For $x=5$ $\quad \ell'(p) = \frac{5}{p} > 0$ and $\ell(p)$ is increasing

So for $\Theta = [0,1]$ the MLE is $\frac{X}{5}$

One context in which the likelihood equations are especially nice is the exponential family — for $X_1, X_2, \ldots, X_n$

$$f(x|\eta) = c(\eta)^n \prod_{i=1}^n h(x_i) \exp \sum_{j=1}^k \eta_j \left(\sum_{i=1}^n T_j(x_i) \right)$$

B+D from $A(\eta) = -\log c(\eta)$ $\Leftrightarrow$ $\exp(-nA(\eta)) \prod_{i=1}^n h(x_i) \exp \sum_{j=1}^k \eta_j \left(\sum_{i=1}^n T_j(x_i) \right)$

so $\quad \ell(\eta) = -nA(\eta) + \sum_{i=1}^n \ln h(x_i) + \sum_{j=1}^k \eta_j \left(\sum_{i=1}^n T_j(x_i) \right)$

and $\quad \frac{\partial}{\partial \eta_j} \ell(\eta) = -\frac{\partial}{\partial \eta_j} A(\eta) + \sum_{i=1}^n T_j(x_i)$

By Corollary 1.6.1 B+D, if $\mathcal{E}$ has nonempty interior

(or claim 9 on Exponential Family handout)

$$\frac{\partial}{\partial \eta_j} A(\eta) \left(= \frac{\partial}{\partial \eta_j} (-\log c(\eta)) \right)$$
$$\| $$
$$E_\eta T_j(X)$$

Thus, the jth of these equations is

$$-n E_\eta T_j(X) + \sum_{i=1}^n T_j(x_i) = 0$$

$$E_\eta T_j(X) = \frac{1}{n} \sum_{i=1}^n T_j(x_i) \quad \Big) \text{ jth likelihood equation}$$

So, the ML estimating equation in an exponential family with nonempty interior in $\mathcal{E}$ is "set the theoretical mean of T equal to the empirical mean of T" — a natural exponential family is "MOM"

The question of whether there must be a solution to the exponential family likelihood equations and if any solution must be unique and maximize the likelihood is discussed in B+D section 2.3 — one simple result from there is

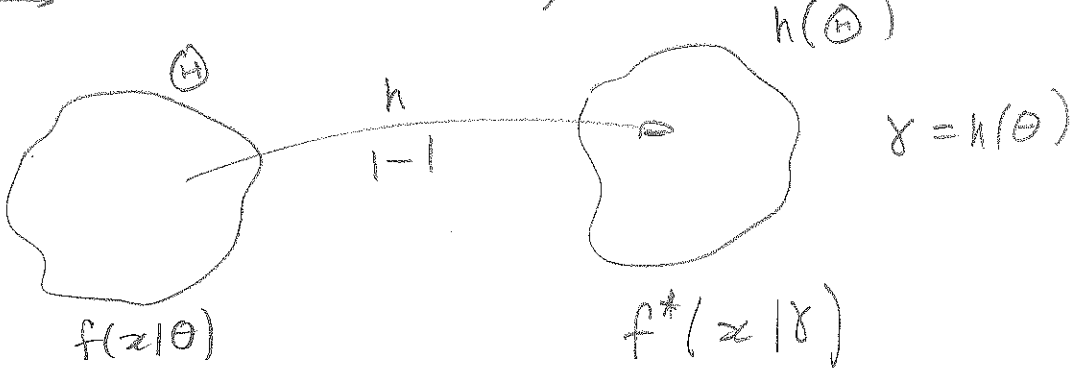
Corollary 2.3.2 If the equations

$$E_\eta T_j(X) = \frac{1}{n}\sum_{i=1}^{n} T_j(x_i) \qquad j = 1, 2, \ldots, k$$

have a solution η in the interior of $\mathcal{E}$ it is the unique MLE of η

Both for application in exponential families where the "natural" parameterization isn't the "standard" or "usual" parameterization and in other contexts, there is the lemma

Lemma (Problem 2.2.16a)



clearly $\hat{\theta}$ maximizes $f(x|\theta)$ over Θ

$\Rightarrow \hat{\gamma} = h(\hat{\theta})$ maximizes $f^*(x|\gamma)$ over $h(\Theta)$

