CSE 527
Autumn 2007
Lectures 6-7
MLE, EM, Expression

Outline

• MLE: Maximum Likelihood Estimators
• EM: the Expectation Maximization Algorithm
• Bio: Gene expression and regulation
• Next: Motif description & discovery

Probability Basics, I

MLE

Maximum Likelihood Estimators

Sample Space

\{1, 2, \ldots, 6\}

\mathbb{R}

Distribution

\text{e.g.} \quad \begin{align*}
p_1, \ldots, p_6 & \geq 0; \quad \sum_{1 \leq i \leq 6} p_i = 1 \\
f(x) & \geq 0; \quad \int_{\mathbb{R}} f(x) dx = 1 \\
p_1 = \cdots = p_6 & = 1/6 \\
f(x) & = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}
\end{align*}

\text{pdf, not probability}
Probability Basics, II

<table>
<thead>
<tr>
<th>Ex.</th>
<th>Ex.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expectation</strong></td>
<td><strong>Expectation</strong></td>
</tr>
<tr>
<td>$E(g) = \sum_{i=1}^{n} g(x_i)p_i$</td>
<td>$E(g) = \int g(x)f(x)dx$</td>
</tr>
<tr>
<td><strong>Population mean</strong></td>
<td><strong>Population mean</strong></td>
</tr>
<tr>
<td>$\mu = \sum_{i=1}^{n} x_i/p_i$</td>
<td>$\mu = \int x f(x)dx$</td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td><strong>Variance</strong></td>
</tr>
<tr>
<td>$\sigma^2 = \sum_{i=1}^{n} (x_i^2 - \mu^2)p_i$</td>
<td>$\sigma^2 = \int (x - \mu)^2 f(x)dx$</td>
</tr>
</tbody>
</table>

Sample

| mean | $\bar{x} = \sum_{i=1}^{n} x_i/n$ |
| variance | $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n$ |

Parameter Estimation

- Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution $f(x|\theta)$, estimate $\theta$.
- E.g.:
  $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
  $$\theta = (\mu, \sigma^2)$$

Maximum Likelihood Parameter Estimation

- One (of many) approaches to param. est.
- **Likelihood** of (indep) observations $x_1, x_2, ..., x_n$
  $$L(x_1, x_2, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta)$$
- As a function of $\theta$, what $\theta$ maximizes the likelihood of the data actually observed
- Typical approach: $\frac{\partial}{\partial \theta} L(\bar{x} | \theta) = 0$ or $\frac{\partial}{\partial \theta} \log L(\bar{x} | \theta) = 0$

Example 1

$n$ coin flips, $x_1, x_2, ..., x_n$; $n_0$ tails, $n_1$ heads, $n_0 + n_1 = n$;

$\theta$ = probability of heads

$$L(x_1, x_2, \ldots, x_n | \theta) = (1-\theta)^{n_0}\theta^{n_1}$$

$$\log L(x_1, x_2, \ldots, x_n | \theta) = n_0 \log(1-\theta) + n_1 \log \theta$$

$$\frac{\partial}{\partial \theta} \log L(x_1, x_2, \ldots, x_n | \theta) = \frac{-n_0}{1-\theta} + \frac{n_1}{\theta}$$

Setting to zero and solving:

$$\theta = \frac{n_1}{n}$$

(Also verify it's max, not min, & not better on boundary)
**Ex. 2:** 
\[ x_i \sim N(\mu, \sigma^2), \quad \sigma^2 = 1, \mu \text{ unknown} \]

\[
L(x_1, x_2, \ldots, x_n | \theta) = \prod_{1 \leq i \leq n} \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x_i - \theta)^2}{2\sigma^2}}
\]

\[
\ln L(x_1, x_2, \ldots, x_n | \theta) = \sum_{1 \leq i \leq n} -\frac{1}{2} \ln 2\pi - \frac{(x_i - \theta)^2}{2\sigma^2}
\]

\[
\frac{d}{d\theta} \ln L(x_1, x_2, \ldots, x_n | \theta) = \sum_{1 \leq i \leq n} (x_i - \theta)
\]

And verify it's max, not min & not better on boundary

\[
\hat{\theta} = \left( \sum_{1 \leq i \leq n} x_i \right) / n = \bar{x}
\]

**Ex. 3, (cont.)**

\[
\ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \leq i \leq n} -\frac{1}{2} \ln 2\pi \theta_2 - \frac{(x_i - \theta_1)^2}{2\theta_2}
\]

\[
\frac{\partial}{\partial \theta_1} \ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \leq i \leq n} (x_i - \theta_1) / \theta_2 = 0
\]

\[
\hat{\theta}_1 = \frac{\sum_{1 \leq i \leq n} x_i}{n} = \bar{x}
\]

\[
\frac{\partial}{\partial \theta_2} \ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \leq i \leq n} \frac{(x_i - \theta_1)^2}{2\theta_2^2} = 0
\]

\[
\hat{\theta}_2 = \frac{\sum_{1 \leq i \leq n} (x_i - \theta_1)^2}{n} = \bar{s}^2
\]

A consistent, but **biased** estimate of population variance.

