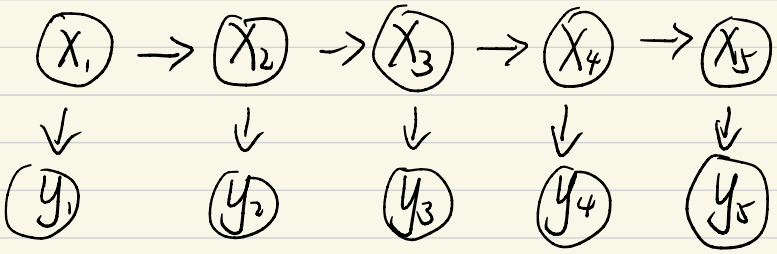
Problem 1 (Hidden Markov Model)

assum 1 $P(X_t \mid X_{t-1}, \cdots, X_1) = P(X_t \mid X_{t-1}) \quad t = 2, \cdots, T$

assum 2 $P(y_t \mid X_1, \cdots, X_t, \text{other } y\text{'s}) = P(y_t \mid X_t) \quad t = 1, \cdots, T$

homogenous: when $P(X_{t+1} = j \mid X_t = i)$ and $P(y_t \mid X_t = k)$ do not change as a function of t

(1)



$$P(X_1, \cdots, X_T, y_1, \cdots, y_T) = \prod_{t=1}^{T} P(y_t \mid pa(y_t)) P(X_t \mid pa(X_t))$$

$$\overset{\text{by assum 1, assum 2}}{=} \left(\prod_{t=1}^{T} P(y_t \mid X_t) \prod_{t=2}^{T} P(X_t \mid X_{t-1}) \right) \cdot P(X_1)$$

(2) let $P(y_t \mid X_t = k) \sim N(\mu_k, \Sigma_k)$ where Σ_k is d by d, μ_k is 1 by d

(a) if cov matrix Σ_k is diagonal

($\setminus$) needs d parameters for each k

μ_k needs d parameter for each k

since each X_{t-1} has K value and X_t has K values and each row sum to 1

thus $P(X_t \mid X_{t-1})$ requires $t \in \{2, \cdots, T\}$ K · degree of freedom $K-1$

$P(X_1)$ has $K-1$ degree of freedom, thus $K-1$ parameters

thus needs

$$Kd + Kd + K(K-1) + K - 1$$

$$= 2Kd + K^2 - K + K - 1$$

$$= 2Kd + K^2 - 1$$

(b) if Σ_k is full, since it's sysmmetric

(Σ) has $1+2+\cdots+d = \frac{(1+d)d}{2}$ degree of freedom

thus total needs

$$K\left(\frac{1+d)d}{2}\right. + kd + K(K-1)+K-1$$

$$= kd\left(\frac{3+d}{2}\right) + K^2-1$$

③

$$P(X_5=k \mid y_1^*,\cdots,y_5^*) = \frac{P(X_5=k, y_1^*,\cdots,y_5^*)}{P(y_1^*,\cdots,y_2^*)}$$

joint prob $= P(X_5=k, y_1^*,\cdots,y_5^*)$

$$= P(y_5^* \mid X_5=k, y_1^*\cdots y_4^*)\cdot P(X_5=k, y_1^*\cdots,y_4^*)$$

$$\overset{\text{by asm2}}{=} P(y_5^* \mid X_5=k)\ P(X_5=k, y_1^*,\cdots,y_4^*)$$

$$= P(y_5^* \mid X_5=k)\sum_{j=1}^{K} P(X_5=k, X_4=j, y_1^*\cdots,y_4^*)$$

$$= P(y_5^* \mid X_5=k)\sum_{j=1}^{K} P(X_5=k \mid X_4=j)\ P(X_4=j, y_1^*,\cdots,y_4^*)$$

$$= P(y_5^* \mid X_5=k)\sum_{j=1}^{K} P(X_5=k \mid X_4=j)\ P(y_4^* \mid X_4=j)\ P(X_4=j, y_1^*,y_2^*,y_3^*)$$

from hint 2

$$P(X_3=i, y_1^*, y_2^*, y_3^*) = P(y_3^* \mid X_3=i)\ P(X_3=i, y_1^*, y_2^*)$$

$$= P(y_3^* \mid X_3=i)\sum_{j=1}^{K} P(X_3=i, X_2=j, y_1^*, y_2^*)$$

$$= P(y_3^* \mid X_3=i)\sum_{j=1}^{K} P(X_3=i \mid X_2=j)\ P(X_2=j, y_1^*, y_2^*)$$