So, e.g., $\hat{\eta}$ MLE of natural parameter if θ can be written as $h(\eta) = \theta$ for a 1-1 function h, Then

$$\hat{\theta} = h(\hat{\eta}) \quad \text{is the MLE of } \theta$$

Solving the likelihood equations

$$\nabla l(\theta) = 0$$

- is "just" a numerical analysis problem — various iterative procedures (bisection, Newton-Raphson, etc.) might be applied to do the job — mostly, their discussion is not a topic for STAT 3 —

one procedure for the problem has some probabilistic/statistical content and can/should be touched here is the "EM algorithm" —

day 17

Basic idea: Sometimes an observable Y has a nasty log likelihood $l_Y(\theta)$ but could be thought of as distributionally equivalent $\forall \theta$ to $S(X)$ for some (potentially completely fictitious) X for which computation and optimization of

$$E_{\theta_0}\left[l_X(\theta) \mid S(X) = y \right]$$

data in hand

is feasible — In such cases I might

0. pick some starting value $\theta^{(0)}$

1. Find

E step $E_{\theta^{(0)}}[\ell_X(\theta) \mid S(X) = y]$

M step 2. optimize this as a function of θ to find $\theta^{(1)}$

3. Replace $\theta^{(0)}$ with $\theta^{(1)}$

4. iterate to convergence

Example $X \sim \exp(\lambda)$ ~~λ~~ $\frac{1}{\text{mean}}$

$$Y = X I[1 < X < 2] + 1 I[X \le 1] + 2 I[2 \le X]$$

(Y a censored version of X) — Suppose $Y_1, Y_2, \ldots, Y_n$ iid with this dsn

$$f(y|\lambda) = \begin{cases} 1 - e^{-\lambda} & y = 1 \\ \lambda e^{-\lambda y} & 1 < y < 2 \\ e^{-2\lambda} & y = 2 \end{cases}$$

If $Y_i = 1$ then X_i has density $\propto \lambda e^{-\lambda x}$ on $(0,1]$ and if $Y_i = 2$ X_i has density $\propto \lambda e^{-\lambda x}$ on $[2, \infty)$

$$\ell_X(\lambda) = n \log \lambda - \lambda \Sigma X_i$$

Then for a particular λ_0 conditional on $Y = y$

$\sum X_i$ has mean

$$\sum_{y_i \in (1,2)} y_i + \#[y_i=1]\, E_{\lambda_0}[X \mid X<1] + \#[y_i=2]\, E_{\lambda_0}[X \mid X>2]$$

$$E_{\lambda_0}[X \mid X<1] = \frac{1}{1-e^{-\lambda_0}} \int_0^1 x \lambda_0 e^{-x\lambda_0} dx$$

$$= \frac{1}{1-e^{-\lambda_0}} \left[-x e^{-\lambda_0 x} \Big|_0^1 + \int_0^1 e^{-x\lambda_0} dx \right]$$

$$= \frac{1}{1-e^{-\lambda_0}} \left[-e^{-\lambda_0} + \frac{1}{\lambda_0}(1-e^{-\lambda_0}) \right]$$

$$= \frac{1}{\lambda_0} - \frac{e^{-\lambda_0}}{(1-e^{-\lambda_0})}$$

$$E_{\lambda_0}[X \mid X>2] = \frac{1}{e^{-2\lambda_0}} \int_2^\infty x \lambda_0 e^{-x\lambda_0} dx$$

$$= e^{2\lambda_0} \left[-x e^{-\lambda_0 x} \Big|_2^\infty + \int_2^\infty e^{-x\lambda_0} dx \right]$$

$$= e^{2\lambda_0} \left[2e^{-2\lambda_0} + \frac{1}{\lambda_0} e^{-2\lambda_0} \right]$$

$$= 2 + \frac{1}{\lambda_0}$$

i.e. the λ_0 conditional mean of $\ell_X(\lambda)$ given $Y=y$ is

$$n \log\lambda - \lambda \left[\sum_{y \in (1,2)} y_i + \#[y_i=1]\left(\frac{1}{\lambda_0} - \frac{e^{-\lambda_0}}{1-e^{-\lambda_0}}\right) + \#[y_i=2]\left(2+\frac{1}{\lambda_0}\right) \right]$$

