Bayesian Inference

- Fit probability model to observed data
- Unknown parameters
  - Summarize using probability distribution
- Use prior information
  - Supply the prior
  - Elicit from data
Outline

• Distributions

• Bayes Rule

• Examples
  – Hemophelia
  – Breast Cancer
  – Nevi
Part I

- Joint, conditional and marginal distributions
- Posterior density and Bayes Rule
General Notations

• **Outcome**: $y$
  - Disease Status ($y = 0, 1$)
  - Tumor size, Number of nevi

• **Parameter**: $\theta$
  - Mutation carrier status ($\theta = 0, 1$)
  - Effect of treatment on tumor size

• **Covariate**: $X$
  - Treatment dosage, amount of sun exposure
Distributions

• **Model: joint probability distribution**
  - density of y and θ
  - $p(y, \theta) = p(\theta) \, p(y|\theta)$

• $p(\theta) = \text{prior density of } \theta$

• $p(y|\theta) = \text{conditional (sampling) density of } y \text{ given } \theta$
Example
(Gelman et al., 1995)

• Hemophelia:
  – Failure of blood to clot
  – X-chromosome linked recessive disorder

• Affected males
  – single copy of X chromosome

• Unaffected female
  – two copies of X chromosome (potential carrier)
Example

- A single male
  - $y = \text{hemophelia status}$
  - $\theta = \text{"bad" hemophelia gene carrier status}$

- $p(\theta) = \text{probability of "bad" hemophelia gene}$

- $p(y | \theta) = \text{probability of hemophelia given gene carrier status}$
Bayes Rule

- Obtain posterior density from prior and sampling densities.

\[ p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \]

\[ p(y) = \sum_\theta p(\theta)p(y|\theta) \quad \text{(when } \theta \text{ is discrete)} \]

\[ = \int_\theta p(\theta)p(y|\theta)d\theta \quad \text{(when } \theta \text{ is continuous)} \]

- \( p(y) \) = marginal probability of data
  - Average over all possible values of \( \theta \)
Example (Hemophelia)

- $\theta$ = carrier status in woman
- Find $p(\theta=1|y_1=0, y_2=0)$

\[
p(\theta = 0) = 0.5 = p(\theta = 1)
\]
\[
p(y_1 = 0, y_2 = 0|\theta = 1) = 0.5 \times 0.5 = 0.25
\]
\[
p(y_1 = 0, y_2 = 0|\theta = 0) = 1 \times 1 = 1
\]
\[
p(\theta = 1|y_1 = 0, y_2 = 0) = \frac{p(\theta = 1)p(y_1 = 0, y_2 = 0|\theta = 1)}{p(y_1 = 0, y_2 = 0)} = \frac{0.125}{0.625} = 0.20
\]
\[
p(y_1 = 0, y_2 = 0) = p(y_1 = 0, y_2 = 0|\theta = 0)p(\theta = 0) + p(y_1 = 0, y_2 = 0|\theta = 1)p(\theta = 1) = 0.625
\]
Example - Breast Cancer
(Satagopan et al., 2001)

- Case-control study

- Sample individuals with $y=1$ (case), and $y=0$ (control)

- Genotype for a BRCA mutation ($\theta=1$ if carrier, $0$ if non-carrier)

- Observed distribution: $p(\theta|y)$

- Want to estimate risk: $p(y=1|\theta=1)$
Example (Breast Cancer)

- **Odds ratio:**

\[
\phi = \frac{p(\theta = 1|y = 1)}{p(\theta = 1|y = 0)}
\]

- **Using Bayes rule:**

\[
p(y = 1|\theta = 1) = \frac{p(y = 1, \theta = 1)}{p(\theta = 1)} = \frac{p(\theta = 1|y = 1)p(y = 1)}{p(\theta = 1|y = 1)p(y = 1) + p(\theta = 1|y = 0)p(y = 0)} = \frac{\phi p(y = 1)}{\phi p(y = 1) + p(y = 0)}
\]
Example (Breast Cancer)

<table>
<thead>
<tr>
<th>BRCA Mutation</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrier</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>Non-carrier</td>
<td>179</td>
<td>1090</td>
</tr>
</tbody>
</table>

- \( p(\theta=1|y=1) = \frac{25}{204} = 0.1225 \)
- \( p(\theta=1|y=0) = \frac{23}{1113} = 0.0207 \)
- \( \Phi = \frac{0.1225}{0.0207} = 5.9 \)
- \( p(y=1) \): disease incidence in a given age group (40-49)
  - from SEER database
  - \( p(y=1) = 0.0138 \)
- \( p(y=0) = 1-p(y=1) \)
- \( p(y=1|\theta=1) = 7.6\% \)

- Age group 40-49
- Satagopan et al., (2001)
Summary

• Probability distributions
  – Sampling density of observed data
  – Probability density of parameter of interest

• Bayes rule helps determine posterior density of interest
Part II

- Probability models for data analysis
- Specifying prior and posterior densities
- Example:
  - Normal distribution
Bayesian Inference

• Fit model to data

• Specify prior density for unknown parameters

• Derive posterior density

• Summaries: posterior mean, variance, confidence interval ...
Studies of multiple risk factors

- Cancer epidemiology studies of multiple risk factors
  - Exposure/behavior: Questionnaires
  - Genetic polymorphisms/variants

- Thematic choice of putative risk factors
  - Sun exposure, sun burn, sun protection behavior
  - Variants in DNA repair, cell cycle control, drug action pathway

- Identify factors associated with disease risk
  - Exploit thematic choice of putative risk factors

- Also high-throughput studies
Some popular analytic methods

- **Regress outcome on the risk factors**
  - Ignores thematic choice of risk factors
  - Potential multi-collinearity issues
  - Unbiased estimates, may lack precision

