangular, uniquely determined (if $r_{ii} > 0$) if X is of full column rank.
   Working through the normal equations simplifies to a back solve.

Class 5

Numerical Linear Algebra topics, continued

Weighted least squares: Extending from the linear model used in last class, suppose each Y_i has variance σ_i^2, possibly different, and put W to be an $n \times n$ diagonal matrix with diagonal entries $1/\sigma_i^2$. Show by vector differentiation that the new normal equations are $(X^TWX)\hat{\beta} = X^TWY$
We comment on the Gauss-Markov theorem, the sense of robustness it carries, and the different sense of robustness as insensitivity to outliers, which these linear model estimators do not have.
Next we move into a topic about reducing dimensionality of a data set; specifically if the data structure is an $n \times p$ matrix X. Consider, e.g. $p = 2$, and show a scatterplot, also suppose the columns are centered to have mean 0. In case there is high correlation between the variables, then the major axis (drawn) explains most of the variation in the data, and using that axis only would entail only a small reduction in the information content.
We discuss the idea of directions of maximal variance and the idea of perpendicular distance from a point to a line (going through the origin), using a simple right triangle; and we note that maximizing the sum of squared projected distance (equals variance) is the same as minimizing the sum of perpendicular distances. To examine this more generally, we first take a diversion.
Suppose we have a $p \times p$ covariance matrix Sigma, say positive definite, which is the covariance matrix of a p-vector Z that also has mean 0. For any other p-vector v, we have the real-valued random variable v^TZ, and we might try to find v to maximize variance. We need to constrain the search to vectors v of common length (say $v^Tv = 1$), otherwise it's not well defined. But doing so we find that the variance of the new variable is $v^T\Sigma v$ and that a Lagrange-multiplier objective function is

$$f(v) = v^T\Sigma v + \gamma(1 - v^Tv)$$

...go through carefully to see the optimal v is an eigenvector of Σ, one corresponding to the largest eigenvalue (which is the variance of the linear combination).
With the most variable direction found, call it v_1, we go on to find a second vector (direction) v_2, which is orthogonal to v_1 ($v_2^T v_1 = 0$), which is also unit, ($v_2^T v_2 = 1$), and which, among these, maximizes the variance $v^T \Sigma v$. So we construct two Lagrange multipliers and have the objective function

$$f(v) = v^T \Sigma v + \gamma_1(1 - v^T v) + \gamma_2(v_1^T v)$$

taking a vector derivative:

$$(df/dv) = 2\Sigma v - 2\gamma_1 v + \gamma_2 v_1,$$

which we set to zero to solve premultiplying by v_1^T gives

$$0 = 2v_1^T \Sigma v - 2\gamma_1 v_1^T v + \gamma_2,$$

and using orthogonality of the solution, it must be

$$0 = 2v_1^T \Sigma v_2 + \gamma_2.$$

This is a scaler equation, so it must be that $\gamma_2 = (-2)v^T \Sigma v_1$, but we know that $\Sigma v_1 = \lambda_1 v_1$, since v_1 is the eigenvector associated with eigenvalue λ_1, so $\gamma_2 = (-2)\lambda_1 v_2^T v_1 = 0$.
Having figured that $\gamma_2 = 0$, we go back to (df/dv) and find that v_2 must be another eigenvector of Σ, and to have maximal variance it must be the one associated with the second largest eigenvalue, say λ_2.
Continuing in this way gives us a sequence of $v_1, v_2, \ldots, v_p$, of eigenvectors of Σ corresponding to decreasing eigenvalues, and these are such that the random variables $v_i^T Z$ reflect random transformations corresponding to ever smaller variance.
Having confirmed this we still are not in a position to say how to reduce the dimension of a data set X, but here we rely on another basic fact from linear algebra, the proof of which we'll outline next time:
For any $n \times p$ X, we may write $X = UDV^T$ such that U is $n \times p$ with orthonormal columns, so $U^T U = I_n$, V is $p \times p$ orthogonal , so $V^T V = I_p$, and D is a diagonal matrix with diagonal entries d_{ii} called the 'singular values'. This decomposition can be arranged so that all $d_{ii} \geq 0$, and $d_1 1 \geq d_2 2 \geq d_3 3$....
Further, if $rank(X) = k$, then $d_{kk} > 0$ but $d_{k+1,k+1} = 0$, so the decomposition tells the rank of X.
Considering the entries of X, it also follows that $X = \sum_{j=1}^{k} d_{jj} u_j v_j^T$ and so X is represented here as a combination of rank 1 matrices; each outer product $u_j v_j^T$ is a rank 1 matrix, since the columns are all multiples of u_j.
It is also true that the best rank q approximation to X, terms of so-called Frobeneous norm (sum squared entries) is

$$\tilde{X} = \sum_{j=1}^{q} d_{jj} u_j v_j^T,$$

so this is a way to reduce the dimensionality of a data set.
Observe further that for centered X, the $p \times p$ matrix $X^T X = V^T D^2 V$ is itself the sample covariance, up to the scaler multiplier $1/n$ (or perhaps $1/(n-1)$), and this may be viewed as the Sigma matrix from the first part of class. Here $V = (v_j)$ are the eigenvectors of Sigma, and the squared singular values are the eigenvalues. The directions in $V = (v_j)$ are called the principal component vectors. The first one is the direction of maximal variance, and subsequent ones are orthogonal carrying decreasing variance.

Week 4

Class 6

Singular value decomposition, continued:

We start with an $n \times p$ matrix $X = (x_{i,j})$ and we claim $X = UDV^T$, noting it's either U $n \times p$ D $p \times p$ and V $p \times p$ or U $n \times n$, D $n \times p$ and V $p \times p$; both expressions are convenient. In either case $U^TU = I$ and $V^TV = I$. And D has nonzero entries only on its diagonal, which are called the singular values. The decomposition can be arranged so that the diagonal entries are ≥ 0 and decreasing. In both versions the columns of U are orthonormal and the columns of V are orthonormal; when square these matrices are both orthogonal. We noted before that X can be expressed as a sum of rank-1 matrices

$$X = \sum_{j=1}^{p} d_{j,j} u_j v_j^T,$$

that $X^TX = VD^2V^T$, and so the squared singular values are eigenvalues of that matrix. The columns of V are called loadings in principal components; V is also called the rotation matrix. As $XV = UD$, the rotation transform of X gives orthogonal columns with decreasing variances associated with the squared singular values. These are sometimes called the principal component values.

We commented that the rank of X equals the number of non-zero $d_{i,i}$, and the best rank k approximation to X (in Frobeneous norm) equals $\tilde{X} = U\tilde{D}V^T$, where $\tilde{D}$ holds the first k positive singular values on the diagonal, and is everywhere else zero.

What we want to do today is to work through the beginnings of a proof that such a decomposition of X can be made.

Given X, note that for any p-vector v, Xv is an n-vector, whose length $||Xv|| = sort\{v^TX^Tv\}$. We say the L_2 norm (or just the norm here) of the matrix $||X||$ is the maximal value of all projections of X, i.e. $||X|| = max_{v:v^tv=1}||Xv||$.

Now suppose that λ_1 is that maximal value (> 0), and that this value occurs at vector v, and is in the direction of unit n-vector u, so that $Xv = \lambda_1 u$.

With this start, begin to construct two orthogonal matrices U ($n \times n$) and V ($p \times p$), yet to be determined but both orthogonal basis of their respective spaces. $U = [u, U_1]$, $V = [v, V_1]$.

Now form $Y = U^TXV$, breaking it out blockwise as

$$Y = \begin{pmatrix} u^TXv & u^TXV_1 \\ U_1^TXv & U_1^TXV_1 \end{pmatrix},$$

being careful about all these matrices conforming.

Observe that the upper left block is λ_1, by construction, the lower left block is a 0 matrix by orthogonality of u with U_1, and call the other pieces

$$Y = \begin{pmatrix} \lambda_1 & w^T \\ 0 & Y_1 \end{pmatrix}$$

Try to compute the norm of Y: by one method, $||Y|| = ||X||$, since Y is an orthogonal transform, and thus lengths are preserved.

By a second method, $||Y|| \geq ||Yc||$ for any vector c, so take c to be

$$c^T = \frac{1}{\sqrt{\lambda_1^2 + w^T w}}[\lambda_1, w^T]$$

and compute $||Y||^2 \geq \lambda_1^2 + w^T w$, forcing $w = 0$.
This begins a computation to diagonalize X through orthogonal transforms. Repeating the calculation on Y_1, and so on, gives $U^T X V = diagonal...$ with non-negative diagonal entries (being maximal projection lengths).
Using that U and V are orthogonal, we know $U^T = U^{-1}$, e.g., so we can left multiply by U and right multiply by V^T to get $X = UDV^T$, as hoped.

Multidimensional scaling:

Given n data points (rows of X) in p-dimensions, the (squared) pairwise distance between points is, say,

$$\delta_{i,j} = (x_i - x_j)^T (x_i - x_j)$$

It is often convenient to attempt to find n points in q-dimensions (e.g. $q = 2$), smaller than p, for which the interpoint distances amongst the y's are as close as possible to those among the x's, especially as a way to help visualize the data.
You will show in homework that two matrices X and Y share inter-row distances iff $XX^T = YY^T$.

Class 7

We begin with a little speculation about the DNA of statistics...recalling the tool metaphor in conjunction with Cavalli-Sforza's work.
We go on to revisit the MDS problem, noting that $UD^2U^T = XX^T$, and the best rank q approximation to this must be $U\tilde{D}^2U^T$, with $\tilde{D}$ containing the first q singular values of D. This is indeed YY^T where $Y =$ first q columns of UD; i.e. the first q principal component variables.
We see the PC's have multiple interpretations: as directions of maximal variance; as directions of minimal perpendicular distances; and now as lower points with optimal interpoint distances approximating the actual interpoint distances.
We comment that MDS goes further with non-Euclidean distances and it is widely used in multivariate analysis.