In general

$$P(X_t=i, y_1^*,\cdots,y_t^*) = P(y_t^* \mid X_t=i)\sum_{j=1}^{K} P(X_t=i \mid X_{t-1}=j)$$

$$\cdot P(X_{t-1}=j, y_1^*, y_2^*)$$

let $g_{t,i} = P(X_t=i, y_1^*,\cdots,y_t^*)$

then systemtically $\begin{cases} g_{t,i} = P(y_t^* | X_t = i) \sum_{j=1}^{K} P(X_t = i | X_{t-1} = j)\, g_{t-1,j} \\ g_{1,i} = P(X_1 = i, y_1^*) \end{cases}$

by this recursion start from $g_{1,i}$ to $g_{t,i}$ we can get

$$P(X_t = k, y_1^*, \cdots, y_t^*) = g_{t,k}$$

for magnal $P(y_1^*, \cdots y_t^*) = \sum_{k=1}^{K} P(X_t = k, y_1^*, \cdots, y_t^*) = \sum_{k=1}^{K} g_{t,k}$

thus $P(X_t = k | y_1^*, \cdots, y_t^*) = \dfrac{g_{t,k}}{\sum_{k=1}^{K} g_{t,k}}$ by bayes formula

Problem3 (Maxmum likelihood for the Multinomial Model)

1. $L(\theta) = P(D|\theta) = \prod_{i=1}^{N} P(X_i|\theta)$

$= \prod_{i=1}^{N} \theta_{X_i}$ where $N = \sum_{i=1}^{V} r_i$, X_i is the ith trial

$= \prod_{k=1}^{V} \theta_k^{r_k}$

2. $l(\theta) = \log L(\theta) = \log \prod_{k=1}^{V} \theta_k^{r_k}$

$= \sum_{k=1}^{V} r_k \log \theta_k$

adding larange operator

$$l(\theta) = \sum_{k=1}^{V} r_k \log \theta_k + \lambda(1 - \sum_{k=1}^{V} \theta_k)$$

$$\frac{dl(\theta)}{d\theta_k} = \frac{r_k}{\theta_k} - \lambda \overset{\text{want}}{=} 0 \quad \text{for } k \in \{1, \cdots, V\}$$

$$\theta_k = \frac{r_k}{\lambda}$$

consider $1 = \sum_{k=1}^{V} \theta_k = \sum_{k=1}^{V} \frac{r_k}{\lambda}$

then $\lambda = \sum_{k=1}^{V} r_k$

thus $\hat{\theta}_k^{ml} = \frac{r_k}{n} = \frac{r_k}{\sum_{i=1}^{k} r_i}$

Problem 4 (Maximum likelihood for Geometric Model)

$P(X=k) = (1-\theta)^{k-1}\theta$, $k = 1, 2, \cdots$ $0 < p < 1$

① $L(\theta) = P(D|\theta) = \prod_{i=1}^{n} P(X_i = k_i|\theta)$

$= \prod_{i=1}^{n} (1-\theta)^{k_i-1}\theta$

$= \theta^n (1-\theta)^{\sum_{i=1}^{n} k_i - n}$

② $\frac{d\,l(\theta)}{d\theta} = \frac{d\log(\theta)}{d\theta} = (n\theta^{n-1}(1-\theta)^{\sum_{i=1}^{n} k_i - n}) - (\sum_{i=1}^{n} k_i - n)(1-\theta)^{\sum_{i=1}^{n} k_i - n - 1}\theta^n$

$\overset{want}{=} 0$

then $n(1-\theta) = (\sum_{i=1}^{n} k_i - n)\theta$

$\sum_{i=1}^{n} k_i \theta = n$

$\hat{\theta}^{ml} = \frac{n}{\sum_{i=1}^{n} k_i}$

Problem 5 (Maximum likelihood for passion Model)

$D = \{X_1, \cdots, X_n\}$, $X_i \in \{0, 1, 2, \cdots\}$

$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$ $\lambda > 0$ $k \in \{0, 1, 2, 3, \cdots\}$

① $L(\theta) = P(D|\theta) = \prod_{i=1}^{n} P(X_i = k_i|\theta)$

$= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{k_i}}{k_i!}$

$= e^{-\lambda n}\lambda^{\sum_{i=1}^{n} k_i} \prod_{i=1}^{n} \frac{1}{k_i!}$

② $l(\theta) = \log L(\theta) = -\lambda n + \sum_{i=1}^{n} k_i \log(\lambda) - \sum_{i=1}^{n} \log(k_i!)$

consider $\frac{\partial l(\theta)}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} k_i}{\lambda} \overset{\text{want}}{=} 0$

thus $\hat{\lambda}^{ml} = \frac{\sum_{i=1}^{n} k_i}{n}$

Problem 6 (Maximum likelihood with Measurement Variance per point)