- **Pre-defined linear combination of risk factors**
  - Pre-defined based on underlying theme
  - Estimate effects of the themes
  - Declare significance (or not) of individual risk factors
  - Trouble if weights are incorrect
Examples

- **Cumulative sun exposure**
  - Lancaster and Nelson 1957; Purdue et al 2005; Karagas et al 2007
  - Definition varies across studies: Elwood and Jopson 1997

- **Cumulative number of variant alleles**
  - Breast cancer: Smith et al 2003; Mechanic et al 2006; **Johnson et al 2007**
  - Skin cancer: Winsey et al 2000; Han et al 2004
  - Second primary melanoma: Millikan et al 2005
  - Bladder cancer: Wu et al 2006; Chen et al 2007; Yang et al 2008
  - Psychiatric disorders: Baum et al 2008

Low allele frequency
Analytic issues

• Pre-determined linear combinations
• Intuitive dimension reduction
• Potential precision gains

• Presumption: weights are accurate
• Biased estimates, false positive findings

• Remedy:
  – Empirical Bayes estimators
Motivating example

• Study of nevi in children
• Questionnaire: sun factors

• $Y = \log(\text{back nevus count})$
• $S = \text{matrix of confounders}$
• $X = \text{questionnaire data}$
• $N = 443 \text{ children}$

• $Y = Sa + Xb + \text{error}$
• Objectives:
  – Precise estimation of $b$
  – Find sun factors associated with $Y$

• Counter-intuitive findings
  – But, perhaps appropriate, reflect lifestyle or skin sensitivity

Oliveria et al 2009
Alternative analysis

• Thematic design of questionnaire
  – Note: correlations

• Sun exposure
  – Time spent at pool/beach, outdoor: weekdays, weekends

• Sunburn
  – Pain, redness of skin: back and other anatomic sites

• Sun protection behavior
  – Use of shirt, hat at pool/beach outdoors, sunscreen use
  – Limiting time in sun

• User-defined linear combinations
Alternative analysis

- **W = (W₁, W₂, W₃)**  
  = (sunexp, sunburn, prot.)
- **XW = linear combinations**
- **Y = Sₐ + (XW)d + error**
- **b_{red} = W_{d_{est}}**
- Sunburn significantly associated with high nevus counts
- Intuitively appealing
- Relatively narrow CIs
- Prior studies: not consistent results
Comparison of the estimates

- Two analyses, two conclusions
- How good is our W?
- Correlated Qx. responses
  - Redness/pain
  - Pool/outdoor time
  - Limited sun / shirt
  - Shirt / Hat @ beach
  - Shirt always / rare sunscreen usage in areas other than back

Correlation or small variance: Dimension reduction beneficial
Frank and Friedman 1993
Empirical Bayes approach

- $b = Wd + r$
- $r = 0$ when $W$ is accurate $\quad r$ stochastic around 0
- Plug-in estimate of $r$: $r_{est} = b_{full} - b_{red}$

- Prior distribution: $f(r) = N(0, t^2I)$
- Conditional distribution: $f(r_{est} \mid r) = N(r, \Sigma)$
- Posterior distribution: $f(r \mid r_{est}) = f(r_{est} \mid r) \ast f(r) / f(r_{est})$
  - Normal
  - mean = $E(r \mid r_{est}) = t^2(\Sigma + t^2I)^{-1}r_{est}$
  - variance = $\Sigma^{-1} + (1/t^2)I$

- $b_{EB} = b_{red} + E(r \mid r_{est}) = t^2(\Sigma+t^2I)^{-1}b_{full} + \Sigma(\Sigma+t^2I)^{-1}b_{red}$
Calculating the components

- $\Sigma = \text{Var}(b_{\text{full}} - b_{\text{red}})$
  
  $= \text{Var}(b_{\text{full}}) + \text{Var}(b_{\text{red}}) - 2 \text{Cov}(b_{\text{full}}, b_{\text{red}})$

  - Easy to write, readily obtained from output

- Estimate of $t^2 = (r_{\text{est}})^T (r_{\text{est}}) / p$

- Plug-in to get $b_{\text{EB}}$

- Parallels to
  - Penalized likelihood method
  - Ridge regression
Inference

• Need to assess significance of $b_{EB}$

• Get confidence intervals
  \[ b_{EB,j} \pm 1.96 \times \sqrt{\text{Var}(b_{EB,j})} \]

• Components with CI including 0 are declared not significantly associated with outcome

• Need variance to calculate confidence intervals
Variance Estimation

- \( b_{EB} = t^2(\Sigma+t^2I)^{-1}b_{full} + \Sigma(\Sigma+t^2I)^{-1}b_{red} \)
- \( b_{EB,true} = f(b, Wd) \)

- Taylor expansion of \( f(b, Wd) \) around \((b_{full}, b_{red})\)

- \( f(b, Wd) = f(b_{true}, Wd_{true}) + (\frac{\partial f}{\partial b} \quad \frac{\partial f}{\partial Wd}) (b-b_{full} \quad Wd-b_{red})^T \)
- \( \text{Var}(b_{EB}) = MRM^T \)

- \( M = \text{matrix of partial derivatives} \)
- \( R = \text{Covariance matrix of } (b_{full}, b_{red}) \)

Application to SONIC data:

Sign of $\text{EB} \approx \text{Sign of } b_{\text{full}}$

Narrow confidence intervals

$\text{EB}$ inference “similar” to $b_{\text{full}}$

Associations may potentially reflect lifestyle or skin sensitivity
Summary

• Specify distributions in a hierarchical manner

• Bayes or empirical Bayes solution

• Some problems: natural solution follows Bayes approach

• Rigorous variance calculation
  – Algebra may be a bit lengthy, but can be done
Some References


