Lecture 7, Jan 29, 2024

Estimators for Multinomial RVs

- The multinomial distribution is a generalization of the binomial distribution
  - In binomial we had 2 outcomes 0 and 1, so $N_0 + N_1 = n$; in multinomial we have k outcomes, $N_1, \dots, N_K = n$
  - The probability of outcome k is θ_k and $\displaystyle\sum_{k=i}^{K} \theta_k = 1$
  - e.g. tossing a die
- The indicator function for multinomial is a k-tuple $\boldsymbol{X}$, with a 1 in the position that the outcome occurred and 0s everywhere else
  - e.g. $\boldsymbol{X} = (0, 0, 1, 0, \dots, 0)$ indicates outcome is 3
- The probability of $\boldsymbol{X}$ is then $P[\boldsymbol{X} = (b_1, \dots, b_K)] = \displaystyle\prod_{k=1}^{K} \theta_k^{b_k}$ where b_k is the number of occurrences of k
- Again consider n independent trials $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ and let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$
- $$\begin{aligned} P[\boldsymbol{X}_1 = \boldsymbol{b}_1, \boldsymbol{X}_2 = \boldsymbol{b}_2, \dots, \boldsymbol{X}_n = \boldsymbol{b}_n; \boldsymbol{\theta}] &= \prod_{j=1}^{n} P[\boldsymbol{X}_j = \boldsymbol{b}_j] \\ &= \prod_{j=1}^{n} \theta_1^{b_{j_1}} \dots \theta_K^{b_{j_K}} \\ &= \theta_1^{\sum b_{j_1}} \dots \theta_K^{\sum b_{j_K}} \\ &= \theta_1^{N_1} \dots \theta_K^{N_K} \end{aligned}$$
  - Where $N_k = \displaystyle\sum_j b_{j_k}$ is the number of times outcome k occurred in n trials
  - The vector $\boldsymbol{N} = (N_1, \dots, N_K)$ is a sufficient statistic for our estimators
- Note $E[\boldsymbol{N}; \boldsymbol{\theta}] = (E[N_1], \dots, E[N_K]) = (n\theta_1, n\theta_2, \dots, n\theta_K)$
  - The expected value of the $\boldsymbol{N}$ vector is simply the number of trials times the probability of each trial
- Consider the MLE estimator:
  - $\log P[\boldsymbol{N}; \boldsymbol{\theta}] = \log(\theta_1^{N_1} \dots \theta_K^{N_K}) = \displaystyle\sum_{k=1}^{K} N_k \log \theta_k$
  - Now we need to optimize this sum with respect to $\boldsymbol{\theta}$, with the constraint that all θ_k are positive the sum of all θ_k is 1
  - Lagrangian: $\displaystyle\sum_{k=1}^{K} N_k \log \theta_k + \lambda \left(\sum_{k=1}^{K} \theta_k - 1 \right)$
    - * For a particular term θ_j, the derivative is $\dfrac{N_j}{\theta_k} + \lambda = 0 \implies \dfrac{N_j}{\theta_j} = -\lambda$
    - * Substituting this into the constraint for θ we get $\lambda = -n$
  - Therefore $\hat{\theta}_{j_{\text{ML}}} = -\dfrac{N_j}{\lambda} = \dfrac{N_j}{n}$
    - * This is expected, since it's the relative frequency of k
- This is for a particular sequence of outcomes; if we only cared about number of occurrences, we have to add the multinomial coefficient
  - $\dbinom{n}{n_1, n_2, \dots, n_K} = \dfrac{n!}{n_1! n_2! \dots n_K!}$ where $n_1 + \dots + n_K = n$
  - For $K = 2$, this reduces to the binomial coefficient
- For the MAP estimate we use the Dirichlet prior, which is a generalization of the beta distribution
  - The Dirichlet distribution is $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \dfrac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \theta_1^{\alpha_1 - 1} \dots \theta_K^{\alpha_K - 1}$ where $\alpha_j > 0, \displaystyle\sum_j \alpha_j = \alpha_0$
    - * This is the conjugate prior for the multinomial distribution since it has the same form

- The posterior is $f(\boldsymbol{\Theta}|n_1, \dots, n_K) = \dfrac{p(n_1, \dots, n_K|\boldsymbol{\theta})f(\boldsymbol{\theta})}{p(n_1, \dots, n_K)}$
$$= c\theta_1^{n_1+\alpha_1-1} \dots \theta_K^{n_K+\alpha_K-1}$$
$$= \frac{\Gamma(\alpha_0 + n)}{\Gamma(\alpha_1 + n) \dots \Gamma(\alpha_K + n_K)} \frac{\prod_{k=1}^K \theta_k^{n_k+\alpha_k-1}}{P(n_1, \dots, n_K)}$$
- We again form the Lagrangian and take derivatives to obtain: $\dfrac{n_j + \alpha_j - 1}{\theta_j} = -\lambda, -\theta_j = \dfrac{n_j + \alpha_j - 1}{\lambda}$
- Therefore $\hat{\theta}_{j_{\text{MAP}}} = \dfrac{n_j + \alpha_j - 1}{n + \alpha_0 - K}$
  - The $-K$ in the denominator gets rid of all the extra 1s in the αs when summed up
  - We can interpret this as a relative frequency, where prior to doing the experiment we did $\alpha_0 - K$ experiments and outcome j occurred $\alpha_j - 1$ times