D consists of N scalar measurements x_i, $1 \le i \le N$

same mean μ, different variance σ_i^2 $1 \le i \le N$ know σ_i^2

brightness μ of a star, measurement x_i taken at different locations i on planet with noise σ_i^2

① $p(x_i \mid \mu, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(x_i-\mu)^2}$

$$L(\theta) = p(D \mid \theta) = \prod_{i=1}^{n} p(x_i = k_i \mid \theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(x_i-\mu)^2}$$

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}(x_i-\mu)^2\right)$$

② $\frac{dl(\theta)}{d\mu} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(x_i-\mu) \overset{\text{want}}{=} 0$

thus $\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\mu = \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}$

$$\hat{\mu}^{ml} = \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \cdot \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)^{-1}$$

③ $\hat{\mu}^{ml} = \sum_{i=1}^{n} \frac{1}{\sigma_i \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)} x_i$

let $w_i = \frac{1}{\sigma_i^2 \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)}$

let $\tau_i = 1/\sigma_i^2$ be precision

thus weights for x_i are its precision normalized by the sum of all precison for x_k $k=\{1,\dots,n\}$

Problem 8

moment matching is trying to find parameters such as the moments of the model (with its estimated parameters) match the empirical moments in the observed data

uniformly distributed $p(x)=\frac{1}{b-a}$

(1) $$E[X]=\int_a^b x\,p(x)\,dx=\frac{1}{b-a}\frac{x^2}{2}\Big|_a^b=\frac{1}{b-a}\frac{b^2-a^2}{2}=$$

$$=\frac{a+b}{2}$$

$$E[X^2]=\int_a^b x^2 p(x)\,dx=\frac{1}{b-a}\frac{x^3}{3}\Big|_a^b=\frac{1}{b-a}\frac{b^3-a^3}{3}$$

$$=\frac{b^2+ab+a^2}{3}$$

want: $$\frac{a+b}{2}=E[X]=\frac{1}{N}\sum_{i=1}^{N}x_i$$

$$\frac{b^2+ab+a^2}{3}=E[X^2]=\frac{1}{N}\sum_{i=1}^{N}x_i^2$$

$$ab=(2E[X])^2-3E[X^2]=4(E[X])^2-3E[X^2]$$

$$a+b=2E[X]\quad\Rightarrow\quad b=2E[X]-a$$

$$a^2-2E[X]a+\left[4(E[X])^2-3E[X^2]\right]=0$$

$$a=\frac{2E[X]\pm\sqrt{-12(E[X])^2+12E[X^2]}}{2}$$

$$=E[X]\pm\sqrt{3}\sqrt{E[X^2]-E[X]^2}$$

$$b=E[X]\mp\sqrt{3}\sqrt{E[X^2]-E[X]^2}$$

$b>0$

thus $a^* = \frac{1}{N}\sum_{i=1}^{N} x_i - \sqrt{3}\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2}$

$b^* = \frac{1}{N}\sum_{i=1}^{N} x_i + \sqrt{3}\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2}$

② $L(\theta) = P(D|\theta) = \prod_{i=1}^{n} p(x_i|a,b)$

$= \prod_{i=1}^{n} \frac{1}{b-a}$

$= \left(\frac{1}{b-a}\right)^n$

$\max L(\theta)$ is equal to minimize $b-a$

WLOG let $\hat{b}_{ML} = \max\{x_i\}$ $i \in \{1,\dots,N\}$

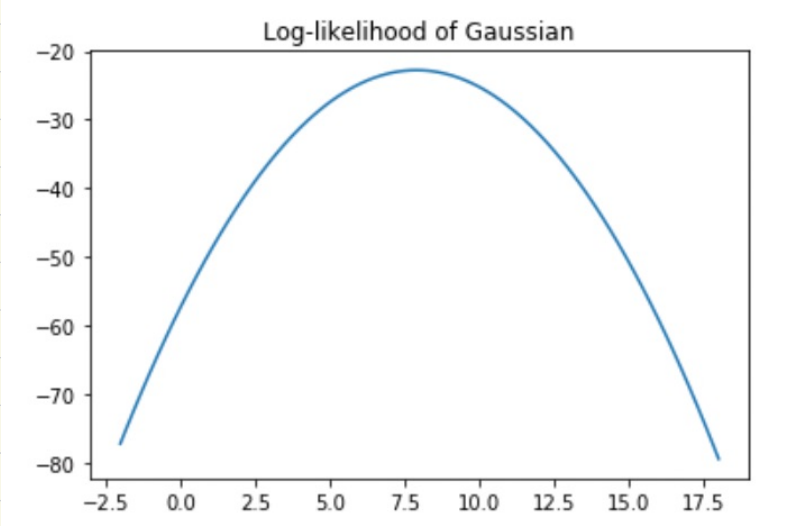
$\hat{a} = \min\{x_i\}$ $i \in \{1,\dots,N\}$