Next we move on to two examples of penalized regression:

Linear model $Y = X\beta + \epsilon$ for iid normal error components, and now consider a Bayesian analysis and posterior distribution

$$p(\beta|\text{data}) \propto p(\text{data}|\beta)p(\beta)$$

$$\log \text{posterior} = \text{constant} + \log\text{-likelihood} + \log \text{prior},$$

and recall the log likelihood is a negative multiple of $S(\beta)$, the total sum of squares . Different priors give different objective functions that we may optimize to yield different estimators:
e.g.1 prior $\beta \sim Normal[0, \tau^2 I_p]$; work out that the mode of the posterior , which, equivalently minimizes:

$$f(\beta) = \frac{1}{2}S(\beta) + \frac{\lambda}{2}\beta^T\beta,$$

where λ is a variance ratio > 0, which, after some calculus, gives

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

We discuss some facts; e.g adding a bit on the diagonal makes that matrix full rank; In sampling theory terms, there is an increased bias and a reduced variance compared to OLS.
In computational terms, $\hat{\beta}_{ridge}$ is shorter as a vector in R^p.
Next we consider a Laplace prior

$$\log p(\beta) = \text{constant} - \sum_{j=1}^{p} |beta_j|$$

We take $p = 1$, with the single X centered and scaled to have $\sum x_i^2 = 1$, so $\hat{\beta}_{OLS} = y^T x$.
Then,

$$f(\beta) = \frac{1}{2} \sum_i [y_i - x_i \beta]^2 + \lambda |\beta|,$$

which is not differerntiable at $\beta = 0$. We study its derivatives elsewhere, finding optimal values

$$\hat{\beta}_{lasso} = \begin{cases} \hat{\beta} - \lambda & \text{if } \hat{\beta} > \lambda \\ \hat{\beta} + \lambda & \text{if } -\hat{\beta} > \lambda \\ 0 & \text{if } \sim \end{cases}$$

and we discuss this shrinkage/selection effect in 1-d.

Week 5

Class 8

On distance covariance: Last time we talked about this without defining. So given data x_i (in R^p) and y_i (in R^q), say for units $i = 1, 2, \ldots, n$, and viewing these as realizations of two distributions, we seek a statistic to measure their dependence. For all i, j write $a_{i,j} =$ distance(x_i, x_j); $b_{i,j}$ same for y_i, y_j, and center these distance matrices over rows/columns as

$$A_{i,j} = a_{i,j} - a_{i,\cdot} - a_{\cdot,j} + a_{\cdot,\cdot},$$

where $\cdot$ is average, and same for $B_{i,j}$, then form

$$\frac{1}{n^2} \sum_i \sum_j A_{i,j} B_{i,j},$$

which is the square of the distance covariance between the two measures: $(dCov(x, y))^2$.
Putting in x, x instead of x, y gets $(dVar(x))^2$, and similarly $(dVar(y))^2$, from which the distance correlation is the non-negative number

$$dCor = \frac{dCov}{\sqrt{dVar \cdot dVar}}$$

This is easy enough to compute, but it also has an important theoretical property, especially compared to Pearson's correlation. Namely, the population value equals 0 iff X and Y are independent.
We don't try to prove it here, but we sketch the proof via characteristic functions $\phi_X(s)$, $\phi_Y(t)$ and $\phi_{X,Y}(s, t)$, the last of which equals the product of the first two iff X and Y are independent. The interesting thing about $dCov$ is that it is exactly a sample version of a norm measuring the distance between these characteristic functions....i.e. it is the norm evaluated on empirical

characteristic functions.

More on Lasso: Last time we worked out, for the simple linear model, normal errors, that a log posterior, under a Laplace prior, is equivalent to the negative of $S(\beta) + \lambda \sum_j |\beta_j|$, where $S(\beta)$ is the sum of squares function and λ is positive. We worked in the simple case $p = 1$, standardized x_i, so that $\hat{\beta}_{OLS} = \sum_i (y_i x_i)$ and the objective function for lasso becomes:

$$(\hat{\beta}_{OLS} - \beta)^2 \frac{1}{2} + \lambda |\beta|,$$

which last time we analyzed to find

$$\hat{\beta}_{lasso} = T(\hat{\beta}_{OLS}, \lambda) = \begin{cases} \hat{\beta}_{OLS} - \lambda & \text{if } \hat{\beta}_{OLS} > \lambda \\ \hat{\beta}_{OLS} + \lambda & \text{if } \hat{\beta}_{OLS} < -\lambda \\ 0 & \text{if } \sim \end{cases}$$

We comment here on an algorithm that works for the general problem with $p > 1$. It's an iterative algorithm, with a number of cycles. Each cycle involves updating each of the coordinates in beta.
In updating β_j, we first construct the predictor

$$\hat{y}_i = \sum_{k \neq j} \beta_k x_{i,k},$$

constructed from non-j variates, the residual $r_i = y_i - \hat{y}_i$, and the OLS value $\beta_{OLS-j} = \sum_{i=1}^n x_{i,j} r_i$. Finally we reset the j'th beta value to $T(\hat{\beta}_{OLS-j}, \lambda)$, and continue cycling.
This so-called 'shooting' algorithm, also called the coordinate descent algorithm, works in a number of special cases like the lasso objective. We'll study it's properties further after learning more about convex optimization.

General optimization: The examples studies so far mostly involve the linear model; statistics is full of optimization problems, as objective functions arise naturally in data analysis, either through likelihoods $p(\text{data}|\theta)$ or through Bayesian posteriors $p(\theta|\text{data})$. [I recall the general view, data $\sim P \in \mathcal{P}$ = model... encoded by θ , as characterizing the statistical approach to data].

Nelder-Mead algorithm: This was introduced as a general optimization tool for multi-parameter, unconstrained objective functions, especially log likelihood $l(\theta)$. It creates a sequence of sets $S_1, S_2, \ldots, S_B$, each of which holds $p+1$ θ_j values, which are points in p-space, and set so that the points do not reside on a (p-1)-dimensional hyperplane in p-space. e.g, 3 points forming a triangle with positive area in 2-space. We sketch out the basic plan, namely to nominate new simplex points which may end up replacing old simplex points, creating a series of simplices that out to converge in some sense to surround the MLE $\hat{\theta} = argmax_\theta l(\theta)$, say.
We go through the NM rules, which are encoded in the R function '`optim`' stopping rules:

$$\begin{cases} \text{either} & \text{points are too close together in } S \\ \text{or} & l(\theta_j) \text{ values are too close together} \\ \text{or} & \text{we have run for too long} \end{cases}$$

On the update rules, three of the simplex points are labeled: θ_b = that value having the best likelihood; θ_s=that value having the smallest, and θ_{s+1}, that value having the 2nd smallest.

Such three points are found, regardless of the dimension.
We find a centering point, average of the non-worst points θ_0=mean of θ_j, not including θ_s and we first propose a reflected point θ_r along a line directed away from θ_s and through θ_0, outside the current simplex.
If the likelihood $l(\theta_r)$ is the best so far, then we think we're heading the right way, so we further expand to a new point θ_e further out along that line.
If the new likelihood $l(\theta_e)$ is best, then take it, so $S_{update} = S_{old} - \theta_s + \theta_e$ (these being set additions and deletions).
Otherwise, as long as the $l(\theta_r)$ exceeds $l(\theta_{s+1})$, take it, so $S_{update} = S_{old} - \theta_s + \theta_r$.
It remains to consider cases where the reflected point has poor properties
either:

1. External Contraction: $l(\theta_r) > l(\theta_s)$, in which case we consider a new point θ_c contracted back from θ_r towards θ_0, or

2. Internal: if $l(\theta_r) < l(\theta_s)$, in which case we find a point $theta_{cc}$ inside the simplex, along that line between θ_s and θ_0

in either case if the contracted point has better likelihood than $l(\theta_s)$, then take it $S_{update} = S_{old} - \theta_s + theta_{contracted}$.
Otherwise we need a 'bail out' clause, which we'll discuss next time.

Week 6

Class 10

We continue with optimization transfer, introduced last week, recalling it's a sequence of minorizing functions $g_m(\theta) \leq l(\theta)$ and a sequence of points $\theta_1, \theta_2, ...$ that touch $g_n(\theta_n) = l(\theta_n)$, and for which $\theta_{m+1} = argmax g_m(\theta)$.
We start with an aside about the Multinomial:
On multinomial: Recall that if we sample N units i.i.d. from a population, and each one takes one of K possible values, with probabilities $\theta = (\theta_1, ..., \theta_K)$, where these sum to 1 and are non-negative, then we may code the output with binary vectors: Z_i, $i = 1, ..., N$, where $Z_i = (Z_{i,1}, Z_{i,2}, ..., Z_{i,K})$ and where each $Z_{i,j} \in 0, 1$ and further, only one per unit is 1; i.e. $\sum_j Z_{i,j} = 1$. With this coding, the likelihood for θ provided by the sample is $\prod_i \prod_j \theta_j^{Z_{i,j}}$, which may be readily computed with. E.g., a Lagrange-multiplier argument readily gives the MLE $\hat{\theta}_j = Z_{\cdot,j}/N$, where $\cdot$ stands for sum.
Aside on a Poisson example: A semi-toy example, simplified from a problem about T-cell receptor sequencing. Suppose we have N wells (embedded in 96 well or 384 well plates for cell-based assays), and into well i an investigator places n_i cells. We discuss the experiment wherein cells are treated with a chemical 6thioguanine (6TG), and such cells will die unless they have one of a number of somatic mutations within their HPRT gene. The rate of mutations in the cell population is, say theta, possibly small (e.g. 10^{-5} or so), but treated as an unknown parameter. The likelihood, upon seeing X_i mutants in well i, is naturally derived from the $Poisson(n_i\theta)$ model, and has logarithm

$$l(\theta) = \sum_i -n_i\theta + x_i \log(n_i\theta) - \log(x_i!),$$

from which via differentiation we get the MLE $\hat{\theta}_{MLE} = \frac{\sum_i x_i}{\sum_i n_i}$, just the total number of mutants over the total number of cells plated out.