E-step

which is maximized at $\lambda^{(j+1)} \leftarrow \frac{n}{[\text{above}]}$

M-step

So ultimately, an iterative E-M algorithm sets

$$\lambda^{(j+1)} \leftarrow \frac{n}{\left[\sum_{y_i \in (1,2)} y_i + \#[y_i = 1]\left(\frac{1}{\lambda^{(j)}} - \frac{e^{-\lambda^{(j)}}}{1 - e^{-\lambda^{(j)}}}\right) + \#[y_i = 2]\left(2 + \frac{1}{\lambda^{(j)}}\right)\right]}$$

and iterates to convergence as opposed to straight-up optimization of

$$\ell_Y(\lambda) = \#[y_i = 1]\log(1 - e^{-\lambda}) + \#[1 < y_i < 2]\log\lambda - \lambda \sum_{y_i \in (1,2)} y_i + \#[y_i = 2]\, 2\lambda$$

day 18

Not Done in Class

Example k possible outcomes, n independent identical trials

p_j = probability that any trial produces outcome j

$$X_{ij} = I[\text{trial } i \text{ produces outcome } j]$$

$$n_j = \sum_{i=1}^{n} X_{ij} = \#\text{ of trials producing outcome } j$$

Suppose that p has each $p_i \geq 0$ w $\sum p_i = 1$

pmf for X is

$$f(x|p) = \prod_{j=1}^{k} p_j^{n_j}$$

(for x with $x_{ij} = 0$ or 1, $\sum_j x_{ij} = 1$, $n_j \geq 0$ integers w/ $\sum n_j = n$)

MLE of p is (as it turns out)

$$\hat{p} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_k}{n}\right)$$

But now, what if, e.g. $k = 4$ and all we observe is this

Trial	Information Available	In Terms of X's
	outcome is	
1	1	$X_{11} = 1$
2	3	$X_{23} = 1$
3	2 or 4	$X_{32} + X_{34} = 1$
4	2	$X_{42} = 1$
5	3	$X_{53} = 1$
6	2 or 3	$X_{62} + X_{63} = 1$
7	1	$X_{71} = 1$
8	1 or 2	$X_{81} + X_{82} = 1$
9	2	$X_{92} = 1$
10	4	$X_{104} = 1$

That is, I don't get the 10 $X_i = (X_{i1}, X_{i2}, \ldots, X_{i4})$ but, e.g.

$$Y_3 = (X_{31}, X_{33}, X_{32} + X_{34})$$

The likelihood function based on Y_i's (information available) not X_i's is

$$L_Y(p) = p_1^2 p_2^2 p_3^2 (1 - p_1 - p_2 - p_3) \times (1 - p_1 - p_3)(p_2 + p_3)(p_1 + p_2)$$

($1 - p_1 - p_3 = p_2 + p_4$)

and I could probably optimize $L_Y(p)$ or $l_Y(p) = \log L_Y(p)$ via some numerical method

Another possibility is to use the EM algorithm

$$l_X(p) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 (1 - p_1 - p_2 - p_3)$$

For the data in hand this is

$$= (2 + X_{31}) \log p_1 + (2 + X_{32} + X_{62} + X_{52}) \log p_2 + (2 + X_{63}) \log p_3 + (1 + X_{34}) \log (1 - p_1 - p_2 - p_3)$$

For any particular $p_0 = (p_{01}, p_{02}, p_{03}, p_{04})$

$$E_{p_0}[X_{31} \mid Y = \text{data in hand}] = \frac{p_{01}}{p_{01} + p_{02}}$$

$$E_{p_0}[X_{32} \mid Y = \text{data in hand}] = \frac{p_{02}}{p_{02} + p_{04}}$$

$$E_{p_0}[X_{62} \mid \quad] = \frac{p_{02}}{p_{02} + p_{03}}$$

$$E_{P_0}\left[X_{82} \mid\right] = \frac{P_{02}}{P_{01}+P_{02}}$$

$$E_{P_0}\left[X_{63} \mid\right] = \frac{P_{03}}{P_{02}+P_{03}}$$

$$E_{P_0}\left[X_{34} \mid\right] = \frac{P_{04}}{P_{02}+P_{04}}$$

Then

$$E_{P_0}\left[l_X(p) \mid Y = \text{data in hand}\right]$$

$$= \underbrace{\left(2 + \frac{P_{01}}{P_{01}+P_{02}}\right)}_{a(P_0)} \log P_1 + \underbrace{\left(2 + P_{02}\left(\frac{1}{P_{02}+P_{04}} + \frac{1}{P_{02}+P_{03}} + \frac{1}{P_{01}+P_{02}}\right)\right)}_{b(P_0)} \log P_2$$

$$+ \underbrace{\left(2 + \frac{P_{03}}{P_{02}+P_{03}}\right)}_{c(P_0)} \log P_3 + \underbrace{\left(1 + \frac{P_{04}}{P_{02}+P_{04}}\right)}_{d(P_0)} \log\left(1 - P_1 - P_2 - P_3\right)$$

