Maximum Likelihood Estimation

- MLE and its recent application computational biology -

Hayan Lee@Schatz Lab

March 17, 2015 Schatz Lab Meeting



Outline

Definition

- Likelihood
- Maximum Likelihood Estimation
- Application
 - Univariate Gaussian Mean
 - Univariate Poisson Mean •
 - C. Pop et al., Causal signals between codon bias, mRNA structure, and ۲ the efficiency of translation and elongation, Molecular systems biology, 2014 -> POTSCON
 - N. Loman et al., A complete bacterial genome assembled de novo mutorier Crassien men using only nanopore sequencing data, BioRxiv, 2015
- Summary

Cold Spring Harbor Laboratory



Probability of the data that you observed given the hypothesis



Maximum Likelihood Estimation

- Estimate parameters that maximize the likelihood
- Choose a hypothesis that gives you the highest probability that you observed the data
- Infer under what model you can get the data that you observed

$$H_{MLE} = \operatorname{argmax} P(D|H)$$

$$\theta_{MLE} = \operatorname{argmax} P(D|\theta)$$

- Settings $p_{x} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\theta)^2}{2\sigma^2})$
- Under what mean θ do we have the highest probability?





- Settings
 - $x \sim N(\theta, \sigma^2)$

$$- P(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\theta)^2}{2\sigma^2})$$

• Under what mean θ do we have the highest probability?

 $P(D|\theta_1)$ $P(D|\theta_2)$ \vdots $P(D|\theta_1)$

 $\theta_{MLE} = argmax P(D|\theta_n)$

- Settings - (, 2, 3) - P(x|\theta) = $\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\theta)^2}{2\sigma^2})$
- Under what mean θ do we have the highest probability?

P(1,2,3,16) 11 z z D $P(110) \cdot P(216) \cdot P(3.6)$ $\frac{3''}{T} P(X;16)$ 1=1





CSH Cold Spring Harbor Laboratory



• Optimization





Univariate Poisson MeanSettingsG(M, G, M, G,

• Under what mean do we have the highest probability?

$$P(D|\lambda_1)$$

$$P(D|\lambda_2)$$

$$P(D|\lambda_3)$$

$$\vdots$$

$$\lambda_{MLE} = argmax P(D|\lambda_n)$$

CSH Cold Spring Harbor Laboratory

Univariate Poisson Mean

- Settings
 - $-x \sim Poiss(\lambda)$

old Spring Harbor Laboratory

$$- P(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

• Under what mean θ do we have the highest probability?

$$P(D|\lambda_1)$$

$$P(D|\lambda_2)$$

$$P(D|\lambda_3)$$

$$\vdots$$

$$\lambda_{MLE} = argmax P(D|\lambda_n)$$

- Settings $- \underbrace{x_1} \underbrace{x_2}_{\lambda} \cdots \underbrace{x_n}_{\lambda} Poiss(\lambda)$ $- P(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$
- Under what mean do we have the highest probability?

 $P(\chi_{1},\chi_{2},...,\chi_{n}|\lambda)$ $= \prod_{\substack{n \in \mathbb{N} \\ \forall i \in \mathbb{N}}} \overline{id}$ $P(\chi_{i}|\lambda) \cdot P(\chi_{2}|\lambda) \cdot P(\chi_{k})$ $= \prod_{\substack{n \in \mathbb{N} \\ \forall i \in \mathbb{N}}} P(\chi_{i}|\lambda)$

Univariate Poisson Mean

- Optimization
 - Derivation
 - Log likelihood





Univariate Poisson Mean



kr Jm.

Univariate Poisson Mean

 $f = \sum_{i=1}^{n} x_i$

Optimization

 γ

mple men

Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation

Molecular systems biology, 2014

Translation Model



Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation Molecular systems biology, 2014

• Notation



- μ_{mk} the dwell time at (m,k)
- J_{mk} : the rate/flow at (m,k)

 $\sum_{k \text{ in } m} \frac{d_{mk}}{\mu_{mk}} /$

old Spring Harbor Laboratory

- proportional to $\frac{\pi_{mk}}{\mu_{mk}} \approx \frac{d_{mk}}{\mu_{mk}}$ - J_m : the protein synthesis rate for message m
 - (ribosome flow out of stop codon)

$$- J_{mk} = J_m$$
 (by flow conservation constraints)

Simons Center for Quantitative Biology

And the source

Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation Molecular systems biology, 2014

$$x \sim Poiss(\lambda)$$
$$P(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Cold Spring Harbor Laboratory



Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation

Molecular systems biology, 2014

$$\arg\max_{\mu_{m,\mu^{c}}^{c},\mu^{c}}\log\prod_{m}\prod_{k}e^{-\mu_{m}^{c}}\mu_{m}^{c}\overset{d_{mk}}{/}_{J_{m}}-penalty$$
$$\arg\max_{\mu_{m,\mu^{c}}^{c},\mu^{c}}\log\prod_{m}\prod_{k}e^{-\mu_{m}^{c}}\mu_{m}^{c}\overset{d_{mk}}{/}_{J_{m}}-\sum_{m,c}w_{m}^{f}\left(\log\mu_{m}^{c}-\log\mu^{c}\right)$$

Optimization

- Convex optimization
- Coordinate decent by iterating through codons c and learning

 μ_m^c, μ^c -L-BFGS *li*matlab wrapper)

6H&Cold Spring Harbor Laboratory

N. Loman, ... J. Simpson

A complete bacterial genome assembled de novo using only nanopore sequencing data,

BioRxiv, 2015





N. Loman, ... J. Simpson

ang unly nan Gragen and how A complete bacterial genome assembled de novo using only nanopore sequencing data,

BioRxiv, 2015



1 1

N. Loman, ... J. Simpson

A complete bacterial genome assembled de novo using only nanopore sequencing data,

 e_i

 e_{i+1}

 S_{i+1}

BioRxiv, 2015

To handle split event, and skip event

 S_{i-1}

$$P(\pi, e_1, e_2, ..., e_n | S, \Theta) = \prod_{i=1}^n P(e_i | \pi_i, \mu_{s_i}, \sigma_{s_i}) P(\pi_i | \pi_{i-1}, S)$$

CSH Cold Spring Harbor Laboratory

N. Loman, ... J. Simpson A complete bacterial genome assembled de nove asing only nanopore sequencing date BioRxiv, 201 Optimization – Block replacement algorithm Sample alternative substring from the reads Mutation algorithm (inspired by Quiver) Generates all strings within given edit distance



H Cold Spring Harbor Laboratory