In the actual experiments, one grows cells in an incubator, and we cannot detect x_i, rather we can measure $Y_i = 1\{X_i > 0\}$: we know whether the well is fertile ($Y_i = 1$) or sterile ($Y_i = 0$). Now the log likelihood is more complicated

$$l(\theta) = \sum_i y_i \log[1 - e^{-n_i\theta}] + (1 - y_i)(-n_i\theta),$$

which does not usually have a closed form expression (unless either all y_i are the same or all n_i are the same). One way to proceed is to take the complete - data log likeliood, from above, and take its expected value given the observed data, and taken under a current estimate of the parameter θ, say θ_m.
Call this $Q_m(\theta) = E\{\log p(X|\theta)|Y = y, \theta_m\}$. In our case we find this readily as

$$Q_m(\theta) = \sum_i \{-n\theta_i + \hat{x}_i \log(n_i\theta) - \text{ stuff }\},$$

where $\hat{x}_i = E\{X_i|Y_i = y, \theta_m\}$, and stuff involves θ_m, y_i, and x_i, but no θ.
Observe that the maximum of this function is readily computed, and named $\theta_{m+1} = \frac{\sum_i \hat{x}_i}{\sum_i n_i}$, using the above formulation. Further it's not too hard to sort out $\hat{x}_i = 0$ if $y_i = 0$ and $= n_i\theta/[1-e^{-n_i\theta}]$ if $y_i = 1$, taken at $\theta = \theta_m$.
We have developed the two key steps of the E-M algorithm; E to construct Q, and M to maximize Q. We thus get a sequence of iterates $\theta_1, \theta_2, ...$, alternating between the two steps.
If we can prove that this is a valid optimization transfer function if we can relate the Q_m's to the g_m's from above.
Consider $g_m(\theta) = Q_m(\theta) - Q_m(\theta_m) + l(\theta_m)$; so clearly $g_m(\theta_m) = l(\theta_m)$; as to the minorization condition, look at the difference

$$\begin{aligned} l(\theta) - g_m(\theta) &= -[Q_m(\theta) - Q_m(\theta_m)] - [l(\theta_m) - l(\theta)] \\ &= -E\left\{\log \frac{p(X|\theta)}{p(X|\theta_m)}|Y = y, \theta_m\right\} - [l(\theta_m) - l(\theta)] \end{aligned}$$

... go through this carefully ...; discrete case use Bayes rule....use that X = complete data, Y= observed data, and $Y = t(X)$ for some function $t()$ that does not involve theta (it reflects information loss) then we get ≤ 0 .... using Jensen's inequality , which implies that EM is an optimization transfer algorithm.

Class 11

We start recalling an interesting statistical facts about likelihood inference. Log likelihood $l(\theta)$, maximized at $\hat{\theta}$... score function $S(\theta) = \frac{d}{d\theta} l(\theta)$ and $i(\theta) = -\frac{d^2}{d\theta^2} l(\theta)$.
We note that in many models $\sqrt{i(\hat{\theta})}[\hat{theta} - \theta_{true}]$ is approximately $Normal[0, 1]$ in its sampling distribution, from which we can derive standard asymptotic confidence intervals [e.g., in 1-d, $\hat{\theta} \pm 2/\sqrt{i(\hat{\theta})}$].
$i(\hat{\theta})$ is usually called the 'observed' information, distinguished from the Fisher information $I(\theta)$ which, in regular models equals $E\{i(\theta)\}$. Thinking that the log likelihood $l(\theta)$ comes along with hat theta, and thus its curvature $i(\hat{\theta})$ comes along with it, we might say that hat theta carries with it a measure of its own fallibility!
The details of such approximations are part of a math stat class [usually discussed for i.i.d. sampling, but in fact the approximation is valid well beyond i.i.d.] This is not the topic of 771;

rather we are interested mostly in how to compute $i(\hat{\theta})$ in various cases.
In the missing-data realm, the EM algorithm does not involve ever even evaluating $l(\theta)$, let alone a curvature, and so it is not clear how from EM output we might approximate $i(\hat{\theta})$.
We start with complete data $X = (Y, Z)$, for observable data Y and missing data Z, each possibly with its own complicated structure, but governed by probabilities

$$f_\theta(x) = f_\theta(y) f_\theta(z|y)$$

Taking logs, differentiating twice, and multiplying by -1, we get

$$i_{complete}(\theta) = i_{obs}(\theta) + i_{miss}(\theta),$$

where

$$\begin{aligned} i_{complete}(\theta) &= E[-\frac{d^2}{d\theta^2} l_{comp}(\theta)|Y = y] \\ i_{obs}(\theta) &= -\frac{d^2}{d\theta^2} l_{obs}(\theta) \\ i_{miss}(\theta) &= E\{-\frac{d^2}{d\theta^2} \log p(Z|y, \theta)|Y = y\} \end{aligned}$$

and so, the object of interest

$$i_{obs}(\theta) = i_{complete}(\theta)\{1 - i_{miss}/i_{complete}\}$$

We note that at $\hat{\theta}$, $i_{complete}$ is often easy to evaluate; e.g. in the Mutation Frequency example $i_{complete}(\theta) = \sum_i \hat{Z}_i/\theta^2$.
The second factor requires a deeper study. First notice that the EM sequence produces $\theta_1, \theta_2, ...$ through two steps E and M per cycle; yet taken together, we really have a mapping $\theta_{m+1} = \psi[\theta_m]$... show in MF example the formula for this ...; indeed, the MLE is a fixed point of this mapping.
An amazing fact about all this the following, which holds under suitable regularity conditions:
Thm: $\psi'(\hat{\theta}) = 1 - i_{miss}(\hat{\theta})/i_{complete}(\hat{\theta})$, and thus by keeping track of the iterates near the MLE, we should be able to approximate psi' and finally produce a value for the observed information.
The proof of this theorem proceeds by noting that the M step, in regular settings, is found by setting the complete-data score to zero: i.e. θ_{m+1} solves $0 = E_{\theta_m}[S_{complete}(\theta_{m+1})|Y = y]$.
Dropping the subscript m, $0 = E_\theta[S_{complete}(\psi(\theta))|Y = y]$ defines the mapping psi implicitly. Next differentiate this (allowing differentiation under the expectation), and one finds the result.

We move next to mixture models:

Complete data $X = (X_i)$; $X_i = (Y_i, Z_i)$... observed data (Y_i) and missing (Z_i), each of which is simple binary vector $Z_i = (Z_{i,1}, ..., Z_{i,K})$, with only one $Z_{i,j}$ equalling 1, the rest 0.
Say $Y_i|Z_{i,j} = 1 \sim f_j$, a component conditional density.
There are several examples. From large-scale hypothesis testing, we have $K = 2$; f_1 may be a null distribution for data on unit i; f_2 is an alternative distribution, the marginal distribution is

$$f(y) = \theta_1 f_1(y) + \theta_2 f_2(y)$$

Having fit the model $P[Z_{i,1} = 1|Y_i = y_i]$ is sometimes called a 'local false discovery rate', since it is the chance of being null for that unit, and it would be a false discovery if that unit were

placed in the list of rejected nulls.
Call this value e_i; notice that the list of rejected nulls is List $= \{i : e_i \leq \text{threshold}\}$ [just like a p-value-style list].
Now the conditional expected number of false discoveries on the list, given the data is

$$E\{\#FD|data\} = \sum_i 1[i \in \text{ List}]P[Z_{i,1} = 1|data] = \sum_i 1[e_i \leq \text{ threshold}] \cdot e_i$$

Notice that e_i being small puts you on the list of discoveries; it also tells you how probable it is that you should not be on the list! As with the MLE at the beginning of class, the e-value carries with it a measure of it's own fallibility!