Note

$$a(P_0) + b(P_0) + c(P_0) + d(P_0) = 10 = n$$

$$P^{(j+1)} = \left(\frac{a(P^{(j)})}{10}, \frac{b(P^{(j)})}{10}, \frac{c(P^{(j)})}{10}, \frac{d(P^{(j)})}{10}\right)$$

iterate to a fixed point

B+D give some arguments as to why EM might work — see, in particular, Lemma 2.4.1 That says that at each step $L_Y(\theta)$ never decreases — the standard complaint w/ EM is that its convergence is often very slow — you need a good starting point and if other methods are possible They may be faster —

One matter that should be raised regarding EM is that B+D phrase their version of it concerns not

$$E_{\theta_0}[\ell_X(\theta) \mid Y=y]$$

but rather

$$E_{\theta_0}[\ell_X(\theta) - \ell_X(\theta_0) \mid Y=y]$$

sometimes I'll be able to compute the 2nd when I couldn't compute the 1st — and optimization of the first is equivalent to optimization of the 2nd

B+D give some arguments why EM might work — see, in particular, Lemma 2.9.1 That says that at each step $\ell_Y(\theta)$ never decreases — the standard complaint about EM is that its convergence is often very slow — you need a good starting point and if other methods are possible, they may be faster —

Bayes Estimators

we've said repeatedly that when $g(\theta)$ specifies a prior dsn

$$g(\theta|x) \propto L_x(\theta) g(\theta)$$

and that characteristics of the postain serve as estimators of $\gamma(\theta)$ — e.g.

under SEL, the Bayes estimator of $\gamma(\theta)$ is

$$\delta(x) = E[\gamma(\theta) \mid X = x]$$

under AEL, the Bayes estimator of $\gamma(\theta)$ is

$$\delta(x) = \text{median of the dsn of } \gamma(\theta) \mid X = x$$

under WEL with weight function $w(\theta) \geq 0$, the Bayes estimator of $\gamma(\theta)$ is

$$\delta(x) = \frac{E[\gamma(\theta) w(\theta) \mid X = x]}{E[w(\theta) \mid X = x]}$$

In fact, this story can sometimes be generalized by not requiring $g(\theta)$ to specify a probability dsn

Example $X \sim N(\theta, 1)$ $\quad g(\theta) = 1$ $\quad \int_{-\infty}^{\infty} 1 \, d\theta = \infty$?!!

$$L_x(\theta) g(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}(x-\theta)^2\right) \cdot 1$$

and we can use $L_x(\theta) g(\theta)$ to specify a "posterior" $(N(x, 1))$ and ∴ "generalized Bayes" estimator $\delta(x) = x$ even though $f(x \mid \theta) g(\theta)$ doesn't define a (joint) probability dsn for (X, θ) — note that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}(x-\theta)^2\right) dx \, d\theta$$

$$= \int_{-\infty}^{\infty} 1 \, d\theta = \infty$$

What is (Bayes) optimal is in principle "clear" The problem of actually computing characteristics of a dsn with density proportional to

$$L(\theta)g(\theta)$$

Note that even finding an exact density for the posterior requires finding

$$\int L(\theta)g(\theta)\,d\theta$$

the normalizer for the function $L(\theta)g(\theta)$

Modern Bayes computation is based on simulation. To substitute for calculus — e.g. if I am interested in

$$Q = \int q(\theta)g(\theta|x)\,d\theta$$

for some $q: \mathbb{R}^k \to \mathbb{R}^1$ and can somehow generate $\theta_1^*, \theta_2^*, \ldots$ iid $G(\theta|x)$ I might approximate Q by

$$\hat{Q}_n = \frac{1}{n}\sum q(\theta_i^*)$$

relying on the LLN to conclude that

$$\frac{1}{n}\sum_{i=1}^{n} q(\theta_i^*) \xrightarrow{P} Q$$

($G(\theta|x)$ probability)

(e.g. I might for $\theta=(\theta_1,\theta_2)$ have $q(\theta)=\theta_1$ or $q(\theta)=I[\theta_1<\theta_2]$)

The problem with this idea is that naive ways of doing simulation would require that I know $\int L(\theta)g(\theta)d\theta$ in order to simulate the θ_i^* —

Happily there is the famous "rejection algorithm"