③

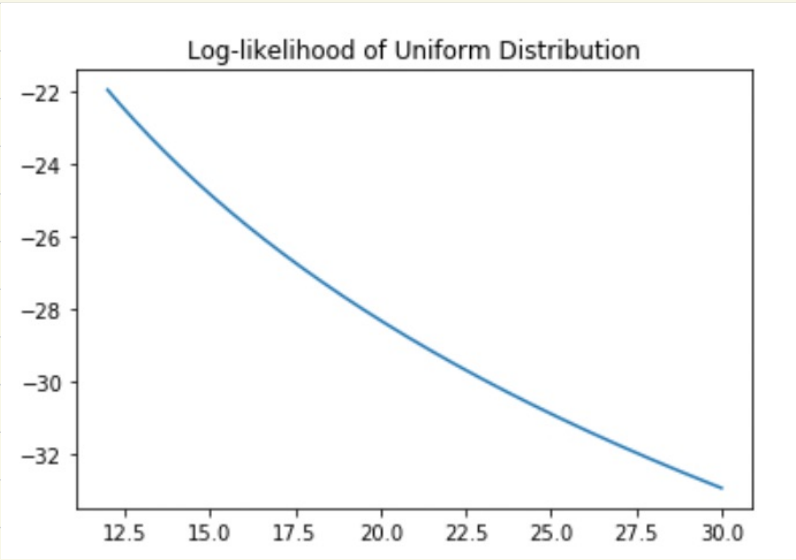
the maximum likehood function estimator is highly depends on data's maximum and minimum however if we add more data, there may have some number less than a or larger than b which $P(\theta)$ will give probibity 0 for prediction than

method of moments gives probibity range which a^*, b^* that $a^* \le \min_i(x)$, $b^* \ge \max_i(x)$ which means that by thus $P(\theta_{no\,moments})$ may predict some number that even not exist in the future data observation

Problem 2



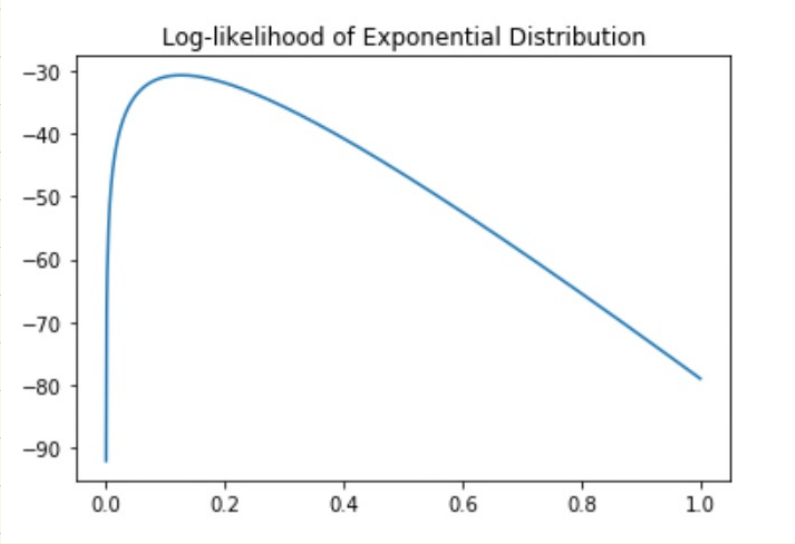
(1)