(An example of **overfitting**) Unbiased estimate is:

\[
\hat{\theta}_2 = \sum_{1 \leq i \leq n} \frac{(x_i - \theta_1)^2}{n-1}
\]

**Moral:** MLE is a great idea, but not a magic bullet.

---

**EM**

The Expectation-Maximization Algorithm
More Complex Example

This?

Or this?

(A modeling decision, not a math problem...)

Gaussian Mixture Models / Model-based Clustering

Parameters \( \theta \)

- means: \( \mu_1, \mu_2 \)
- variances: \( \sigma_1^2, \sigma_2^2 \)
- mixing parameters: \( \tau_1, \tau_2 = 1 - \tau_1 \)

P.D.F.

\[
L(x_1, x_2, \ldots, x_n | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2) = \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)
\]

Likelihood Surface

\( x_i = \{-10.2, -10, -9.8, -0.2, 0, 0.2, 11.8, 12, 12.2\} \)

\( \sigma^2 = 1.0 \)

\( \tau_1 = 0.5 \)

\( \tau_2 = 0.5 \)
**A What-If Puzzle**

Likelihood

\[
L(x_1, x_2, \ldots, x_n | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2) = \Pi_{i=1}^{\tau} \sum_{j=1}^{\sigma_i} \tau_j f(x_i | \mu_j, \sigma_j^2)
\]

- Messy: no closed form solution known for finding \( \theta \) maximizing \( L \)
- But what if we knew the hidden data?

\[
z_{ij} = \begin{cases} 
1 & \text{if } x_i \text{ drawn from } f_j \\
0 & \text{otherwise}
\end{cases}
\]

**EM as Egg vs Chicken**

- If \( z_{ij} \) known, could estimate parameters \( \theta \)
- If parameters \( \theta \) known, could estimate \( z_{ij} \)
- But we know neither; (optimistically) iterate:
  - E: calculate expected \( z_{ij} \), given parameters
  - M: calc “MLE” of parameters, given \( E(z_{ij}) \)

**Simple Idea: “Classification EM”**

- If \( z_{ij} < .5 \), pretend it’s 0; \( z_{ij} > .5 \), pretend it’s 1
  i.e., classify points as component 0 or 1
- Now recalc \( \theta \), assuming that partition
- then recalc \( z_{ij} \), assuming that \( \theta \)
- then re-recalc \( \theta \), assuming new \( z_{ij} \)
- etc., etc.
**Full EM**

$x_i$'s are known; $\theta$ unknown. Goal is to find MLE $\theta$ of:

$$L(x_1, \ldots, x_n \mid \theta)$$

(hidden data likelihood)

Would be easy if $z_{ij}$'s were known, i.e., consider:

$$L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)$$

(complete data likelihood)

But $z_{ij}$'s aren't known.

Instead, maximize expected likelihood of visible data

$$E(L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)),$$

where expectation is over distribution of hidden data ($z_{ij}$'s)

---

**The E-step**

- Assume $\theta$ known & fixed
- A (B): the event that $x_i$ was drawn from $f_1$ ($f_2$)
- D: the observed datum $x_i$
- Expected value of $z_{i1}$ is $P(A \mid D)$

$$P(A \mid D) = \frac{P(D \mid A) \cdot P(A)}{P(D)}$$

$$P(D) = P(D \mid A) \cdot P(A) + P(D \mid B) \cdot P(B)$$

$$= f_1(x_i \mid \theta_1) \cdot \tau_1 + f_2(x_i \mid \theta_2) \cdot \tau_2$$

Repeat for each $x_i$

---

**Complete Data Likelihood**

Recall:

$$z_{ij} = \begin{cases} 
1 & \text{if } x_i \text{ drawn from } f_j \\
0 & \text{otherwise}
\end{cases}$$

so, correspondingly.

$$L(x_i, z_{ij} \mid \theta) = \begin{cases} 
\tau_1 f_1(x_i \mid \theta) & \text{if } z_{i1} = 1 \\
\tau_2 f_2(x_i \mid \theta) & \text{otherwise}
\end{cases}$$

Formulas with “if’s” are messy; can we blend more smoothly?