Week 7

Class 12

Continuing with mixture models, recall the complete data model:
$X_i = (Y_i, Z_i)$, $i = 1, 2, ..., n$, are i.i.d. such that $Z_i = (Z_{i,1}, ..., Z_{i,K}) \sim Multinomial_K\{1, \theta = (\theta_1, ..., \theta_K)\}$ and $Y_i|Z_{i,j} = 1 \sim f_j$, a component density.
Case 1: The components $f_1, ..., f_K$ are known. [Aside, if they are linearly independent functions, then the mixing proportions theta are identifiable].
The log likelihood for θ is

$$l(\theta) = \sum_i \log \left[\sum_j \theta_j f_j(y_i) \right],$$

which is difficult to work with, as one can see in trying some differentiation. Instead, the complete-data log likelihood is

$$l_c(\theta) = \log \prod_i \prod_j \{\theta_j f_j(y_i)\}^{Z_{i,j}} = \sum_i \sum_j Z_{i,j}[\log(\theta_j) + \log f_j(y_i)]$$

Essentially, we go from a sum of logs of sums to a sum of sums of logs, which is much easier to handle. The E-step of the EM algorithm produces

$$Q_m(\theta) = \sum_i \sum_j \hat{Z}_{i,j}[\log(\theta_j) + \log f_j(y_i)],$$

where $\hat{Z}_{i,j} = P_{\theta_m}[Z_{i,j} = 1|y_i] \propto \theta_{m,j} f_j(y_i)$ and further, the M-step gives $\theta_{m+1,j} = \sum_{i=1}^n \hat{Z}_{i,j}/n$.
Iterating from some initial value yields a sequence with non-decreasing $l(\theta)$.
Notice the implied mapping $\theta_{m+1} = \psi(\theta_m)$, and also recall that at the MLE, $\hat{\theta} = \psi(\hat{\theta})$, and so the MLE solves the 'self-consistency' equation:

$$P_{\hat{\theta}}[Z_{i,j} = 1] = \hat{\theta}_j = mean_i[P_{\hat{\theta}}(Z_{i,j} = 1|y_i)]$$

i.e. the estimated marginal rate of class j is the average of posterior probabilities of this class.
From the fitted model, we can also cluster the data points according to the class they are most likely to be derived from:

$$cluster(i) = argmax_j \hat{Z}_{i,j}$$

This is the basic scheme behind the R package 'Mclust', which relies on multivariate Gaussian components on data living in Euclidean space. It's really an instance of what we can call 'Case

II', where the component densities depend themselves on some other parameters; in Mclust, they describe means and covariance matrices of the individual components. When working with parametrized components it's important to be careful about things such as identifiably [usually, constraints on the parameter space help to fix this] [e..g with univariate normals]. We note, without proof, that statistical properties improve with shared parameters among components, but computational problems increase.
Notice the boundary between clusters may be a curve, and can lead to unusual behaviour [e.g. two univariate normals with different means and different variances] [I didn't talk about the unbounded likelihood, but should have mentioned it here.]
Also, Mclust uses B.I.C. to guess K, and has some nice features for finding simpler models.
We go on to consider a special case involving Gaussian components, with a common covariance matrix $\sigma^2 I_p$ for all components, but with different means. We sketch in 2-d and then 1-d, to get at the nature of the clustering. We suppose that the current parameters put data point y_i closest to μ_1, say, and we work out for the E step

$$\hat{Z}_{i,1} = 1/\left\{1 + \sum_{j\neq 1}[\theta_j/\theta_1]f_j(y_i)/f_1(y_i)\right\}$$

Simplifying and taking limit as $\sigma^2 \to 0$, we see $\hat{Z}_{i,1} \to 1$; i.e. the E-step gives points i probability 1 to be assigned to the group that they are closest to. The new M-step would be an average of the so-classified points; this special limiting EM is exactly the same as the K-means algorithm for clustering.
We comment that many algorithms, especially ones that have good operating characteristics, are probably Bayes rules or ML rules under some probability model; the advantage of this perspective is that upon identifying the model, we can make improvements through improving the model rather than through riskier adjustments to the algorithm itself.
Next we go on to review for Wednesday's quiz.

Class 13

QUIZ!!!

Week 8

Class 14

On Generalized Linear Models
First, a brief discussion of four examples:

e.g.1 British Doctor's Study. In 1951, all registered physicians in the UK were contacted by letter and asked to participate in a prospective cohort study looking at the effects of smoking. More than 2/3 agreed, which amounted to more than 40,000 physicians. Among various information collected on the participants were smoking data and age, but also the investigators tapped records on death certificates and kept track of deaths associated with various causes. The study was the first large statistical study to show an association between lung cancer and smoking. We have summary data on the incidence of death by heart disease, for each age class and smoking group. The data amount to counts of deaths as well as counts of the number of 'person years' covered in both the smoking and non-smoking groups.

e.g.2 British Institute of Radiology. In the early 1990's data were combined from several large-scale clinical trials to do with the effectiveness of different radiotherapy protocols in the treatment of certain neck cancers. Data recorded properties of the cancer (e.g. stage) as well as characteristics of the radiotherapy (total dose, # fractions), and the response indicator of whether or not the therapy controlled tumor growth over a three year period. Interest is in how to fractionate a total dose most effectively.

e.g.3 AIDs diagnoses. Another British study, this time on rates of diagnosis of AIDs in a study from 1983 to 1992. Each quarter, investigators aimed to identify the number of new diagnoses in that quarter. The clinics where diagnoses occurred would report them to a public health authority, but the reporting was often subject to administrative delays. The basic data are counts per quarter of undelayed reports, those which were delayed by 1 quarter, by 2 quarters, and so on. There is missing data towards the end of the study, as delayed reports have not yet been counted. A goal of the study was to estimate the incidence of AIDS, accounting for this reporting delay.

e.g.4 Mutant frequencies, continued. We say a simplified version of this problem earlier this semester. T cells are isolated from blood cells sampled from a patient, and they are plated out and treated with 6TG to asses whether or not the deposited cells have any mutants at the HPRT locus. The binary data per well i is $Y_i = 1[fertile]$, which we modeled as Bernoulli with success probability $1 - e^{-n_i\lambda}$, for n_i = # cells per well and λ = mutation rate per cell. In the more realistic situation, λ itself depends on factors about the cells, such as whether or not the donor was treated with some medication, or some other factor (e.g. whether the cells were surrounding a tumor).

In all cases we have a count or binary response as well as covariate data: $\{Y_i, x_i\}$. A linear regression might seem reasonable, except that the response is not continuous. The GLM is a class of models used for such data.
GLM:

- Systematic component: $\mu_i = E(Y_i|x_i)$ is such that for some $g()$ $g(\mu_i) = x_i^T\theta$, for some regression parameters θ.
- We call $x_i^T\theta$ the linear predictor, and use notation η_i.
- Random component: $Y_i|x_i \sim$ exponential family; independent over i.

Suppressing i and x_i, e-family means $\log f_\theta(y) = \eta(\theta)t(y) - b(\theta) + c(y)$; we discuss this decomposition, and the importance of closure to independence sampling and the constancy of the dimension of the sufficient statistics.
Examples for binomial, Poisson: the link function g could be canonical [e.g. logit for binary, log for Poisson] if matches the canonical parameterization of the expo family, but this isn't necessary [e.g clog log link for MF example].
log likelihood: $l(\theta) = \sum_i l_i(\theta)$ using independence over units ; typically $l(\theta)$ is differentiable in θ; if also the MLE occurs on the interior of the parameter space, then it satisfies $l'(\hat{\theta}) = S(\hat{\theta}) = 0$, where S is the score function (vector of first derivatives of $l(\theta)$).

Root finding
Various root finding schemes may be considered to solve $S(\theta) = 0$. These are iterative algorithms producing a sequence $\theta_1, \theta_2, ..., \theta_B$, with some starting and stopping rules, and an update rule. Focusing on the updates, we could do

1. Scaling: for $\alpha > 0$, $\theta_{t+1} = \theta_t + \alpha S(\theta_t)$ [sketch].

2. Newton's method [also called Newton-Raphson]... $\theta_{t+1} = \theta_t + [-S'(\theta_t)]^{-1} S(\theta_t)$, where S' is the matrix of second partials of $l(\theta)$, the observed information matrix.

3. Could also replace $-S'(\theta_t)$ by the Fisher information $I(\theta_t)$; this is called Fisher scoring.

4. We could use several iterates to approximate the S'; secant method 1-d, quasi-Newton generally [sketch].

The R function 'optim' uses Nelder mead as default and a quasi-newton as a second option.
These methods work well if you have a good start, but they can fail badly if you don't [sketch].
In GLM's Newton's method simplifies to a series of regression problems, via the 'iteratively reweighed least squares algorithm'.

IRLS
Use $l = \sum_i l_i$. In differentiating in theta we use a chain rule and first differentiate in η_i, the i-th linear predictor, then differentiate η_i in θ; owing to the linearity of η_i in θ there are some nice simplifications.
First use $D = (x_{i,j})$ for the $n \times p$ design matrix, and let $u = n \times 1$ vector holding $\frac{dl_i}{d\eta_i}$.
Show 1, via chain rule, that $S(\theta) = D^T u$.
By a second derivative, show that $[-S'] = D^T W D$ where W is diagonal with negative $\frac{d^2 l_i}{d\eta_i^2}$ on its diagonal.
Then simplify the Newton update to: $D^T W D \theta_{t+1} = D^T W \tilde{y}$, where $\tilde{y} = D^T \theta_t + W^{-1} u$, and so the Newton iterates are obtained through a series of weighted least squares solutions.

