BIOS 7731 Homework 4 Tim Vigers

1.6.12) The multinomial distribution can be written:

$$p(\vec{x}|\vec{\theta}) = \frac{n!}{x_1! \dots x_K} \theta_1^{x_1} \theta_2^{x_2} \dots \theta_K^{x_K}, \quad 0 < \theta_j < 1, \quad 1 \leq j \leq K, \quad \sum_{i=1}^{K} \theta_i = 1$$

and be thought of as modeling the probability of counts in K "bins," with n independent "trials" leading to a count for exactly one of the K categories. First, show that the distribution can be written in exponential family (EF) form:

$$p(\vec{x}|\vec{\theta}) = h(x) \exp\left(x_1 \log(\theta_1) + \dots + x_K \log(\theta_K)\right)$$

where $h(x) = \frac{n!}{x_1! \dots x_K}$

This can be re-parameterized using the fact that $x_K = n - x_1 - x_2 \dots - x_{K-1}$ and $\theta_K = 1 - \theta_1 - \theta_2 \dots \theta_{K-1}$:

$$p(\vec{x}|\vec{\theta}) = h(x) \exp\left(\sum_{i=1}^{K-1} x_i \log(\theta_i) + \left(n - \sum_{i=1}^{K-1} x_i\right) \log\left(1 - \sum_{i=1}^{K-1} \theta_i\right)\right)$$

This simplifies to:

$$p(\vec{x}|\vec{\theta}) = h(x) \exp\left(\sum_{i=1}^{K-1} x_i \log(\theta_i) + n \log\left(1 - \sum_{i=1}^{K-1} \theta_i\right) - \left(\sum_{i=1}^{K-1} x_i\right) \log\left(1 - \sum_{i=1}^{K-1} \theta_i\right)\right)$$

$$= h(x) \exp\left(\sum_{i=1}^{K-1} x_i \left(\log(\theta_i) - \log\left(1 - \sum_{i=1}^{K-1} \theta_i\right)\right) + n \log\left(1 - \sum_{i=1}^{K-1} \theta_i\right)\right).$$

This is a K-1 parameter EF where:

$h(x) = \frac{n!}{x_1! \dots x_K!}$ $\quad \eta_j(\theta) = \log\left(\frac{\theta_j}{1 - \sum_{j=1}^{K-1} \theta_j}\right)$ and $T_j(x) = x_j$ for $1 \leq j \leq K-1$

$B(\theta) = -n \log\left(1 - \sum_{j=1}^{K-1} \theta_j\right)$ can be re-written in terms of $\vec{\eta}$

but it is unnecessary to prove the rank of an EF.

Using the definition of rank of an EF, we must show that:

$$P\left[\sum_{j=1}^{k-1} a_j T_j(x) = a_k\right] < 1 \quad \text{unless all } a_j = 0.$$

This can be shown using the distribution of X_1 conditioned on $X_2 = x_2 \ldots X_{k-1} = x_{k-1}$:

$$p(x_1, x_k \mid X_2 = x_2 \ldots X_{k-1} = x_{k-1}) = \frac{n_0!}{x_1!\, x_k!} \theta_1^{x_1} \theta_k^{x_k} \quad \text{with } n_0 = n - x_2 - x_3 \cdots x_{k-1}$$

If we consider "bin" k as an unknown number of "failures" and bin 1 a count of "successes," it follows that X_1 has a binomial distribution.

Therefore, at least one of the X_i is a random variable with density. As a result, the only way that

$$P\left[\sum_{j=1}^{k-1} a_j T_j(x) = a_k\right] = 1 \quad \text{is if all of the } a$$

are 0, which is not allowed under this definition of EF rank. In other words, because $\vec{T}(x)$ is a random variable there is no way to guarantee it equals some constant a. If the final "bin" k was observed, then the X_i's would no longer be independent, therefore the EF is not of minimal rank k. and is rank $k-1$.

1.6.18) Let $Y_i \sim N(\beta_1 + \beta_2 z_i, \sigma = 1)$ where $z_1 \ldots z_n$ are observed covariate values and not all equal. The density of $\vec{Y}$ is:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}(y_i - (\beta_1 + \beta_2 z_i))^2\right)$$

This simplifies to:

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \exp\left(\frac{-y_i^2}{2} + \frac{2y_i(\beta_1 + \beta_2 z_i)}{2} - \frac{(\beta_1 + \beta_2 z_i)^2}{2}\right)$$

