

# Recitation: maximum likelihood estimator for multinomial

## Recitation problem statement

Consider a finite space  $E = \{a_1, a_2, \dots, a_r\}$  of size  $r \geq 2$  and let  $X$  be a random variable taking values in  $E$ . For  $j = 1, \dots, r$ , let  $p_j^* = \mathbf{P}[X = a_j]$ . Consider a sample of  $n$  i.i.d. copies  $X_1, \dots, X_n$  of  $X$ . Based on this sample, we would like to estimate the multivariate parameter  $p^* = (p_1^*, \dots, p_r^*)$ .

1. What is the parameter space ?
2. Write the likelihood associated with the model described above.
3. Compute the maximum likelihood estimator  $\hat{p}$  of  $p^*$ .
4. Using the central limit theorem, show that  $\hat{p}$  is asymptotically normal. Compute the asymptotic covariance matrix. Denote it by  $\Sigma$ .
5. Prove that  $\Sigma$  is not invertible. Conclude that the theorem for the MLE could not have been applied here. What condition is not satisfied ?

Maximum Likelihood estimator for multinomial model

$E = \{1, 2, \dots, r\}$ . pmf  $p_j, j=1, \dots, r$ . Observe  $X_1, \dots, X_n$  iid.  $\mathbf{P}(X_i=j) = p_j, j=1, \dots, r$

- binomial is a case of multinomial where  $r=2$

$X_i$  must all equal one of the values of  $E$  (1 to  $r$ )

Parameter space  $\left\{ p_j \geq 0, \sum_{j=1}^r p_j = 1 \right\} = \mathcal{P}$

$$\textcircled{1} \quad \mathbf{P}(X_1=x_1, \dots, X_n=x_n) = \prod_{i=1}^n \mathbf{P}(X_i=x_i) = \prod_{i=1}^n \prod_{j=1}^r p_j^{\mathbb{1}\{x_i=j\}} = \prod_{j=1}^r \prod_{i=1}^n p_j^{\mathbb{1}\{x_i=j\}}$$

- indicator function  $\mathbb{1}\{x_i=j\}$  means that if  $x_i = j$  then  $p_j$  will remain, otherwise  $p_j = 1$  due to exponent

$$= \prod_{j=1}^r p_j^{\sum_{i=1}^n \mathbb{1}\{x_i=j\}} = \prod_{j=1}^r p_j^{T_j}$$

- $T_j$  means how many times have we observed each element  $j$
- in binomial this would be number of heads and tails in coin flip example

$$\log L = \log P(X_1=x_1, \dots, X_n=x_n) = \log \prod_{j=1}^r p_j T_j = \sum_{j=1}^r T_j \log p_j$$

- log of product becomes a sum and the  $T_j$  can be pulled outside the log
- 

Compute MLE

- take the likelihood given the parameters,
- plug in observations
- maximise wrt the parameters ( $p_j$  in this case)

Maximum likelihood estimator for multinomial model

$$\log L = \sum_{j=1}^r T_j \log p_j, \quad T_j = \sum_{i=1}^n \mathbb{1}\{X_i=j\}, \quad \mathcal{P} = \{p \in \mathbb{R}^r : p_j \geq 0 \ \forall j, \sum p_j = 1\}$$

$$= f(p)$$

① Calculating MLE:  $\max_p f(p)$

Assume  $T_j > 0 \ \forall j$

Necessary conditions:  $0 = \nabla f(\hat{p})$

$$\partial_{p_j} f(p) = \frac{T_j}{p_j} - r = 0$$

- $\hat{p}_j$  is the optimum we want to find
- if  $p_j$  is strictly positive then this derivative will never become 0, only approach it as  $p_j \rightarrow \infty$
- but they have to lie between 0 and 1, so it is a constrained optimisation problem, not unconstrained

$\hat{f}(\rho) = f(\rho)$

① Calculating MLE:  $\max_{\rho \in \mathcal{P}} f(\rho) \Leftrightarrow \max f(\rho)$  st.  $h(\rho) - \sum_{j=1}^r p_j - 1 = 0$

