BP-203 Foundations for Mathematical Biology

Statistics Lecture III

By Hao Li

Nov 8, 2001

Statistical Modeling and Inference

- data collection
- constructing probabilistic model
- inference of model parameters
- interpreting results
- making new predictions

Maximum likelihood Approach

Example A: Toss a coin N times, observe m heads in a specific sequence

Model: binomial distribution
Inference: the parameter p
Prediction: e.g., how many heads will be observed
for another L trials

Prob. of observing a specific sequence of m heads

$$P(m \mid p) = p^m (1-p)^{N-m}$$

Find a p such that the above prob. is maximized

$$\frac{\partial \log P(m \mid p)}{\partial p} \Big|_{\hat{p}} = 0 \qquad \hat{p} = m / N$$

$$\log P(m \mid \hat{p}) = N\left[\hat{p} \log \hat{p} + (1-\hat{p}) \log(1-\hat{p})\right]$$

↑
-entropy

How good is the estimate?

Distribution of $\hat{p}$ under repeated sampling

Central limit theorem → distribution of m approaches normal for large N

$$m \sim Np \pm \sqrt{Np(1-p)}$$

$$\hat{p} \sim p \pm \sqrt{p(1-p)/N}$$

Thus the estimate converges to the real p with a square-root convergence

Maximum likelihood Approach

Example B: $x_1, x_2, ..., x_N$

independent and identically distributed (i.i.d) sample drawn from a normal distribution $N(\mu, \sigma^2)$

Estimate the mean and the variance

Maximizing the likelihood function (show this is true in the homework)

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{N} x_i / N$$

$$\hat{\sigma}^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2 / N$$

General formulation of the maximum likelihood approach

D: observed data

M: the statistical model

θ parameters of the model

$P(D \mid M, \theta)$ probability of observing the data given the model and parameters

$L(\theta; D) \equiv P(D \mid M, \theta)$ the likelihood of θ as a function of data

Maximum likelihood estimate of the parameters

$$\hat{\theta} = \arg\max L(\theta; D)$$

Theorem: $\hat{\theta}$ converges to the true θ_0 in the large sample limit with error $\sim 1/\sqrt{N}$

Example C: Segmentation

a sequence of head (1) and tail (0) is generated by first using a coin with p_1 and then change to a coin with p_2 the change point unknown

Data = (00101000000001011101111100010)

$$P(seq, x \mid p_1, p_2) = {p_1}^{m_1(x)}(1-p_1)^{x-m_1(x)}\,{p_2}^{m_2(x)}(1-p_2)^{N-x-m_2(x)}$$

x position right before the change

$m_1(x)$ number of 1's up to x

$m_2(x)$ number of 1's after x

N total number of tosses

Example C continued

For fixed x maximize $P(seq, x \mid p_1, p_2)$ with respect to p_1 and p_2

$$\log P(seq, x \mid \hat{p}_1, \hat{p}_2) = x\left[\hat{p}_1 \log \hat{p}_1 + (1-\hat{p}_1)\log(1-\hat{p}_1)\right] + (N-x)\left[\hat{p}_2 \log \hat{p}_2 + (1-\hat{p}_2)\log(1-\hat{p}_2)\right]$$

$$\hat{p}_1 = m_1(x)/x$$
$$\hat{p}_2 = m_2(x)/(N-x)$$

Then maximize $P(seq, x \mid \hat{p}_1, \hat{p}_2)$ with respect to x

The above approach is sometime referred as “entropic segmentation”, as it tries to minimize the total entropy

A generalization of the above model to 4 alphabet and unknown number of breaking points can be used to segment DNA sequences into regions of different composition. more naturally described by a hidden Markov model.