Which can be written in canonical EF form:

$$h(y) \exp\left(\beta_1 \Sigma y_i + \beta_2 \Sigma y_i z_i - \frac{\Sigma(\beta_1 + \beta_2 z_i)^2}{2}\right) \quad \text{with } \Sigma = \sum_{i=1}^{n}$$

where $\eta_1 = \beta_1$ $\quad T_1(Y) = \sum Y_i$ $\quad$ and

$\eta_2 = \beta_2$ $\quad T_2(Y) = \sum Y_i z_i$

$$A(\eta) = \frac{\sum(\beta_1 + \beta_2 z_i)^2}{2} = \frac{\sum(\eta_1 + \eta_2 z_i)^2}{2}$$

First use $A(\eta)$ to find the mean vector and variance matrix of T:

$$A'(\eta) = \left[\frac{d}{d\eta_1} A(\eta), \ \frac{d}{d\eta_2} A(\eta) \right]$$

$$\frac{d}{d\eta_1} A(\eta) = \frac{2\sum(\eta_1 + \eta_2 z_i)(1)}{2} = \sum(\beta_1 + \beta_2 z_i)$$

$$\frac{d}{d\eta_2} A(\eta) = \frac{2\sum(\eta_1 + \eta_2 z_i) z_i}{2} = \sum z_i(\beta_1 + \beta_2 z_i)$$

$$A'(\eta) = \left[\sum_{i=1}^{n} \beta_1 + \beta_2 z_i, \ \sum_{i=1}^{n} z_i(\beta_1 + \beta_2 z_i) \right] = \text{mean } T(Y)$$

Next the variance matrix:

$$A''(\eta) = \begin{bmatrix} \frac{d^2}{d\eta_1^2} A(\eta) & \frac{d^2}{d\eta_1 d\eta_2} A(\eta) \\ \frac{d^2}{d\eta_1 d\eta_2} A(\eta) & \frac{d^2}{d\eta_2^2} A(\eta) \end{bmatrix}$$

$$\frac{d^2}{d\eta_1^2} A(\eta) = \frac{d}{d\eta_1} \sum(\eta_1 + \eta_2 z_i) = \sum_{i=1}^{n} 1 = n$$

$$\frac{d^2}{d\eta_2^2} A(\eta) = \frac{d}{d\eta_2} \sum z_i(\eta_1 + \eta_2 z_i) = \sum_{i=1}^{n} z_i^2$$

and $\frac{d^2}{d\eta_1 d\eta_2} A(\eta) = \sum_{i=1}^{n} z_i$

So the variance of $T(Y)$ is:

$$\begin{bmatrix} n & \sum z_i \\ \sum z_i & \sum z_i^2 \end{bmatrix}$$

The determinant of the variance matrix var(T) is:

$n \sum z_i^2 - \sum z_i \sum z_i$, which is always $>$ because $n \sum z_i^2 > \sum z_i \sum z_i$. I don't think this is enough to show that var(T) is positive definite, but it's a necessary part. Proving that Var(T) is positive definite is equivalent to saying the canonical EF $\mathcal{P}$ is rank 2 (in this case).

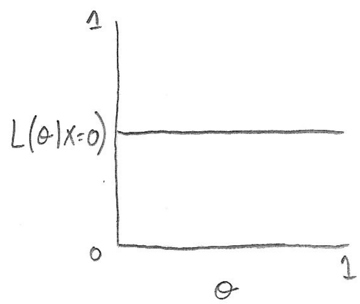
Another equivalent statement is that A is strictly convex or Ξ. This is clearly the case here, as Ξ is open, and the function $A(\eta) = \frac{\sum(B_1 + B_2 z_i)^2}{2}$ is a positive quadratic function.

Therefore, this is a canonical EF of rank 2 and the variance of T is positive definite.

3) For this discrete distribution, the likelihood is almost always a straight line increasing or decreasing as a function of θ, except in the case of $X=0$. So, we can make a table of the MLE $\hat{\theta}$ for each value of x:

X	-2	-1	0	1	2
$\hat{\theta}$	0	1	$\hat{\theta}^*$	0	1

Not only do both 0 and 1 maximize the likelihood given different values of x, the likelihood at $X=0$ is a constant:



So in the case of $X=0$ the likelihood is the same for all values of $\hat{\theta}$, and there is no unique maximizer.

1.6.27) First, define the canonical family generated by $\vec{T}$ and h_0:

$$q(x, \eta^*) = h_0(x) \exp(\eta^* T(x) - A^*(\eta^*))$$

Now replace $h_0(x)$ with $q(x, \eta_0)$:

$$q(x, \eta^*) = h(x) \exp(\eta_0 T(X) - A(\eta_0)) \exp(\eta^* T(x) - A^*(\eta^*))$$

which is equivalent to:

$$h(x) \exp((\eta_0 + \eta^*) T(x) - A^*(\eta^*) - A(\eta_0)) \qquad (1)$$

Next find $A^*(\eta^*)$:

$$A^*(\eta^*) = \log\left(\int h_0(x)\exp(\eta^* T(x))\,dx\right)$$ which expands to:

$$= \log\left(\int h(x)\exp(\eta_0 T(x) - A(\eta_0))\exp(\eta^* T(x))\,dx\right)$$