Assume  $T_j > 0 \forall j$

Necessary conditions:  $0 = \nabla f(\hat{\rho}) + \lambda \cdot \nabla h(\hat{\rho})$ ,  $\lambda \in \mathbb{R}$

$$\partial_{p_j} f(\rho) = \frac{T_j}{p_j} = 0, \quad p_j \rightarrow \infty \quad ???$$

$$\partial_{p_j} h(\rho) = 1 \Rightarrow 0 = \frac{T_j}{p_j} + \lambda$$

- add in additional  $h()$  function, which the derivative gives us lambda as an additional term
- using Lagrange multiplier
- there is only one lambda for all dimensions

$$\partial_{p_j} h(\rho) = 1 \Rightarrow 0 = \frac{T_j}{p_j} + \lambda \Rightarrow \lambda \neq 0 \Rightarrow \hat{p}_j = -\frac{T_j}{\lambda}$$

$$1 = \sum_{j=1}^r \hat{p}_j = \sum_{j=1}^r \left( -\frac{T_j}{\lambda} \right) = -\frac{1}{\lambda} \sum_{j=1}^r T_j = -\frac{n}{\lambda} \Rightarrow \lambda = -n$$

$$\Rightarrow \boxed{\hat{p}_j = \frac{T_j}{n}}$$

- this is just the frequency of each element
- this is assuming none of the  $T_j$  are 0

Maximum Likelihood estimator for multinomial model

$$f(\rho) \log L_h = \sum T_j \log p_j, \quad T_j = \sum_{i=1}^n \mathbb{1}\{X_i=j\}, \quad \mathcal{P} = \left\{ \rho \in \mathbb{R}^r : p_j \geq 0 \forall j, \sum p_j = 1 \right\} \text{ (Simplex)}$$

$$h(\rho) = \sum p_j - 1, \quad \partial_{p_j} f(\rho) = \frac{T_j}{p_j}, \quad \partial_{p_j} h(\rho) = 1$$

$$\textcircled{1} \quad T_j > 0 \Rightarrow \hat{p}_j = \frac{T_j}{n}. \quad \text{Global max?}$$

$$\partial_{p_k} \partial_{p_j} f(p) - \partial_{p_k} \frac{\partial_{p_j}}{p_j} = \begin{cases} -\frac{T_j}{p_j^2}, & j=k \\ 0, & j \neq k \end{cases} \Rightarrow \nabla^2 f(p) < 0 \Rightarrow f \text{ concave}$$

- the log likelihood is concave so we know that  $\hat{p}_j$  is a global maximum
- the Hessian is strictly concave

$$T_j = 0 ? \quad (\text{Karush-Kuhn-Tucker conditions})$$

$$P(X_1=x_1, \dots, X_n=x_n) = \prod_{j=1}^n p_j^{T_j}$$

- if  $T_j$  is 0 then it will not enter the product
- gives same result  $T_j / n$

Asymptotic variance

$$\text{Maximum Likelihood estimator for multinomial model}$$

$$\boxed{\hat{p}_j = \frac{T_j}{n}}, \quad P(X_i=j) = p_j, \quad \mathcal{P} = \{P : p_j \geq 0, \sum p_j = 1\}, \quad T_j = \sum_{i=1}^n \mathbb{1}\{X_i=j\}$$

$$\textcircled{2} \quad \hat{p}_j = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i=j\}, \quad E[(Y_i)_j] = E[\underbrace{\mathbb{1}\{X_i=j\}}_{\sim Be(p_j)}] = P(X_i=j) = p_j$$

$$\text{CLT: } \sqrt{n} \cdot (\hat{p} - E[Y_i]) \xrightarrow[n \rightarrow \infty]{D} \mathcal{N}(0, \text{Cov}(Y_i))$$