Example D: detecting weak common sequence patterns in
a set of related sequences

e.g., local sequence motifs for functionally or structurally
related proteins (no overall sequence similarity)

regulatory elements in the upstream regions of
co-regulated genes, could be genes clustered together
by microarray data

the simplest situation: each sequence contain one realization of the
motif with given length, but the starting positions are unknown

Example: 22 genes identified as pho4 target by microarray, O'shea lab

YAR071W:600:-600

\catcaagatgagaaaataaagggatttttttcgttcttttatcattttctctttctcacttccgactacttcttatatctactttcatcgtttcattcatcgtgggtgtctaataaagtttta
atgacagagataaccttgataagctttttcttatacgctgtgtcacgtatttattaaattaccacgttttcgcataacattctgtagttcatgtgtactaaaaaaaaaaaaaaaaaa
gaaataggaaggaaagagtaaaaagttaatagaaaacagaacacatccctaaacgaagccgcacaatcttggcgttcacacgtgggtttaaaaaggcaaattacacag
aatttcagaccctgtttaccggagagattccatattccgcacgtcacattgccaaattggtcatctcaccagatatgttatacccgttttggaatgagcataaacagcgtcgaa
ttgccaagtaaaacgtatataagctcttacatttcgatagattcaagctcagtttcgccttggttgtaaagtaggaagaagaagaagaagaggaacaacaacagcaaa
gagagcaagaacatcatcagaaatacca\

YBR092C:600:-600

\aatcaatgacttctacgactatgctgaaaagagagtagccggtactgacttcctaaaggtctgtaacgtcagcagcgtcagtaactctactgaattgaccttctactgggac
tggaacactactcattacaacgccagtctattgagacaatagttttgtataactaaataatattggaaactaaatacgaatacccaaatttttatctaaattttgccgaaagatta
aaatctgcagagatatccgaaacaggtaaatggatgtttcaatccctgtagtcagtcaggaacccatattatattacagtattagtcgccgcttaggcacgcctttaattagca
aaatcaaaccttaagtgcatatgccgtataagggaaactcaaagaactggcatcgcaaaaatgaaaaaaaggaagagtgaaaaaaaaaaaattcaaaagaaatttacta
aataataccagtttgggaaatagtaaacagctttgagtagtcctatgcaacatatataagtgcttaaatttgctggatggaagtcaattatgccttgattatcataaaaaaaata
ctacagtaaagaaagggccattccaaattacct\

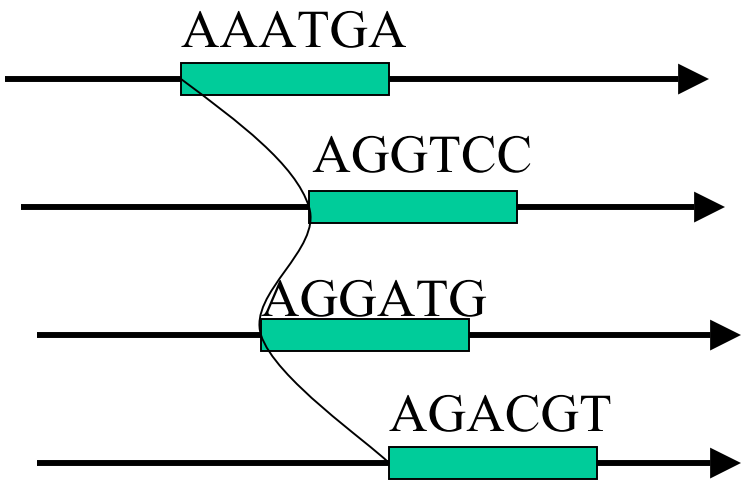
YBR093C:600:-600

\cgctaatagcggcgtgtcgcacgctctctttacaggacgccggagaccggcattacaaggatccgaaagttgtattcaacaagaatgcgcaaatatgtcaacgtatttgg
aagtcatcttatgtgcgctgctttaatgttttctcatgtaagcggacgtcgtctataaacttcaaacgaaggtaaaaggttcatagcgcttttctttgtctgcacaaagaaatata
tattaaattagcacgttttcgcatagaacgcaactgcacaatgccaaaaaaagtaaaagtgattaaaagagttaattgaataggcaatctctaaatgaatcgatacaaccttg
gcactcacacgtgggactagcacagactaaatttatgattctggtccctgttttcgaagagatcgcacatgccaaattatcaaattggtcaccttacttggcaaggcatatac
ccatttgggataagggtaaacatctttgaattgtcgaaatgaaacgtatataagcgctgatgttttgctaagtcgaggttagtatggcttcatctctcatgagaataagaacaa
caacaaatagagcaagcaaattcgagattacca\