Yes, many possibilities. Idea 1:

$$L(x_i, z_{ij} \mid \theta) = z_{i1} \cdot \tau_1 f_1(x_i \mid \theta) + z_{i2} \cdot \tau_2 f_2(x_i \mid \theta)$$

Idea 2:

$$L(x_i, z_{ij} \mid \theta) = (\tau_1 f_1(x_i \mid \theta))^{z_{i1}} \cdot (\tau_2 f_2(x_i \mid \theta))^{z_{i2}}$$

---

**M-step Details**

(For simplicity, assume $\sigma_1 = \sigma_2 = \sigma_1 = \tau_2 = .5 = \tau$)

$$L(\vec{x}, \vec{z} \mid \theta) = \prod_{1 \leq i \leq n} \frac{\tau}{\sqrt{2\pi}\sigma^2} \exp \left( - \sum_{1 \leq i \leq 2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

$$E[\log L(\vec{x}, \vec{z} \mid \theta)] = E \left[ \sum_{1 \leq i \leq n} \left( \log \tau - \frac{1}{2} \log 2\pi\sigma^2 - \sum_{1 \leq j \leq 2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right) \right]$$

$$= \sum_{1 \leq i \leq n} \left( \log \tau - \frac{1}{2} \log 2\pi\sigma^2 - \sum_{1 \leq j \leq 2} E[z_{ij}] \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

Find $\theta$ maximizing this as before, using $E[z_{ij}]$ found in E-step. Result:

$$\mu_j = \sum_{i=1}^n E[z_{ij}] x_i / \sum_{i=1}^n E[z_{ij}]$$

(intuit: avg, weighted by subpop prob)
2 Component Mixture

$$\sigma_1 = \sigma_2 = 1; \ \tau = 0.5$$

<table>
<thead>
<tr>
<th>mu1</th>
<th>-20.00</th>
<th>-6.00</th>
<th>-5.00</th>
<th>-4.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu2</td>
<td>6.00</td>
<td>0.00</td>
<td>3.75</td>
<td>3.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x1</th>
<th>-6</th>
<th>z11</th>
<th>5.11E-12</th>
<th>1.00E+00</th>
<th>1.00E+00</th>
</tr>
</thead>
<tbody>
<tr>
<td>x2</td>
<td>-5</td>
<td>z21</td>
<td>2.61E-23</td>
<td>1.00E+00</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>x3</td>
<td>-4</td>
<td>x31</td>
<td>1.33E-34</td>
<td>9.98E-01</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>x4</td>
<td>0</td>
<td>z41</td>
<td>9.09E-80</td>
<td>1.52E-08</td>
<td>4.11E-03</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>x5</td>
<td>4</td>
<td>z51</td>
<td>6.19E-125</td>
<td>5.75E-19</td>
<td>2.64E-18</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>x6</td>
<td>5</td>
<td>z61</td>
<td>3.16E-136</td>
<td>1.43E-21</td>
<td>4.26E-22</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>x7</td>
<td>6</td>
<td>z71</td>
<td>1.62E-147</td>
<td>3.53E-24</td>
<td>6.69E-26</td>
</tr>
</tbody>
</table>
Gene Expression

- Recall a gene is a DNA sequence for a protein.
- To say a gene is expressed means that it:
  1. is transcribed from DNA to RNA
  2. the mRNA is processed in various ways
  3. is exported from the nucleus (eukaryotes)
  4. is translated into protein
- A key point: not all genes are expressed all the time, in all cells, or at equal levels.

Transcription

- RNA polymerase complex
  - E. coli: 5 proteins (2α, β, β′, σ)
    - σ is initiation factor; finds promoter, then released/replaced by elongation factors
  - Eukaryotes: 3 pols, each >10 subunits
- Attaches to DNA, melts helix, makes RNA copy (5′ → 3′) of template (3′ → 5′) at ~30nt/sec

5′ Processing: Capping

- methylated G added to 5′ end, and methyl added to ribose of 1st nucleotide of transcript
- probably helps distinguish protein-coding mRNAs from other RNA junk
- prevents degradation
- facilitates start of translation
### 3’ Processing: Poly A (Eukaryotes)

- Transcript cleaved after AAUAAA (roughly)
- pol keeps running (until it falls off) but no 5’ cap added to strand downstream of poly A site, so it’s rapidly degraded
- 10s - 100s of A’s added to 3’ end of transcript - its “poly A tail”

### More processing: Splicing

- Also in eukaryotes, most genes are spliced: protein coding exons are interrupted by non-coding introns, which are cut out & degraded, exons spliced together
- More details about this when we get to gene finding

### Nuclear Export

- In eukaryotes, mature mRNAs are actively transported out of the nucleus & ferried to specific destinations (e.g., mitochondria, ribosomes)
Regulation

- In most cells, pro- or eukaryote, easily a 10,000-fold difference between least- and most-highly expressed genes.
- Regulation happens at all steps. E.g., some transcripts can be sequestered then released, or rapidly degraded, some are weakly translated, some are very actively translated, some are highly transcribed, some are not transcribed at all.
- Below, focus on 1st step only: transcriptional regulation.

DNA Binding Proteins

A variety of DNA binding proteins ("transcription factors"; a significant fraction, perhaps 5-10%, of all human proteins) modulate transcription of protein coding genes.

Summary

- Learning from data:
  - MLE: Max Likelihood Estimators
  - EM: Expectation Maximization (MLE w/hidden data)
- Expression & regulation
  - Expression: creation of gene products
  - Regulation: when/where/how much of each gene product; complex and critical.
- Next: using MLE/EM to find regulatory motifs in biological sequence data.