Grouping terms in the exponents produces:

$$A^*(\eta^*) = \log\left(\int h(x)\exp((\eta_0+\eta^*)T(x) - A(\eta_0))\,dx\right)$$

Because $A(\eta_0)$ is not a function of x, we can write:

$$A^*(\eta^*) = \log\left(\int h(x)\exp((\eta_0+\eta^*)T(x))\,dx\ \exp(-A(\eta_0))\right)$$

This is equivalent to:

$$A^*(\eta^*) = \log\left(\int h(x)\exp((\eta_0+\eta^*)T(x))\,dx\right) + \log(\exp(-A(\eta_0)))$$

The first term in this formula is the definition of $A(\eta_0+\eta^*)$, so

$$A^*(\eta^*) = A(\eta_0+\eta^*) - A(\eta_0)$$

Then we can plug this back into (1):

$$q(x,\eta^*) = h(x)\exp((\eta_0+\eta^*)T(x) - (A(\eta_0+\eta^*) - A(\eta_0)) - A(\eta_0))$$

The $A(\eta_0)$ terms cancel, leaving us with:

$$q(x,\eta^*) = h(x)\exp((\eta_0+\eta^*)T(x) - A(\eta_0+\eta^*)) = q(x,\eta^*+\eta_0)$$

From this it's clear that if $\eta^* = \eta - \eta_0$, then

$$q(x,\eta^*) = q(x,\eta) \quad \square.$$

2.3.1) Given $P(Y_i = 1) = p(x_i, \alpha, \beta)$ with $\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$

and $\theta = (\alpha, \beta)$, we can write the likelihood:

$L_x(\theta) = \prod_{i=1}^{n} p(x_i, \theta)^{Y_i}(1-p(x_i, \theta))^{1-Y_i}$. Taking the log

of this gives us:

$$\ell_x(\theta) = \sum_{i=1}^{n} Y_i \log(p(x_i, \theta)) + (1-Y_i)\log(1-p(x_i, \theta)).$$

This rearranges to:

$$\ell_x(\theta) = \sum_{i=1}^{n} Y_i \log(p(x_i, \theta)) + \log(1-p(x_i, \theta)) - Y_i \log(1-p(x_i, \theta))$$

$$= \sum_{i=1}^{n} Y_i \left(\log(p(x_i, \theta)) - \log(1-p(x_i, \theta))\right) + \log(1-p(x_i, \theta))$$

$$= \sum_{i=1}^{n} Y_i \log\left(\frac{p(x_i, \theta)}{1-p(x_i, \theta)}\right) + \sum_{i=1}^{n} \log(1-p(x_i, \theta))$$

Because $\log\left(\frac{p}{1-p}\right)(x_i, \theta) = \alpha + \beta x$, we can rewrite in

terms of α and β:

$$\log\left(\frac{p}{1-p}\right)(x_i, \theta) = \alpha + \beta x \quad \text{and}$$

$$\frac{p}{1-p} = \exp(\alpha + \beta x), \quad \text{so} \quad p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

$$1 - p = \frac{1}{1 + \exp(\alpha + \beta x)}.$$

Therefore, $\ell_x(\theta) = \sum_{i=1}^{n} Y_i(\alpha + \beta x_i) - \sum \log(1 + \exp(\alpha + \beta x_i))$.

This is a rank 2 EF with open parameter space $\mathcal{E}$ (I've run out of time to show this but the general approach would be similar to 1.6.12 with each Y_i the result of a Bernoulli trial).

This is ~~not quite~~ in canonical EF form, but it gives us the parts we need for proving the existence of MLEs. By Theorem 2.3.1 in B&D, the MLE $\hat{\eta}$ exists and is unique if:

$$P(c^T T(y) > c^T t_0) > 0 \quad \text{where } y \text{ is the observed data and } t_0 = T(y).$$

Based on the previous page, we have:

$$\eta_1 = \alpha \qquad T_1(Y) = \sum Y_i$$

$$\eta_2 = \beta \qquad T_2(Y) = \sum Y_i x_i$$

From the hint we know that:

$$c_1 \sum y_i + c_2 \sum y_i x_i = \sum_{i=1}^{n} (c_1 + c_2 x_i) y_i \le \sum_{i=1}^{n} (c_1 + c_2 x_i) 1(c_2 x_i + c_1 \ge 0)$$

and that the bound is sharp and only attained when $y_i = 0$ for $x_i \le \frac{-c_1}{c_2}$ and $y_i = 1$ for $x_i \ge -\frac{c_1}{c_2}$.