YBR296C:600:-600

\gaaatctcggtttcacccgcaaaaaagtttaaatttcacagatcgcgccacaccgatcacaaaacggcttcaccacaagggtgtgtggctgtgcgatagacctttttttttctt
tttctgctttttcgtcatccccacgttgtgccattaatttgttagtgggcccttaaatgtcgaaatattgctaaaaattggcccgagtcattgaaaggctttaagaatataccgtac
aaaggagtttatgtaatcttaataaattgcatatgacaatgcagcacgtgggagacaaatagtaataatactaatctatcaatactagatgtcacagccactttggatccttcta
ttatgtaaatcattagattaactcagtcaatagcagatttttttacaatgtctactgggtggacatctccaaacaattcatgtcactaagcccggttttcgatatgaagaaaattat
atataaacctgctgaagatgatctttacattgaggttattttacatgaattgtcatagaatgagtgacatagatcaaaggtgagaatactggagcgtatctaatcgaatcaatat
aaacaaagattaagcaaaaatg\

A model for the motif



alignment matrix

	1	2	3	4	5	6
A	4	1	2	1	0	1
C	0	0	0	1	1	1
G	0	3	2	0	2	1
T	0	0	0	2	1	1

position specific probability matrix

	1	2	3	4	5	6
A	1.00	0.25	0.50	0.25	0.00	0.25
C	0.00	0.00	0.00	0.25	0.25	0.25
G	0.00	0.75	0.50	0.00	0.50	0.25
T	0.00	0.00	0.00	0.50	0.25	0.25

$\leftarrow f_{i,\sigma}$

Model: probability of observing certain base inside the motif is given by the above matrix

probability of observing certain base outside the motif is given by the background frequency f_{σ}^{0}

Starting positions of the motif unknown $\vec{x} = (x_1, x_2, ..., x_N)$

Position specific probability matrix unknown $f_{i,\sigma}$

need to be inferred from the observed sequence data

$$P(seq, \vec{x} \mid f_{i,\sigma}) = \prod_{i=1}^{N} \left(\prod_{j=1}^{x_i - 1} f^0_{\sigma_{ij}} \prod_{j=x_i}^{x_i + w - 1} f_{j - x_i + 1, \sigma_{ij}} \prod_{j=x_i + w}^{L} f^0_{\sigma_{ij}} \right)$$

N	Number of sequences
L	Length of the sequence
w	Width of the motif
σ_{ij}	Base of sequence i at position j

$$P(seq, \vec{x} \mid f_{i,\sigma}) = const \prod_{j=1}^{w} \prod_{\sigma} \left(\frac{f_{j,\sigma}}{f^0_\sigma} \right)^{n_{j,\sigma}(\vec{x})}$$

likelihood ratio

$n_{j,\sigma}(\vec{x})$ Total number of count for base σ at position j in the alignment

Maximizing $P(seq, \vec{x} \mid f_{i,\sigma})$ w.r.t. $f_{i,\sigma}$ With $\vec{x}$ fixed

$$\log P(seq, \vec{x} \mid \hat{f}_{i,\sigma}) = N \sum_{j=1}^{w} \hat{f}_{j,\sigma} \log\left(\frac{\hat{f}_{j,\sigma}}{f_{\sigma}^{0}}\right)$$

log likelihood ratio
relative entropy

$$\hat{f}_{j,\sigma} = \frac{n_{j,\sigma}(\vec{x})}{\sum_{\sigma} n_{j,\sigma}(\vec{x})}$$

in reality, this formula is modified by adding pseudo counts due to Baysian estimate

Then maximize the above relative entropy w.r.t $\vec{x}$
→ Alignment path.