Class 15

Generalized Estimating Equations (GEE)
Consider a data structure that seems to be amenable to GLM analysis: sketch a vector Y of responses and a matrix X of covariates, e.g. each entry of Y a binary; but suppose the data are structured into blocks, with Y_i and X_i for block i, corresponding to m, say different measurements on the same sampling unit, say for $i = 1, 2, ..., n$. We may want to relate predictors with responses via logistic regression, but we know that measurements on the same block may be correlated. E.g., repeated measures or longitudinal measures on a subject [we think there may be independence among blocks].
GEE is a way to use the GLM to develop an estimating equation for parameters relating predictors and response, and also to still accommodate dependencies.
To explain, let's go back to square one on information. For a single r.v. Y with density $f_\theta(y)$, the log likelihood $l(\theta) = \log f_\theta(y)$ and the score is $S(\theta) = \frac{d}{d\theta} l(\theta)$. In regular models it's easy to show that $E[S(\theta)] = 0$, and also that $var(S) = E[-S'(\theta)]$. The Fisher Information $I(\theta)$ is defined to be the variance of the score, and it usually equals the expected negative 2nd derivative (matrix). The next 'square' , suppose $Y_1, Y_2, ..., Y_n$ are i.i.d. from f_θ, and do a Taylor expansion of the the score $S(\theta)$ around the true parameter θ_0, evaluated at MLE $\hat{\theta}$ to get

$$(\hat{\theta} - \theta_0) \overset{app}{\sim} [-S'(\theta_0)]^{-1} S(\theta_0) \tag{1}$$

It often makes sense to scale S by $1/\sqrt{n}$, thus making it approximately $Gaussian[0, I(\theta_0)]$, and to scale $-S'(\theta_0)$ by $(1/n)$, to make it approximately $I(\theta_0)$ by the LLN. Recalling that for any

random vector Z with covariance matrix Σ, and any fixed matrix A (conforming), the covariance matrix of AZ is $A\Sigma A^T$. Thus looking at (1), (with the $\sqrt{n}$ scaling), we have

$$var(Z) \simeq i^{-1} var(S) i^{-1},$$

where $i = S'$ is the observed information and $var(S)$ is the variance of the score. Note that $var(S)$ is sandwiched in between two common factors (the bread!).
In the context of the model, $var(S)$ = information, and so the $var(Z) \simeq i^(- 1)$, as we usually consider it. So tests and confidence intervals may be developed using approximate normal theory with i^{-1} as the covariance matrix of the MLE. This is how we also operate for GLMs, noting that when IRLS converges, the limiting $D^T W D$ is the required i^{-1}.
Now GEE is developed in the context of some dependence. To see how, take one more small step, this time using binary Y_i, $i = 1, 2, ..., n$, independent and with covariates x_i, according to a GLM, so $\mu_i = E(Y_i | x_I)$ is such that for some link (e.g. logit or clog log) $g(\mu_i) = x_i^T \theta$, the linear predictor. We used a chain rule in the linear predictor to develop IRLS; for GEE we instead use a chain rule in μ_i.

$$S(\theta) = \frac{d}{d\theta} l(\theta) = \sum_i \left(\frac{dl_i}{d\mu_i} \right) \left(\frac{d\mu_i}{d\theta} \right),$$

for the Bernoulli Y_i, find that $\left(\frac{dl_i}{d\mu_i} \right) = V_i^{-1} [y_i - mu_i]$ where $V_i = \mu_i (1 - \mu_i)$ is the variance of Y_i [as in GLM's , it's a function of the μ_i].
So for univariate, independent Y_i we can write

$$S(\theta) = \sum_i \left(\frac{d\mu_i}{d\theta} \right) V_i^{-1} (y_i - \mu_i) \tag{2}$$

To make the connection to GEE and the more elaborate data structure introduced at the beginning of class, now use this same score equation but think of it with y_i the vector of binaries on block i (possibly dependent), μ_i the vector of means from a standard GLM, and V_i the diagonal matrix from the GLM, as if within block variables were independent, and $d\mu_i / d\theta$ the model function determined by the link and the linear predictors. I.e., had you the more elaborate data structure and you derived an ordinary GLM ignoring the dependence, you would get (2) as the score function.
Now the MLE $\hat{\theta}$ has variance, following (1) as $i^{-1} var(S) i^{-1}$. The GEE trick is to replace $var(S)$ by something other than i^{-1}, something that accounts for dependence among entries of each y_i. Applying 'var' to (2) we see

$$var(S[\theta]) = \sum_i \left(\frac{d\mu_i}{d\theta} \right) V_i^{-1} var(Y_i) V_i^{-1} \left(\frac{d\mu_i}{d\theta} \right)^T,$$

[a $p \times p$ matrix, assuming p covariates].
Of course we do not know the variance covariance matrix of Y_i, and we expect possible dependencies. A rank-1 quantity $[Y_i - \mu_i][Y_i - \mu_i]^T$ is $p \times p$ and has expected value $var(Y_i)$, so one idea is to say

$$var[S] \simeq \sum_i \left(\frac{d\mu_i}{d\theta} \right) V_i^{-1} (y_i - \mu_i)(y_i - \mu_i)^T V_i^{-1} \left(\frac{d\mu_i}{d\theta} \right)^T \stackrel{def}{=} M \text{ (for 'meat')}$$

Then the $var[\hat{\theta}] \simeq i^{-1} M i^{-1}$. This is the so-called 'sandwich' estimator of the the variance of the statistic $\hat{\theta}$ which is the root of (2) in the nested data structure.

There are other elements of the theory. On one hand some descriptions do not start with a probability model, but instead start with an equation, like (2), that an estimator ought to satisfy... a so-called estimating equation. One may tinker variously with elements of the estimating equation, for instance by replacing V_i with a so-called 'working covariance matrix' that may guide the anticipated dependence structure. Large sample theory for GEE's addresses properties such as regularity conditions needed for the estimates to be consistent and for the variance estimates to be accurate. Often they are, and the methods are popular in applied statistics, though they can have poor finite sample properties.
Other approaches are through explicit probability models [e..g random/mixed effects, GLMs], which we'll try to study later this semester. Doing something to handle dependence is important, since confidence intervals and tests will be invalid (too liberal) if we act as if the variables are independent, especially if there is some positive association.
Other facts about GLMs: Use of calculus to study scores and information and to obtain MLEs or solve estimating equations works fine when the log-likelihood is differentiable and when the MLE occurs in the interior of a parameter space, since then the score will be zero. But often the MLE occurs on the boundary of the space, and then neither is the score zero nor does the information guide the construction of confidence sets/tests. E.g. logistic regression, 1 - d or 2-d, with perfect split of 0's and 1's... leads to some θ_j's going off to infinity. I say the MLE still exists, it's just at the boundary of the parameter space.
One more binary-regression style data structure: Type-I interval censoring.
Random variables T_i are i.i.d. from a distribution F and associated with them are i.i.d (and independent from T's) $X_i \sim$ i.i.d. G. The data are pairs (Y_i, X_i), but where $Y_i = 1[T_i <= X_i]$... we don't get to see T_i's. More about this next time, in an example from stem cell biology...

Class 16

Examples involving constrained optimization:

Ordered binomials and current-status data: Last time we introduced the type-I interval censoring data structure (Y_i,X_i) where X_i is observed and $Y_i = 1[\ T_i \ ¡= X_i\ ]$ is also observed, but the event time of interest is not. We summarized the stem-cell aberration data, where T_i is time to occurrence of specific DNA damage in a stem cell sample i, X_i is the passage number of the sample. Of primary scientific interest is to compare the rate of damage between iPS and HES cells, but part of this is the problem to estimate the distribution F of T_i. Conditioning on x_i, (wlog are distinct, and ordered), we have

$Y_i \sim$ Bernoulli($\theta_i = \mathrm{F}(x_i)$); and since F is a cdf the θ_i's must be non-decreasing.

One option is to model F parametrically; e.g. $\mathrm{F(t)} = \exp(\ \alpha + \beta\ \mathrm{t}\ )/[\ 1 + \exp(\alpha + \beta\ \mathrm{t})\ ]$, which would be the same as logistic regression on the Y_i. An alternative approach is to attempt to estimate theta non parametrically. We would be aiming to maximize

$$L(\theta_1, ..., \theta_n) = \prod_i \theta_i^{y_i}(1-\theta_i)^{n_i - y_i} \tag{3}$$

subject to ordering constraint $\theta_1 <= \theta_2 <= ... <= \theta_n$ [here n_i's are all 1, but the binomial case is analogous]

This is a classical problem, whose solution was first presented by Ayer et al 1955, Ann Math Statist. Problems like this fit into the realm of optimization of linear (log L) functions subject to linear inequality constraints, and modern optimization theory, via KKT conditions, shows us how to find the solution. In the present case the solution has some interesting statistical

elements, so we present it without an attempt to justify it via KKT.

First, the solution in fact has a formula:

$$\hat{\theta}_i = max_{s<=i} min_{t>=i} A(s,t)/B(s,t) \tag{4}$$

where A(s,t) = $\sum_{j=s}^{t} y_j$ and B(s,t) = $\sum_{j=s}^{t} n_j$
This is not so illuminating, but we make headway by considering the cumulative sum diagram, which plots A(1,i] vs B(1,i) for all i in a scatterplot, also including the origin (0,0).
Next we consider

h(t) = sup $\phi(t)$: ϕ is convex and $\phi(t) <= CSD(t)$, the greatest convex minorant [sketch for data $n_i = 1$, y_i = (0,1,0,0,1,1,1)]

We claim that theta.hat equals the left derivative of h, say $h^{left(i)}$, in the case with all n_i=1.

Next we present the pool adjacent violators algorithm, applied to the example data, and we sketch and show how intermediate proportion sequences so produced create CSD's that move the original CSD towards a fixed point, the GCM h. Indeed, if the proportions start out non-decreasing then the PAVA has no effect, since there are no violating pairs.

A similar development holds for isotonic regression, to be investigated in homework. The R package 'iso' is a good one for these computations.