The right hand side of this inequality is essentially the maximum value possible for $c^T t_0$, so when the bound is attained $p(c^T T(Y) > c^T t_0) = 0$ and the MLE does not exist.

This bound is only possible to attain if the y_i are a sequence of 1s then 0s (or vice versa) when the X_i are ordered $x_1 < x_2 \dots < x_n$. If this is not the case, then $y_i \ne 0$ for all $x_i \le \frac{-c_1}{c_2}$ and $P(c^T T(Y) > c^T t_0) > 0$. If the y_i are perfectly separated (or vice versa) such that $y_i = 1$ for all $x_i \le c' \wedge$, then it is possible to choose c_1 and c_2 such that the bound is attained and $P(c^T T(Y) > c^T t_0) = 0$.

This is why logistic regression does not work with perfect separation between groups.

2.3.8. b) Let $X_1 \dots X_n$ be iid with density

$$f_\theta(x) = c(\alpha)\exp(-|x-\theta|^\alpha) \quad \theta \in R^p, \quad \alpha = 1, \quad x \in R^p$$

The likelihood for $\alpha = 1$ and $p = 1$ is:

$$L_x(\theta) = \prod_{i=1}^{n} c(\alpha)\exp\left(-|x_i-\theta|\right) = c(\alpha)^n \prod_{i=1}^{n} \exp\left(-|x_i-\theta|\right)$$

Taking the log of the likelihood gives us:

$$\ell_x(\theta) = n\log(c(\alpha)) - \sum |x_i - \theta|$$

The absolute value can be split in terms of the X_i's relation to θ, giving us:

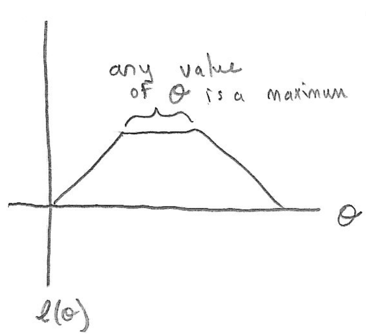
$$\ell_x(\theta) = n\log(c(\alpha)) - \left(\sum_{X_i > \theta} X_i - \theta + \sum_{X_i < \theta} \theta - X_i\right)$$

Taking the derivative of the log likelihood results in:

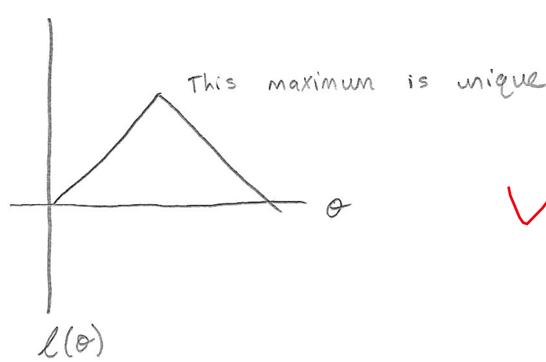
$$\ell'_x(\theta) = -\left(\sum_{X_i > \theta} -1 + \sum_{X_i < \theta} 1\right) = \sum_{X_i > \theta} 1 - \sum_{X_i < \theta_i}$$

Or, the number of $X_i > \theta$ minus the number of $X_i < \theta$. So, when n is an even number, $\ell'_x(\theta) = 0$ when the number of $X_i > \theta$ is equal to the number $X_i < \theta$. For example, if $n = 10$ with $X_{(5)} = 3$ and $X_{(6)} = 6$, then any value of $\hat{\theta} \in (3, 6)$ will maximize $\ell_x(\theta)$. If $n = 9$, then $\ell_x(\theta)$ will be maximized at $X_{(5)}$ (the median). I found it helpful to plot these two situations:

<u>n is even</u>



<u>n is odd</u>



1.6.27) An alternative approach if the previous is too circular:

Write the density:

$$q(x, \eta) = h(x) \exp\left(\eta T(x) - A(\eta)\right).$$

We can rewrite this with $\eta = \eta + \eta_0 - \eta_0$:

$$q(x,\eta) = h(x) \exp\left((\eta + \eta_0 - \eta_0) T(x) - A(\eta - \eta_0 + \eta_0)\right)$$

and rearrange:

$$q(x,\eta) = h(x) \exp\left(\eta_0 T(x) - A(\eta_0)\right) \exp\left((\eta - \eta_0) T(x) - A(\eta - \eta_0)\right)$$

Because $h_0(x) = q(x, \eta_0)$, this is equal to:

$$q(x,\eta) = h_0(x) \exp\left((\eta - \eta_0) T(x) - A(\eta - \eta_0)\right).$$

Therefore $q(x, \eta)$ is also canonical EF generated by $T(\cdot)$ and h_0 with $\eta_* = \eta - \eta_0$.