Stormo-Hartzell Algorithm: Consensus

- each of the length w substrings of the first sequence are aligned against all the substrings of the same length in the second sequence, matrices derived, N top matrices with highest information contents are saved

- the next sequence on the list is added to the analysis, all the matrices saved previously are paired with the substrings of the added sequence and top N matrices saved

- repeat the previous step until all the sequences have been processed

Consensus output for Pho4 regulated genes

MATRIX 1
number of sequences = 22
information = 8.80903
ln(p-value) = -153.757 p-value = 1.67566E-67
ln(expected frequency) = -13.357 expected frequency = 1.58165E-06

A |	6	5	20	3	0	3	0	0	0	6
G |	11	0	0	5	22	0	21	15	14	2
C |	4	17	0	14	0	0	1	2	8	1
T |	1	0	2	0	0	19	0	5	0	13
	G	C	A	C	G	T	G	G	G	T

1|1 : 1/317 ACACGTGGGT
2|2 : 2/55 AAAGGTCTGT
3|3 : 3/347 ACACGTGGGA
4|4 : 4/274 GCACGTGGGA
5|5 : 5/392 CAACGTGTCT
6|6 : 6/395 ACAAGTGGGT
7|7 : 7/321 ACACGTGGGA
8|8 : 8/536 GCAAGTGGCT
9|9 : 9/177 GCTGGTGTGT
10|10 : 10/443 GCACGTGTCT
11|11 : 11/14 CCAGGTGCCT
12|12 : 12/502 GAAAGAGGCA
13|13 : 13/354 GCACGAGGGA
14|14 : 14/257 GCACGTGCGA
15|15 : 15/358 TCACGTGTGT
16|16 : 16/316 ACACGTGGGT
17|17 : 17/479 GCACGTGGCT
18|18 : 18/227 GATGGTGGCT
19|19 : 19/186 GCACGTGGGG
20|20 : 20/326 GAAGGAGGGG
21|21 : 21/307 CCACGTGGGC
22|22 : 22/255 CCACGTGGCT

Maximum likelihood estimate with missing data
General formulation
Expectation and Maximization (EM) algorithm

Missing data: in example C, the point where the coin is changed
in example D, the starting positions of the motif

in the maximum likelihood approach, there is a crucial distinction between parameters (population) such as the position specific probability matrix and the missing data, since missing data grow with the sample size and in general can not be recovered precisely even if the sample size goes to infinity

For many problems, it is necessary to sum over all missing data

$$L(x;\theta) = \sum_{y} P(x, y \mid \theta)$$

Where x is the observed data and y is the missing data

To estimate the parameters, one maximize the likelihood function $L(x;\theta)$ however, it is often difficult to perform the summation over missing data explicitly

Expectation Maximization (EM) algorithm

Improve the estimate of the parameters iteratively
Given an estimate θ^t find θ^{t+1} that increases the likelihood function

E step: calculate the Q function, the expectation of $\log P(x, y \mid \theta)$ over missing data with prob. given by the current parameter

M step: maximize the Q function to get an new estimate $\theta^{t+1} = \arg\max Q(\theta \mid \theta^t)$

$$Q(\theta \mid \theta^t) \equiv \sum_y P(y \mid x, \theta^t) \log P(x, y \mid \theta)$$

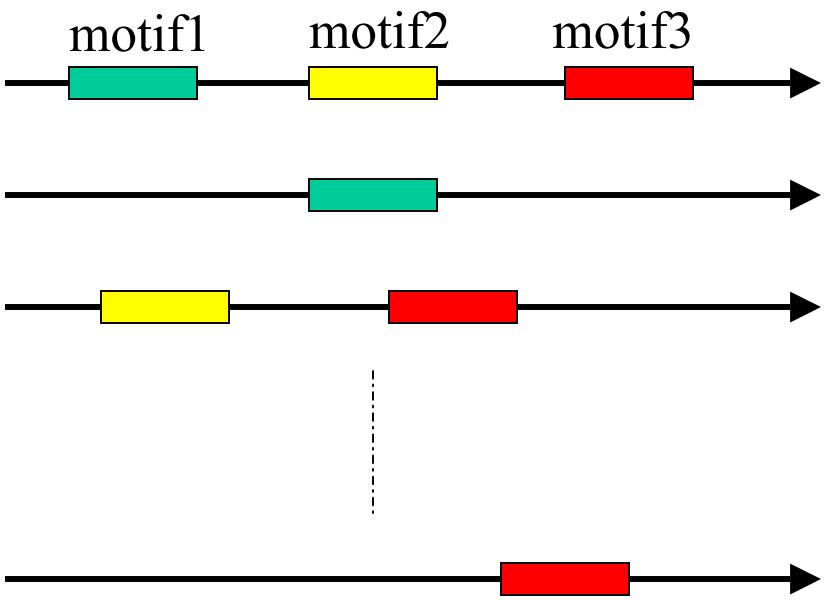
That the EM algorithm always increase the likelihood function
Can be proved by the following equation and inequality