- Consider the LMS estimator:
  - $E[\boldsymbol{\Theta}|\boldsymbol{N}] = \displaystyle\int \dots \int (\theta_1, \dots, \theta_K) c\theta_1^{n_1+\alpha_1-1} \dots \theta_K^{n_K+\alpha_K-1} \,\mathrm{d}\theta_1 \dots \mathrm{d}\theta_k$
$$= (E[\Theta_1|n_1 + \alpha_1 - 1], \dots, E[\Theta_K|n_K + \alpha_K - 1]$$
$$= \left(\frac{n_1 + \alpha_1}{n + \alpha_0}, \dots, \frac{n_K + \alpha_K}{n + \alpha_0}\right)$$
    - Note $E[\Theta_j|n_j + \alpha_j - 1] = c\displaystyle\int_0^1 \theta_j \theta_j^{n_j+\alpha_j-1} \,\mathrm{d}\theta_j = \frac{n_j + \alpha_j}{n + \alpha_0}$
  - Therefore $\hat{\theta}_{j_{\text{LMS}}} = \dfrac{n_j + \alpha_j}{n + \alpha_0}$
- Again notice that as $n \to \infty$, all 3 of these estimators converge to the ML estimator

Binary Hypothesis Testing

- Hypothesis testing is like a more constrained version of parameter estimation; instead of estimating the value of θ, we are testing whether θ_0 or θ_1 is more likely
- Given two hypotheses H_0 (the *null hypothesis*, or the "default" to be proved or disproved) and H_1 (the *alternative hypothesis*), we want to know which one is more likely
- We would like to find $g\colon S_{\boldsymbol{X}} \mapsto \{ H_0, H_1 \}$ mapping from observations to hypotheses based on $P[\boldsymbol{X} \in A; H_j]$
  - g divides the sample space into 2 parts, the *acceptance region* R^c where H_0 is accepted and *rejection region* R where H_0 is rejected
- If g is not perfect, then 2 types of error can occur:
  - *Type I error*: H_0 is rejected when it is true
    - Also known as the *significance level* of a test
    - $\alpha(R) = P[\boldsymbol{X} \in R; H_0]$
    - We typically pick this to be 10%, 5%, 1%, etc
  - *Type II error*: H_0 is accepted when H_1 is true (i.e. H_0 is false)
    - $\beta(R) = P[\boldsymbol{X} \in R^c; H_1]$
- We can do this partitioning using our 3 estimators
- Using MLE, we simply pick the H that gives us the maximum likelihood
  - We just need to test $p_{\boldsymbol{X}}(\boldsymbol{x}|H_0)$ and $p_{\boldsymbol{X}}(\boldsymbol{x}|H_1)$
  - The *likelihood ratio* is $L(\boldsymbol{x}) = \dfrac{p_{\boldsymbol{X}}(\boldsymbol{x}|H_1)}{p_{\boldsymbol{X}}(\boldsymbol{x}|H_0)}$ (alternative divided by null)
  - With the maximum likelihood rule we reject H_0 when $L(\boldsymbol{x}) > 1$
  - This can be generalized to rejecting when $L(\boldsymbol{x}) > \xi$ where ξ is the *critical value*
    - Use this when we know one hypothesis is more likely (i.e. a prior)
    - As we increase ξ, α decreases while β increases
- Example: $H_0 : X \sim \mathcal{N}(0, 1), H_1 : X \sim \mathcal{N}(1, 1)$
  - The hypothesis changes the mean of the Gaussian

- $L(x) = \frac{f_X(x;H_1)}{f_X(x;H_2)} = \frac{e^{-(x-1)^2/2}}{e^{-x^2/2}} = e^{-\frac{1}{2}(-2x+1)}$
- In this case the threshold rule is $x \lessgtr \gamma = \ln\xi + \frac{1}{2}$
- Type I error: $\alpha(\gamma) = \int_\gamma^\infty \frac{1}{\sqrt{2\pi}} e^{-x'^2/2}\,\mathrm{d}x' = Q(\gamma)$
  * This decreases with γ
- Type II error: $\beta(\gamma) = \int_{-\infty}^{\gamma} = \frac{1}{\sqrt{2\pi}} e^{-(x'-1)^2/2}\,\mathrm{d}x' = Q(1-\gamma)$
  * This increases with γ
- Note $Q(x) = 1 - \Phi(x)$ where $\Phi(x)$ is the standard normal CDF

- So far we've only divided the region into 2, where one side is accept and the other is reject; we could also do a more complex division where we have pockets of accept in the rejection region, etc; is this better?

Theorem

Neyman Pearson Lemma: Given the likelihood ratio test $L(X), \xi$ such that

$$P[L(x) > \xi; H_0] = \alpha \quad \text{and} \quad P[L(X) \leq \xi; H_1] = \beta$$

then for any other test (region R) with $P[X \in R; H_0] \leq \alpha$ it must be that $P[X \notin R; H_1] \geq \beta$ and

$$P[X \in R; H_0] < \alpha \implies P[X \notin R; H_1] > \beta$$

That is, the LRT achieves the best possible tradeoff between α and β.

- The Neyman Pearson lemma states that given any value of α, the likelihood ratio test gives the smallest possible β to achieve that α
  - This is a constrained minimization problem of minimizing β subject to a certain α
    * Lagrangian: $\int_A f_X(x;H_1)\,\mathrm{d}x + \lambda\left(\int_R f_X(x;H_0)\,\mathrm{d}x - \alpha\right) = \lambda(1-\alpha) + \int_A (f_X(x;H_1) - \lambda f_X(x;H_0))\,\mathrm{d}x$
    * To minimize this we include x in A if $\frac{f_X(x;H_1)}{f_X(x;H_0)} < \lambda$ to make the term in the integral always negative, which is the LRT