Comment: The NPMLE in the interval censoring case has the interesting property (in an iid model, with non-fixed x_i's) that it converges at cube root n rate

Empirical likelihood: (Owen)

Aside 1: Profile likelihood: Often θ=(μ,ψ), where μ is of interest and ψ is a nuisance, dealing with such ψ parameters is important, and distinguishes different approaches to inference [e.g. Bayesian inference involves integrating uncertainty in ψ]. One approach is to create profile likelihood

$l_p(\mu) = sup_\psi$ l(μ,ψ) , which requires optimization at each value of μ, over the space of ψ's

Aside 2: Likelihood ratios: We've discussed a lot about the likelihood for estimation and computing approximate standard errors (through information), but we have spoken little about likelihood ratios. Start with the simple case of i.i.d. data and a one-dimensional parameter theta, then expand the log-likelihood l(θ) in a 2 term Taylor series about the MLE $\hat{\theta}$ to get, after evaluating at the truth and ignoring errors,

2[l($\hat{\theta}$) - l(θ_0)] .approx. $[Normal(0,1)]^2$ (owing to a cancellation in information quantities)

i.e. the LR is approximately χ^2. S Wilks (1938) proved this as an instance of a more general result that

2[$sup_{\theta \in M1}$ l(θ) - $sup_{\theta \in M0}$ l(θ)] $\sim$ Distribution $\chi^2_{df=p-q}$

where M1 is a p dimensional model, and M0 is a q dimensional sub model (under certain regularity conditions). This supports the chi-squared goodness of fit test that is often used to assess whether or not a small model sufficiently well explains variation in data, and it's related to deviance in GLMs.

Aside 3: On likelihood-based confidence intervals. So far we've only considered normal-based approximate confidence intervals based on point estimate +/- 2 SE. We can also get likelihood based CI's by

CI = θ: l(θ) >= c for some well chosen c. In order for the CI to have a targeted nominal coverage, it's necessary to know (at least approximately) the sampling distribution of the likelihood ratio,

which Wilks assures is chi-square in nice parametric models.

Empirical likelihood: Consider data X_1, X_2,...$X_n \sim$ iid F in some Euclidean space, and our target of inference is $\mu = \mathrm{E}(X_i)$. Rather than model F parametrically, Art Owen proposed the following nonparametric likelihood

$$L(F) = \prod_{i=1}^{n} w_i \tag{5}$$

for an F that puts mass w_i on x_i [ignore ties in the x_i's for now].

Not so interestingly, argmax L(F) is the empirical distribution, which puts equal mass 1/n at each x_i. It gets interesting, however, with the empirical likelihood for the parameter of interest:

EL(μ) = $sup_{w:E(X)=\mu}$ L(F) for same F's defined by w_i

That is, we consider the profile likelihood for μ, maximizing over discrete distributions F having $\sum_i w_i x_i = \mu$. This constrained optimization may be solved by Lagrange multipliers, and creates a wonderfully balanced [between smooth and nonparametric] likelihood surface over the parameter space. What is especially amazing is that

2(log EL(sample mean) - log EL(mu0)) $\sim$ Distribution to χ^2_1, thus generalizing Wilks's famous result to a highly nonparametric case.

Class 17

Note on class project: consider selecting a recent paper from AoS, AoAS, Biometrika, Biostatistics, JRSSb, JASA, JCGS, Biometrics and identifying a statistical method addressing a specific inference problem using a specific data structure and working with a specific computational approach

Finishing up from last time, recall the duality between tests and confidence intervals. If for data X you can test a simple null hypothesis H_0: $\theta = \theta_0$ via the rejection region $R_\alpha(\theta_0) =$ X: reject H0 if we see this X ,

such that $P_{\theta_0}[\ \mathrm{X} \in R_\alpha(\theta_0)\ ] = \alpha$ [type-I error rate], then this test can be inverted to construct the confidence set $C_\alpha(X) = \theta$: X not in $R_\alpha(\theta)$. Clearly $P_\theta[\theta \in C_\alpha(X)] = 1 - \alpha$, and thus we have a valid CI.

In the LR context, the test might be based on T = 2[$\mathrm{l}(\hat{\theta}) - l(\theta_0)$] which is approximately chi-square distributed under various situations (e.g. Wilks) , and so the CI is the set of theta that are within a certain distance of the ML. Solving this out, with p=1,

and 95%, the cutoff for T is at 3.84, which amounts to about 2 units of log likelihood drop for an approximate 95% interval. See also Empirical likelihood.

Next we're moving on to motivate Bayesian inference; we'll do this with a series of short stories:

1. On nuisance parameters: Suppose $\theta = (\mu, \psi)$, a parameter of interest mu and a nuisance parameter ψ. And suppose for the MLE $\hat{\theta}$ we use the Fisher information to approximate it's inverse covariance. ... show that the variance of $\hat{\mu}$ is

$var(\hat{\mu}) = 1/[I_{1,1} - I^2_{1,2}/I_{2,2}] >= 1/I_{1,1}$, which would be the variance had we known the nuisance parameter psi; i.e., not knowing psi may inflate our uncertainty about μ [as long as $I_{1,2} \neq 0$; may equal 0, e.g. in regression with orthogonal covariates.]

2. On many nuisance parameters: Recall the Neymann Scott problem. X_i, Y_i pairs are independent and normal variates all with variance σ^2, but within pair each X_i, Y_i have mean ψ_i. The parameter of interest is σ^2. Write down the likelihood, argue

that $\hat{\psi} = (X_i + Y_i)/2$, show the profile log likelihood, obtain $\hat{\sigma}^2 = (1/4)(1/n)\sum_{i=1}^{n}(y_i - x_i)^2$, which is inconsistent for σ^2, since it converges to $\sigma^2/2$. The large number of nuisance parameters has messed up the otherwise nice properties (consistency/efficiency) of the MLE

3. On basic decision theory: data X, parameter theta, actions a, decision rules δ mapping data to actions; loss function L(θ,a)>=0, a value you lose if you take action a when θ is the state of nature;

Risk R$(\theta, \delta) = E_\theta[L(\theta, \delta(X))]$ the expected loss of using 'method' δ when theta is the state of nature; e.g. estimation, squared error loss, Risk of X in normal data is constant; Risk equals mean squared error in that case

Admissibility: δ is admissible if there is no other rule δ^* such that $R(\theta, \delta^*) <= R(\theta, \delta)$ at all θ, with strict inequality at some θ. Otherwise δ is inadmissible.

Why use an inadmissible method? Note the difficulty in comparing some [e.g. a shrinkage estimator vs a constant risk estimator for estimating a mean parameter]

An interesting thing about Bayes rules (to be defined), is that they are admissible under weak conditions.

Discuss the Stein phenomena: if X $\sim$ Normal[θ, 1], then $\hat{theta} =$ X is a great estimator; e.g. under L(a,θ) $=(a-\theta)^2$, it is constant risk, and it's unbiased, and minimum variance amongst the unbiased estimators, as well as being MLE

If the pair (X_1, X_2) is jointly normal, independent components, with means (θ_1, θ_2) and variances 1, then similarly (X_1, X_2) is an ideal estimator of (θ_1, θ_2).

Now consider$X = (X_1, X_2, X_3)$ similarly independent normals with means $\theta = (\theta_1, \theta_2, \theta_3)$; also unbiased, ML, minimum variance amongst the unbiased estimators, but under loss$L(a, \theta) = \sum_j (a_j - \theta_j)^2$, Stein showed it is inadmissible.

wow!

Later he and a student James constructed an estimator with uniformly smaller risk than X, which turned out to be something called an 'empirical Bayes' estimator, which we'll talk about more later, but which has the form of a Bayes estimator, but where we estimate

prior elements using data. More on that later.

Class 18

More on Bayesian inference

Exchangeability: I hand out a bunch of tacks and discuss an experiment: each person will toss his/her tack once, getting data $X_i \in$ 0,1

Now the outcome $(X_1, X_2, ..., X_n)$ is obviously unknown to us before we proceed, and the Bayesian analysts job is in part to specify a set of probability masses

$p(x_1, ..., x_n) = Pr[X_1 = x_1, ..., X_n = x_n]$ for all2^n possible outcomes.

Remember, in Bayesian analysis probability is not an intrinsic property of some experiment but it's some expression of uncertainty about the outcome; it turns out that the standard i.i.d. assumption is not primary, but something else is:

In simplifying the prob assignments, we might be willing to assume $p(x_1, ..., x_n) = p(x_{\pi_1}, ..., x_{\pi_n})$ for any permutation π of 1,2...,n, as an expression of the idea that the labels may be irrelevant in assigning probabilities.

We further say that an infinite sequence of binary variables is exchangeble if every finite sequence is. Then Bruno de Finetti proved that

$X_1, X_2, ...$ is exchangeable if and only if there exists $\theta \sim$ F such that $X_1..., X_n|\theta \sim$ iid Bernoulliθ.

And further that the average $\bar{X_n}$ converges a.s. to θ as n diverges.

This basic result about probability distributions anchors the connection between Bayesian and non-Bayesian approaches; it provides a new object 'F', called the prior, which represents uncertainty we have in the long-run relative frequency of success.

As an example, we consider the Polya learning model for binary sequence $X_1, X_2, ...$, showing that it gives an exchangeable sequence, and showing that it's de Finetti measure is the Beta distribution. [we didn't show, but could that the Beta mixture of Binomials gives rise to the Polya]

The whole kit and caboodle:

The basic elements of a Bayesian analysis include the data X, the parameter θ, the likelihood $p(x|\theta)$, but now also the prior f(θ), the posterior p(θ—x) and the prior-predictive p(x); example in a discrete case, flat prior, where the posterior is proportional to the likelihood.

On inference, let's revisit decision theory, with losses, risk, and so on.

A procedure $\delta = \delta$(x) is a Bayes rule if it minimizes Bayes risk, which is $E[R(\theta, \delta)]$, the average taken against the prior f, which may be analyst dependent.