$$\log P(x \mid \theta) - \log P(x \mid \theta^t) = Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t) + \sum_y P(y \mid x, \theta^t) \log \frac{P(y|x,\theta^t)}{P(y|x,\theta)}$$

$$\log P(x \mid \theta) - \log P(x \mid \theta^t) \geq Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t)$$

Example E: identifying combinatorial motifs by word segmentation

A set of Regulatory Sequences



chapterptgpbqdrftezptqtasctmvivwpecjsnisrmbtqlmlfvetl
loomingsfkicallxjgkmekysjerishmaeljplfsomeylqyearstvh
njbagoaxhjtjcokhvneverpmqpmindhowzrbdlzjllonggbhqi
preciselysunpvskepfdjktcgarwtnxybgcvdjfbnohavinglittl
ezorunozsoyapmoneyyvugsgtsqintmyteixpurseiwfmjwgj
nyyveqxwftlamnbxkrsbkyandrnothingcgparticularwtzao
qsjtnmtoqsnwvxfiupinterestztimebymonlnshoreggditho
ughtyxfxmhqixceojjzdhwouldsailpcaboutudxsbsnewtpg
gvjaasxmsvlittleplvcydaowgwlbzizjlnzyxandzolwcudthjd
osbopxkkfdosxardgcseebbthefzrsskdhmawateryjikzicim
ypartmofprtheluworldvtoamfutitazpisagwewayrqbkiosh
avebojwphiixofprmalungipjdrivingpkuyoikrwxoffodhicb
nimtheixyucpdzacemspleenqbpcrmhwvddyaiwnandada
bkpgzmptoregulatingeetheslcirculationvsuctzwvfyxstuzr
dfwvgygzoejdfmbqescwheneverpitfindfmyselfcgrowingne
ostumrydrrthmjsmgrimcczhjmgbkwczoaboutjbwanbwzq
thehrjvdrcjjgmouthuutwheneveritddfouishlawwphxnae

Bussemaker/Li/Siggia Model:

Probabilistic Segmentation/Maximum likelihood

A probabilistic dictionary
Words
probabilities

$$\left\{\begin{array}{ll} A & \rightarrow P_A \\ C & \rightarrow P_C \\ G & \rightarrow P_G \\ T & \rightarrow P_T \\ GC & \rightarrow P_{GC} \\ TATAA & \rightarrow P_{TATAA} \end{array}\right\}$$

A | G | T | A | T | A | A | G | C
A | G | T A T A A | G C
A | G | T A T A A | G | C

word boundary is missing

maximizing the likelihood function

$$Z = \sum_{Seg} P_{w_1} P_{w_2} P_{w_3} \ldots P_{w_n}$$

Dictionary Construction

Parameter inference: given the entries in the dictionary, find P_W by maximizing the likelihood function. Starting with a simple dictionary with all possible words

Model improvement: do statistical test on longer words based on the current dictionary, add the ones that are over-represented re-assign P_W by maximizing the likelihood function