(2)

since data has maximum 12, if $b < 12$ then likehood $= 0$

log(likehood) doesn't exist

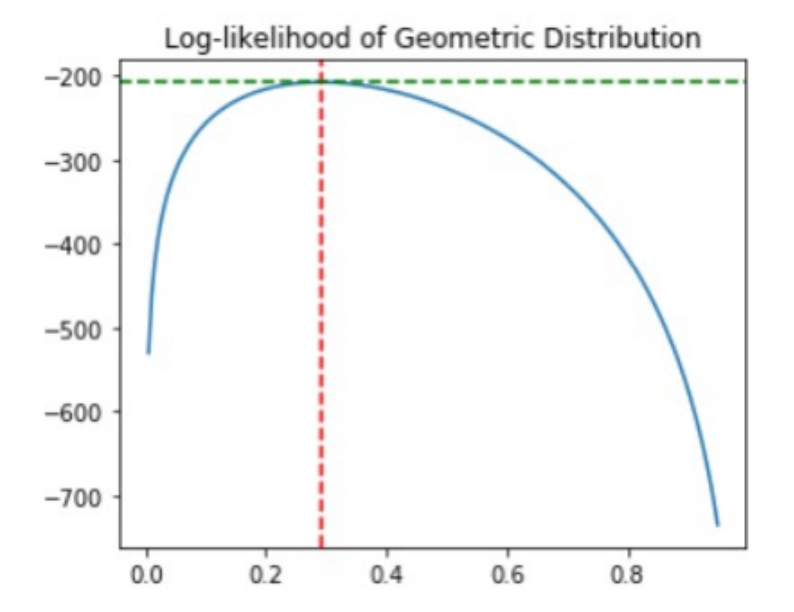


Problem 7

1. Gemotric Model $P(X=k)=(1-\theta)^{k-1}\theta \quad k=0,1,2\cdots$

   possion model $P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!} \quad \lambda>0 \quad k\in\{0,1,2,3,\cdots\}$

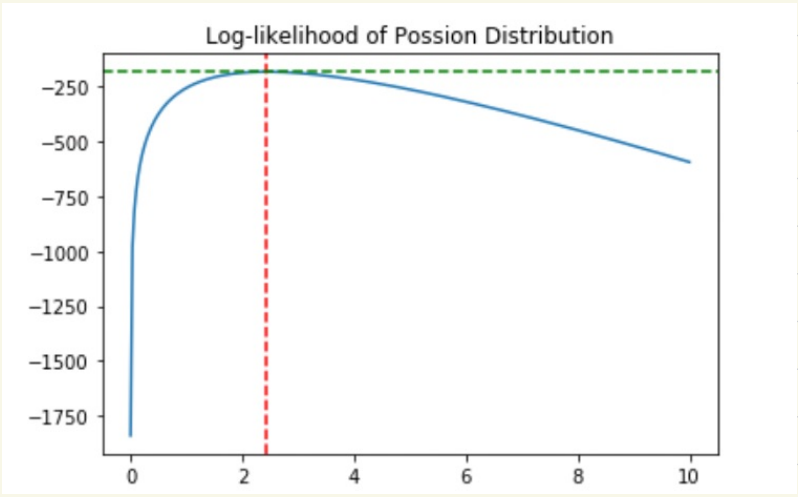
2.



Geometric Distribution

$\hat{\theta}^{ml} = 0.29069767$

$\ell(\theta)_{max} = -207.35$

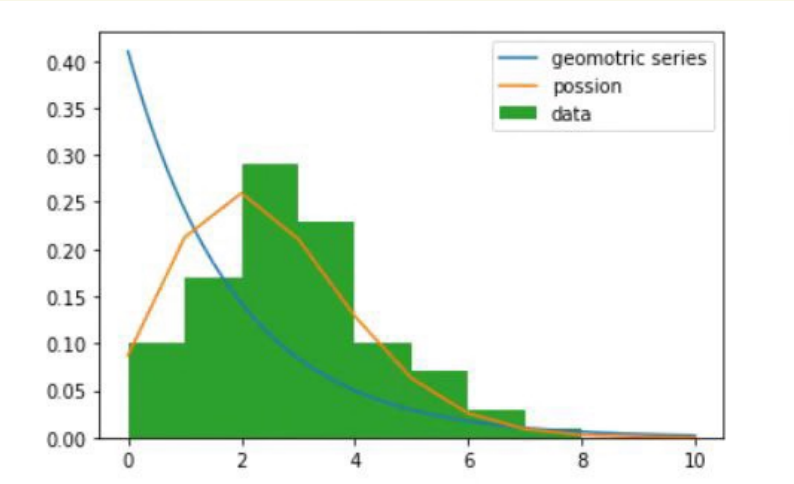


Poisson Distribution

$\hat{\lambda}^{ml} = 2.44$

$\ell(\theta)_{max} = -181.22$

3.



4. Poisson is better to fit data both by observation and has smaller maximaxum bihebihood between two types.

5. No. since model with two paramemters has degree of freedom 2.

we can use extra-sum-of-squares F test compas or use Information theory approach Akaike's Grterion AIC to help compare them.

graphpad.com/guides/prism/7/curve-fitting/reg_approaches to comparing_models.htm?tocno&printWindow