Writing this out as a double average we see that finding the Bayes rule is possible by minimizing posterior expected loss for each data set [we also comment on admissibility]

We take some examples: $L(\theta, a) = (\theta - a)^2$, which leads to the posterior mean as the Bayes estimate of location

$L(\theta, a) = 1[\theta \neq a]$, 0-1 loss for a discrete parameter, and $L(\theta, a) = \sum_j 1[\theta_j \neq a_j]$, the Hamming loss for a vector discrete parameter

We show that for the 0-1 loss that the Bayes rule is the parameter value which maximizes the posterior (over the parameter space).

We introduce a toy example with $\theta = (\theta_1, \theta_2, \theta_3)$, each component binary, with a posterior on the 2^3 lattice

p1 on (1,1,1)

p2 on each of (0,0,0), (0,0,1), (0,1,0), and (1,0,0), and zero mass elsewhere,

and under the condition that $p2 < p1 < 2{*}p2$, which, with p1 + 4p2 = 1, gives p2 in (1/6,1/5). The Bayes estimator under Hamming loss involves the marginal posterior distribution of each θ_j, and gives

$\hat{theta_bayes}$ = (0,0,0), even though the joint MAP estimate is (1,1,1). This is the first example to show how marginalization enters into many Bayesian computations, but it also is relevant to examples where there

is substantial uncertainty in the parameter value.

Class 19, Bayesian inference continued:

e.g. $X_1, X_2, ..., X_m \sim$ iid Binomial[N, ϕ], where θ = (N,ϕ) are both unknown,

a basic model related to schemes used in ecology and elsewhere (e.g. genomics).

L(θ) = $\prod_{i=1}^{m} choose(N, x_i)\psi_i^x(1-\psi)^{1-x_i}$

which has banana-like contours typically, show for $N >= max(x_i)$

Bayesian inference starts with a prior $p(N, \phi)$ ... various are possible

then posterior $p(N, \phi|data)$, but then, importantly, if N is parameter of interest, we

integrate p(N|data) = $\int_0^1 p(N, \phi|data)d\phi$, and it's the marginal posterior distribution that gives us various Bayes rules (e.g. point estimate$E(N|data)$or $median(N|data)$, or CI's). Bayesian machinery may work even if the prior is improper, e.g. p(N,ϕ) $\propto$ 1/N, which, after integrating gives

$$p(N|data) \propto (1/N) \prod_i choose(N, x_i)(mN - s)!/(mN + 1)!1[N >= max(x_i)] \qquad (6)$$

where s = $\sum_i x_i$....sketch

See Raftery, 1989, Biometrika for extensions and discussion.

e.g. if m=1, we get $p(N|x) = x/[N(N+1)]$, which happens to have median 2x ; related to the example of being dropped in a city knowing nothing, seeing a bus labeled x, and wondering

how many buses are in the city.

A point for us is just that Bayesian rules come through marginal posterior computations.

On confidence intervals:

With a posterior distribution p(theta—x) one can define a highest posterior density (HPD) region by $\theta : p(\theta|x) >= c$, for c chosen such that P[$\theta \in HPD(x)|x]$ = pre-set like 0.95 [bayesian coverage] [sometimes called credible intervals or sets]. Note the distinction from regular confidence sets and frequency coverage.

Since HPD(X) is a procedure for making a set, one can ask about its frequency coverage

$P[\theta \in HPD(X)|\theta]$, which in general is not 0.95, the preset value, but which does have a certain average property. In fact, if theta's arise from the same prior as used in the construction of HPD(X), then

$EP[\theta \in HPD(X)|\theta] = EP[\theta \in HPD(X)|X = x] = 0.95$,

so Bayes intervals have an average frequency coverage property. [see e.g. Rubin and Schenker, 1984, Applied Stat.]

On testing:

Consider two hypotheses, e.g. H and K, disjoint, viewed as subsets of a parameter space. The Bayes rule for selecting H or K based on data, under 0-1 loss, uses the posterior odds

$p[H|x]/p[K|x] = (p[H]/p[K])(p(x|H)/p(x|K))$

the first factor is the prior odds, and the next is the so-called 'Bayes factor'. It's like a likelihood ratio, but notice that parameters are integrated rather than optimized away.

e.g.$p(x|H) = \int_H p(x|\theta)p(\theta|H)d\theta$

e.g. think of two - d param space, and perhaps $H = \theta : \theta[1] = 0$

with K = H^c

Special case:

suppose data arise from i.i.d. sampling of n units, and log likelihood satisfies

$$l(\theta) = l(\hat{\theta}) - [n\lambda/2]||\theta - \hat{\theta}||^2 + somethingboundedinn \tag{7}$$

look at

$$logp(x|H) = log \int p(x|\theta)d\theta \tag{8}$$

ignoring $p(\theta|H)$, or at least treating it as flat.

Substituting into the integral, and recognizing the normal integrand

$$logp(x|H) = l(\hat{\theta}) - (p/2)log(n) + somethingboundedinn \tag{9}$$

The two terms on the right, times -2, equal BIC = -2 $l(\hat{\theta}) + plog(n)$

Finding models (hypotheses) with small BIC thus balances goodness of fit with model complexity, and is seen as an approximate posterior inference tool. This calculation was reported by Gideon Schwartz in 1978.

On other examples:

e.g. Lawrence/Liu, 1992. Data$X = (X_{i,j})$all independent and each one$X_{i,j}$in A,C,T,G. Parameters are $\sigma = (\sigma_1, ..., \sigma_m)$, m is the extent of i, [the number of sequences], each σ_i in 1,2,...,n-K+1, where n is the extent of j [length of each sequence] and K is a smallish (e.g. 7) length of a binding site. the sigma's tell us the position of binding sites in the data. There are a slew of probability parameters

$\theta = [\theta_0, \theta_1, ..., \theta_K]$, and each theta.k is a vector over 4 letters A,C,T,G,

The model is

$P[X_{i,j} = c] = \theta_{k_c}$: k=0 if j is not in a binding site, and otherwise k = position in binding site.

This gives a simple likelihood. The task is to compute marginal posterior inferences about positions of the binding elements σ_i given the data, and this requires some fancy integration, which we'll discuss after the break.

Class 20

On graphs in statistics....the node/edge kind:

A. For representing dependencies among variables

B. For organizing statistical computations

Recall a graph G = (V,E), for a vertex set V and an edge set E. We are concerned with graphs that are either directed and acyclic, for which edges go in one direction only between nodes, or undirected, and we use simple graphs with no self loops or multi-edges. The graph G encodes somehow dependencies among a collection of node-specific random variables $X = X_V = X_v : v \in V$, governed by some joint distribution (density/mass function) p(x).

Directed Acyclic Graphs (DAG). For any node $v \in V$, write $pa(v) = j : (j, v) \in E$, which may be empty. A joint distribution is represented by the graph if

$$p(x) = \prod_v p(x_v | x_{pa(v)}) \tag{10}$$

which also means that X_j is conditionally independent of non-descendants given parents. We discuss some examples [introducing plate notation]

X_i mutually independent [no edges]

X_i conditionally independent given theta, θ random [like in de Finetti]

Markov chain, Hidden Markov chain,

Regression; logistic regression; the model via Richardson et al. 2002 involving logistic regression with some observable predictors, some missing predictors that are measured with noise, a mixture model on the missing predictors involving a discrete mixture identifier, an unknown number of components, and parameters for the component distribution, the mixing distribution, the covariate noise, and the regression.

We note that the DAG is not unique; e.g. U,V $\sim$ iid N(0,1), then Y = Normal[U+V, 1]....two different DAGs at least

DAGs often useful in expressing a model in generative form, using some intrinsic natural ordering of variables.

Undirected graphical representations:

Now edges are undirected. There are two key perspectives.

1. On local characteristics, or full conditional distributions:

A collection X = X_v is a Markov random field (MRF) relative to a graph G if $p(x_v|x_{-v}) = p(x_v|x_n bv)$, where nb(v) are the neighbors of v in the graph

This is the local Markov property, and implies that X_j is conditionally independent of X_v given $X_{nb(v)}$, for nodes j not in the neighborhood of v.

Questions naturally arise about whether/how a set of putative local characteristics can determine a joint distribution. It should be clear that given a

joint distribution, all of the local characteristics are determined, but it's not clear if the argument can be reversed. First note two examples:

e.g. 1: for two RV's X_1, X_2both in -1,+1, with the following local characteristics:

$P(X_1 = 1|x_2)$ — x_2

1/2 -1

1/2 +1

P($X_2 = 1|x_1$) — x_1
———————
1/2 -1
1/4 +1
This does not correspond to a joint distribution because
X_2
-1 +1
———————
-1 — O = O —
X_1 —-———————
+1 — O ¿ O —
———————

Conclusion: The collection of local characteristics must satisfy a number of internal constraints.

But even if the putative local characteristics correspond to some joint distribution, they may not determine that distribution.

Example 2: Again, let X_1, X_2, X_3 be binary, taking values +/-,

and consider two joint distributions:

Distribution 1 Distribution 2

X_3=+

X_2 X_2

- + - +

——————— ———————-

- — 1/4 0 — - — 1/8 0 —

X_1 ——————— ———————-

+ — 1/4 0 — + — 1/8 0 —

——————— ———————-

X_3=-

X_2 X_2

- + - +

——————— ———————-

- — 0 1/4 — - — 0 3/8 —

X_1 ——————— ———————-

+ — 0 1/4 — + — 0 3/8 —

——————— ———————-

These distributions have identical local characteristics. That is because

conditioning on two other variables fixes a row or column or level. You

may work throught the cases individually.