Iterate the above

EM algorithm for the word segmentation

$N_w(Seg)$ Number of word w in a given segmentation

$$L(\{p_w\}; seq) = \sum_{Seg} \prod (p_w)^{N_w(Seg)}$$

E step $$Q(\{p_w\} \mid \{p_w(t)\}) = \sum_w \langle n_w \rangle_t \log p_w$$

M step $$p_w(t+1) = \frac{\langle N_w \rangle_t}{\sum_w \langle N_w \rangle_t}$$

Ditionary1		Ditionary2		Dictionary3	
e	0.065239	e	0.048730	e	0.042774
t	0.055658	s	0.042589	s	0.040843
a	0.052555	a	0.040539	a	0.038595
o	0.050341	t	0.040442	i	0.036897
n	0.049266	i	0.038550	t	0.036871
i	0.048101	d	0.038547	d	0.036323
s	0.047616	o	0.036486	l	0.035336
h	0.047166	l	0.036300	c	0.034818
r	0.043287	g	0.034509	m	0.034650
l	0.041274	r	0.034496	y	0.034482
d	0.039461	c	0.033916	b	0.034396
u	0.034742	m	0.033724	r	0.034105
m	0.034349	n	0.033321	p	0.034044
g	0.034001	y	0.033227	w	0.033819
w	0.033967	p	0.033156	n	0.033817
c	0.032934	f	0.032863	g	0.033676
f	0.032597	b	0.032780	f	0.033534
y	0.031776	w	0.032009	o	0.033206
p	0.031711	h	0.031494	h	0.033200
b	0.031409	v	0.030727	k	0.032103
v	0.028268	k	0.030445	v	0.031498
k	0.028113	u	0.030379	j	0.031209
j	0.026712	j	0.029268	u	0.031186
q	0.026561	z	0.028905	z	0.031003
z	0.026542	x	0.028404	x	0.030544
x	0.026357	q	0.028123	q	0.030244
		th	0.009954	the	0.005715
		in	0.006408	ing	0.003237
		er	0.004755	and	0.003128
		an	0.004352	in	0.002968
		ou	0.003225	ed	0.002547
		on	0.003180	to	0.002496
		he	0.003108	of	0.002486
		at	0.002851	en	0.001331
		ed	0.002804	an	0.001313
		or	0.002786	th	0.001270
		en	0.002538	er	0.001250
		to	0.002511	es	0.001209
		of	0.002475	at	0.001181
		st	0.002415	it	0.001171
		nd	0.002297	that	0.001165

Words	<Nw>	quality factor
abominate	2.0000	1.0000
achieved	2.0000	1.0000
aemploy	2.0000	1.0000
affrighted	2.0000	1.0000
afternoon	2.0000	1.0000
afterwards	5.0000	1.0000
ahollow	2.0000	1.0000
american	3.0000	1.0000
anxious	2.0000	1.0000
apartment	2.0000	1.0000
appeared	4.0000	1.0000
astonishment	4.0000	1.0000
attention	2.0000	1.0000
avenues	2.0000	1.0000
bashful	2.0000	1.0000
battery	2.0000	1.0000
beefsteaks	2.0000	1.0000
believe	2.0000	1.0000
beloved	2.0000	1.0000
beneath	6.0000	1.0000
between	12.0000	1.0000
boisterous	3.0000	1.0000
botherwise	2.0000	1.0000
bountiful	2.0000	1.0000
bowsprit	2.0000	1.0000
breakfast	5.0000	1.0000
breeding	2.0000	1.0000
bulkington	3.0000	1.0000
bulwarksb	2.0000	1.0000
bumpkin	2.0000	1.0000
business	6.0000	1.0000
carpenters	2.0000	1.0000

Table 1. Known cell cycle sites and some metabolic sites that match words from our genomewide dictionary

MCB	ACGCGT	AAACGCGT ACGCGTCGCGT CGCGACGCGT TGACGCGT
SCB	CRCGAAA	ACGCGAAA
SCB'	ACRMSAAA	ACGCGAAA ACGCCAAA AACGCCAA
Swi5	RRCCAGCR	GCCAGCG GCAGCCAG
SIC1	GCSCRGC	GCCCAGCC CCGCGCGG
MCM1	TTWCCYAAWNNGGWAA	TTTCCNNNNNNGGAAA
NIT	GATAAT	TGATAATG
MET	TCACGTG	RTCACGTG TCACGTGM CACGTGAC CACGTGCT
PDR	TCCGCGGA	TCCGCGG
HAP	CCAAY	AACCCAAC
MIG1	KANWWWWATSYGGGGW	TATATGTG CATATATG GTGGGGAG
GAL4	$CGGN_{11}CCG$	$CGGN_{11}CCG$

our dictionary vs. known TF binding sites

Yeast promoter database 443 non-redundant sites
(Zhu and Zhang, cold spring harbor)

	# of matches	Expected (standard deviation)
Our dictionary	**114**	25 (4.8)
Scrambled dictionary	**33**	14 (3.3)
Brazma et al.	**30**	9 (2.9)