- can apply CLT as  $Y$  is an average of iid random variable and becomes Bernoulli with parameter  $p_j$

$$\Sigma_{jk} = \begin{cases} \text{Var}((Y_i)_j), & j=k \\ \text{Cov}((Y_i)_j, (Y_i)_k), & j \neq k \end{cases}$$

$$\text{Var}((Y_1)_j) = p_j(1-p_j)$$

$$E[(Y_1)_j(Y_1)_k] = E[1_{\{X_1=j\}} 1_{\{X_1=k\}}] = 0, \quad j \neq k$$

- 0 because one of the identity functions will always be 0 since  $j$  does not equal  $k$

$$\text{Cov}((Y_1)_j, (Y_1)_k) = E[(Y_1)_j(Y_1)_k] - E[(Y_1)_j]E[(Y_1)_k]$$

$$= 0 - p_j p_k, \quad j \neq k$$

covariance matrix

$$= \begin{cases} p_j(1-p_j), & j=k \\ -p_j p_k, & j \neq k \end{cases}$$

- $p_j(1-p_j)$  on the diagonal where  $j=k$
- negative on the off-diagonals
- negative shows an anti-correlation since when  $j$  is big  $k$  is small (similar to the identity functions)

Another method of calculating asymptotic variance

$$\textcircled{2} \quad \hat{\theta} \text{ MLE, } \log \ln f(\theta)$$

- think of theta as p here

$$\ln(\hat{\theta} - \theta^*) \xrightarrow[n \rightarrow \infty]{D} N(0, I(\theta^*)^{-1})$$

$$I(\theta) = -\mathbb{E}[\nabla^2 f(\theta)]$$

$$f(p) = \sum_{j=1}^r T_j \cdot \log p_j, \quad \partial_{p_j} f(p) = \frac{T_j}{p_j}$$

$$\partial_{p_j} \partial_{p_k} f(p) = \begin{cases} -\frac{T_j}{p_j^2}, & j=k \\ 0, & j \neq k \end{cases}$$

- the Hessian is a diagonal matrix as shown here

$$I(p)_{jk} = \mathbb{E}[(\nabla^2 f(p))_{jk}] = \begin{cases} \frac{p_j}{p_j^2} = \frac{1}{p_j}, & j=k \\ 0, & j \neq k \end{cases}$$

$$I(p)^{-1}_{jk} = \begin{cases} p_j, & j=k \\ 0, & j \neq k \end{cases}$$

- the inverse Fisher Information matrix does not look like the covariance matrix as calculated before (below)

$$\sum_{j,k} = \begin{cases} p_j(1-p_j), & j=k \\ -p_j p_k, & j \neq k \end{cases}$$

Explanation

some assumptions were not satisfied

1. Model identified ✓
2.  $\text{supp } \tilde{P}_\theta$  does depend on  $\theta$  ✓

get second assumption by assuming new  $p_{\tilde{\theta}}$  (below) that is strictly above 0

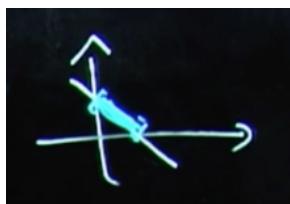
$$\tilde{\mathcal{P}} = \left\{ p : p_i \geq 0, \sum_{j=1}^r p_j = 1 \right\}$$

3.  $\theta^*$  does not lie on the boundary of  $\mathcal{H}$   $\times$

- bounded by linearity, this is a problem?

4.  $I(\theta)$  is invertible in a neighbourhood of  $\theta^*$   
 5. Technical asspts.

- technical assumptions is continuously differentiable



- since these assumptions aren't fulfilled then covariance is not equal to Fisher Information

Can get around this

$$\Sigma \neq I(p)^{-1}$$

$$\tilde{\mathcal{P}} = \left\{ p \in \mathbb{R}^r : p_j > 0, \underbrace{\sum_{j=1}^{r-1} p_j}_{p_r} < 1 \right\}$$

- parameterise the set differently with 1 less parameter

$$\tilde{I}(p)^{-1} = (\Sigma)_{\substack{j=1, \dots, r-1 \\ k=1, \dots, r-1}}$$