We ask, when do local characteristics determine a joint. E.g. imagine that an oracle has a joint distribution p(x) for some collection X, and

she gives to you the collection of local characteristics $p(x_v|x_{-v})$ for all variables. Can you tell what p(x) must be? We work through Besag's

illustration, involving two points a=(a1,a2,...,an) and b=(b1,b2,...,bn), showing that, under a certain condition

$$p(a)/p(b) = \prod_{i=n}^{1} p(a_i|a_1, ..., a_{i-1}, b_{i+1}, ..., b_n)/p(b_i|a_1, ..., a_{i-1}, b_{i+1}, ..., b_n) \tag{11}$$

and so, together with the unity sum constraint it must be that joint probabilities are constructible for full conditional probabilities. We've required that

denominators not vanish in this argument, which happens under the 'positivity condition', whereby any events that have positive marginal probabilities have

positive joint probability.

2. MRF's and local characteristics have to do with conditional probability, but an undirected graph may also encode a joint probability in other ways.

We say a distribution is Gibbs wrt G if $p(x) = \prod_{C incliques} h_C[x_C]$

a clique C is either a singleton node or a set of nodes that are all connected by edges in G [i.e. a complete subgraph]

The Hammersly Clifford Theorem says that X is MRF wrt G iff X is Gibbs wrt G, under positivity.

This result is a general statement about how complex can be a joint distribution determined by full conditionals , and has been useful both for computation

and for a general understanding of the nature of dependencies of multiple random variables. More next time.

Class 21

Graphs continued:

First: please send in project proposals to me by Friday.

Aside on numbers: a(b+c) takes either one sum and one product or two products and one sum;

given a_i and b_i, j , $S = \sum_i \sum_j a_i b_{i,j}$ takes either m^2 products then m^2 - 1 sums, or, using $S = \sum_i a_i \sum_j b_{i,j}$ m(m-1) sums , m products and another m-1 sums, where m is the extent of each index

given $a_i, b_{i,j}, c_{j,k}$, then $S = \sum_i \sum_j \sum_k a_i b_{i,j} c_{j,k}$ takes naively about m^3 operations, but can be done in O(m^2) by pushing sums, and the problem continues for more indices.

Back to graphs, we had discussed both undirected and directed representations [quick review]; note a MVN random vector is associated with a graph having edges where inverse covariance values are non zero [prove by pacman]; recall the logistic regression DAG, a generative expression of joint distribution of unknowns.

Aside from undirected graphs being useful for some model specifications, they're also central objects in inference from DAG - based models. There are three main approaches

1. exact comptuations
2. Monte Carlo
3. Variational

On exact calculations, they start by converting the DAG to an undirected representation. This may be tricky, but it's easy in one case, when no node v has more than one parent pa(v). In this simple case

notice that (using pacman) that the conditional distribution of x_v givenx_{-v} depends on any parents of v and also any children of v, and so a valid undirected graph is obtained by simply converting every directed edge in G to an undirected edge.

Next we puzzle about posterior computation on a fixed node v, say given all the data D. We do this through a message passing scheme.

1. view v as the root of a tree, and consider directions along every edge toward the root

2. on any edge e = i $->$ j, let D_e denote all the data upstream of that edge, possibly empty; this is information we will condition upon

3. let $m_e(x_i) = P[D_e|x_i]$ denote the message to be sent along edge e, recording the probability of all upstream data given the value of the state at the receiving node

4. We'll observe how to compute these messages shortly, but first note that by Bayes rule and the conditional independence of different data entries given the root, $P[X_v = x_v|D]$ may be computed from

all the messages entering the root [at least assuming that the marginal probability of the root state is at hand]

5. Observe that at any internal edge $e = i- > l$for which we also have multiple edges $e = j - - > i$entering i, that the message $m_e(x_l)$ is computable from the messages entering i

$$m_{i->l}(x_l) = P[D_{i->l}|x_l] = \sum_{x_i} P[D_{i--->l}, x_i|x_l] = \sum_{x_i}[\prod_{j--->i} P[D_{j--->i}|x_i]]p(x_i|x_l) \qquad (12)$$

here using conditional independencies induced in the graph.

and so these internal probabilities are yet other messages entering node i. Furthermore the conditional probability $p(x_i|x_l)$ is available from he original DAG. [assuming marginals $p(x_i)$ are also available]

Thus inference probabilities may be derived from messages computed along a flow through the undirected graph, towards the parameter of interest node. It happens that we can work all nodes at once, recognizing that

messages will be passed in two directions over each node [as in the forward-backward Baum-Welch scheme]. This general scheme is called the sum-product algorithm.

When nodes in the DAG have more than one parent, more is required to get an undirected representation. More about that in homework.

Class 22

Monte Carlo (1).

It's about doing computations by making the computer realize the experiment $X_1, X_2, ...$, [vary somehow] distribution P, and then representing the target quantity, say $\mu = E_P[g(X)]$, so that empirical averages

approximate the target quantity. 'vary somehow' is usually i.i.d., then we have the law of large numbers for justification, and the CLT for error. There are other schemes, including Markov chain sampling (that we'll discuss), which eliminate the independence part, and quasi-MC, where points are more regular than random.

Smiley face: LLN

frown face: CLT....error goes down very slowly

We comment on importance sampling, where in the X_i's come from a different distribution, where you tilt things so that you may average a different object to get the same target mu, but where you may have reduced variance.

Pseudo-random numbers: Ideal random numbers exist only mathematically; we know of no constructive definition. So we instead use deterministic schemes that produce sequences that have the same relevant statistical

properties as truly random sequences (discuss period and distribution tests, and that Mersenne Twister, R's default has a long period and passes these tests well beyond 600 dimensions; though we can onlyhope that more complex experiments are not being fooled by existing pseudo-random generators [maybe if target quantity is a complex high-dim summary?]

Statistical uses:

- simulation to study sampling properties of a method [e.g. bias, power, coverage]
- methods for data analysis

* diagnostics

* p-values, confidence sets,

* anything involving conditional probability [e.g. Bayesian inference]

Uniform[0,1]: We discuss that all Monte Carlo realizations can come from uniform[0,1] realizations, and note three major techniques for 'non-uniform variate generation'

1. transformation: e.g U $\sim$ Unif[0,1], log(1/U) $\sim$ Exp(1); more generally $F^{-1}(U) \sim F$

2. rejection sampling: we go through the argument in steps

a. if $X_1, X_2, ... \sim p$, and A is a set, then N=min n: $X_i \in$ A is a geometric RV, and $Y = X_N$ has the distribution of X_i given X_i in A

b. What it means to be uniform over an open set in Euclidean space ; P[X $\in$ A] pronto volume[A]

c. One example is { (x,y): x real and 0 <= y <= $h(x)$ } for an envelope function h that is non-negative and integrable, and from this the marginal density of X is proportional to the envelope

d. Starting with a target density p(x) and an envelope $h(x) >= p(x)$, make pairs uniform in set under p by first making point uniform under h and retaining those in the first set, by a,b, then project.

3. composition: p(x) = sum or integral $\pi_s p(x|s) ds$; think of DAG representation; e.g. $S \sim \chi^2_v$, $X|S \sim Normal[0, v/S]$, the marginally $X \sim t_v$

Example: N(0,1)

1. transformation via qnorm or Box-Muller

2. rejection via Marsaglia's polar method

3. composition via Marsaglia-Bray

CLass 23

Monte Carlo, continued.

We note that midterm 2 will be on Wed April 29, hw 5 will be due Mon April 27, and the following week (May 4/6) will be student presentations:

First a few notes on multivariate simulation:

Methods

1. DAG...if one exists, simulate variables sequentially following the DAG, using whatever marginal or conditional distributions are encoded [e.g. HMM]

2. Specials: e.g. Multivariate normal via iid normal + linear (Cholesky) transformation..

3. Other: often harder, more next week [e.g. where one has an undirected representation]

Bootstrap methods:

Data D $\sim$ P, but that's unknown, estimate it from observed data somehow, say via P_{boot}.

A statistic T(D) has some induced 'sampling distribution' Q, depending on P and T, which we don't know, but which if we had we could use to advantage for testing or CI construction.

Bootstrap methods are plug-in methods, involving the use of Q_{boot}, the distribution induced by P_{boot} and T. They have nothing intrinsically to do with MC, but since they are

complicated, they are most easily approximated by sampling $D* \sim P_{boot}$, repeatedly, and averaging: e.g. $Q_{boot}(A) \stackrel{app}{\sim} (1/B) \sum_{b=1}^{B} 1[T(D*_b) \in A]$

Such samples $D*_b$ are called bootstrap samples; typically they have the same structure as the original data, but are generated via sampling from P_{boot} (either model based estimate or an empirical distribution

in case D is composed of a random sample, and P_{boot} is the empirical)

One application of this is for CI construction, say for a real-valued parameter theta, with an estimator $\hat{\theta} = \hat{\theta}(D)$. Q is the sampling distribution of the estimator (e.g. encoding its bias an MSE), and Q_{boot}

is the bootstrap distribution of $\hat{\theta}(D*)$. The standard deviation of bootstrap estimates equals the bootstrap estimate of the standard error of $\hat{\theta}$, and so that quantity provides for one simple CI

$\hat{\theta}$ +/- 2 $\hat{SE}$ [a. $\hat{SE}$ by information argument or some asymptotic argument; b $\hat{SE}$ by bootstrap]

Another is the bootstrap percentile method, funding L and U s.t. $Q_{boot}[L, U] = 0.95$ later we offer some justification of this

We go on to discuss the Pearson correlation example, and then an example on nonparametric estimation of the aberration probability distribution arising in the stem cell example.

We finish with Efron's argument if $m(\hat{\theta}) - m(\theta) \sim Normal[0, c^2]$for some monotone increasing m and some constant c, regardless of P....