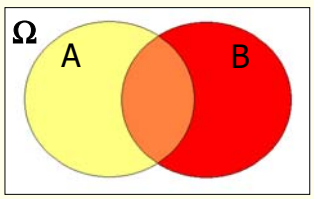
MIXTURE MODELS, BAYES AND EMPIRICAL BAYES APPROACHES

- A Sketch of Bayesian Inference
- Applications (Microarray Experiments)
  - Mixture model approach
  - Empirical Bayes methods
- Discussion

BAYESIAN INFERENCE AND MCMC

Conditional Probability (Bayes' Rule)



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B \mid A)}{P(B)}$$

Bayesian Inference

y: observed data

θ: parameters
(all unobserved quantities)

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

$$p(\theta \mid y) \propto p(\theta)p(y \mid \theta)$$

posterior distribution

prior distribution

sampling distribution

Marginal Posterior Distributions

$$p(\theta_1 \mid y) \propto \int_{\theta \neq \theta_1} p(\theta) p(y \mid \theta) d\theta_{\theta \neq \theta_1}$$

Gibbs Sampling

$$\theta = (\theta_1', \theta_2', \ldots, \theta_r')' \rightarrow p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_r)$$

$$\theta^{(0)} = (\theta_1^{(0)'}, \theta_2^{(0)'}, \ldots, \theta_r^{(0)'})'$$

$$\theta_1^{(1)} \mid \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_r^{(0)}$$
$$\theta_2^{(1)} \mid \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_r^{(0)}$$
$$\vdots$$
$$\theta_r^{(1)} \mid \theta_2^{(1)}, \theta_3^{(1)}, \ldots, \theta_{r-1}^{(1)}$$

- Burn-in & Convergence
- Tinning interval & Lag correlations
- Sample size & Monte Carlo error

NORMAL MIXTURE MODEL

(Pan et al., 2002)

$X_{1i}, \ldots, X_{mi}$ (gene i, m arrays under condition 1)

$Y_{1i}, \ldots, Y_{mi}$ (gene i, m arrays under condition 2)

$$X_{ji} = \mu_{(1)i} + \varepsilon_{ji}$$
$$Y_{li} = \mu_{(2)i} + e_{li}$$

$\Rightarrow$

- $\mu_{(1)i}$ and $\mu_{(2)i}$: mean expression levels for gene i under the two conditions
- $\varepsilon_{ji} \sim (0, \sigma^2_{(1)i})$, $e_{ji} \sim (0, \sigma^2_{(2)i})$: usually symmetric distributions

H_0: $\mu_{(1)i} = \mu_{(2)i}$

$$Z_i = \frac{\sum_{j=1}^{m} X_{ji}/m}{\sigma_{(1)i}} - \frac{\sum_{l=1}^{m} Y_{li}/m}{\sigma_{(2)i}} = \frac{\mu_{(1)i}}{\sigma_{(1)i}} - \frac{\mu_{(2)i}}{\sigma_{(2)i}} + \frac{\sum_{j=1}^{m} \varepsilon_{ji}}{m\sigma_{(1)i}} - \frac{\sum_{l=1}^{m} e_{li}}{m\sigma_{(2)i}}$$

$$E[Z_i] = \frac{\mu_{(1)i}}{\sigma_{(1)i}} - \frac{\mu_{(2)i}}{\sigma_{(2)i}} \quad \text{and} \quad Var[Z_i] = \frac{2}{m}$$

Estimating the Null Distribution (f_0)

Nonparametric method:

- Null score for each gene

$$z_i = \frac{X_{1i} - X_{2i} + \ldots + X_{m-1,i} - X_{mi}}{m\sigma_{(1)i}} + \frac{Y_{1i} - Y_{2i} + \ldots + Y_{m-1,i} - Y_{mi}}{m\sigma_{(2)i}}$$

$$= \frac{\varepsilon_{1i} - \varepsilon_{2i} + \ldots + \varepsilon_{m-1,i} - \varepsilon_{mi}}{m\sigma_{(1)i}} + \frac{e_{1i} - e_{2i} + \ldots + e_{m-1,i} - e_{mi}}{m\sigma_{(2)i}}$$

$$\begin{cases} \varepsilon_{ji} \text{ and } -\varepsilon_{ji} : \text{ same (symmetric) distribution} \\ e_{ji} \text{ and } -e_{ji} : \text{ same (symmetric) distribution} \end{cases}$$

⇨ z_i values across all genes are used to estimate f_0

EMPIRICAL BAYES (EB) APPROACH

(Newton and Kendziorski, 2003)

- Well-suited to high-dimensional inference problems (n x p)
- Borrowing of information across genes (p >> n)

  particularly information related to variability in the system
- Probability distributions are specified in several layers (accounting for multiple sources of variation).
- Methods provide estimates of differential expression, hypothesis tests and assessment of patterns of differential expression among multiple conditions, and rank ordering genes (cDNA or Affymetrix).
- Comparison of multiple groups: posterior probability of differential expression using a parametric mixture model.

➡ Canonical EB example

(Scaled) biological + other sources of variation

- Single array (J genes)
- X_j observed (measured) expression level ~ (μ_j, σ^2) (j = 1, 2, …, J)

True expression level of gene j

$\hat{\mu}_j = \mathrm{MLE}(\mu_j) = x_j$ Least variance (among unbiased estimators)

☞ These obvious estimates has some problems when J is relatively large (Stein, 1956)

Squared Euclidean length of the vector **x** $\sum_j x_j^2$ > Squared length of the target profile **μ**

☞ Better estimates of **μ** can be found…

Empirical Bayes argument

- Treat μ_j as random;
- Work out a summary feature of $p(\boldsymbol{\mu}|\mathbf{x})$;
- Plug in estimates of remaining parameters using the whole dataset.

e.g. $\hat{\mu}_j = wx_j + (1-w)\bar{x}$

Data-dependent weight $0 \le w \le 1$

$\bar{x} = \sum x_j / J$

Sharing information among genes

➡ General Model Structure: Two Conditions

- I samples
- $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jI})'$
- Probability distribution of $\mathbf{x}_j$?
- Baseline hypothesis: I samples are exchangeable, i.e., x_{ji} are independent random deviations from a gene specific mean value μ_j as $f_{obs}(.|\ \mu_j)$.

☞ Two groups (conditions) with s_k (k = 1, 2) samples each; $I = s_1 + s_2$

EE_j: equivalent expression for gene j (baseline hypothesis above holds)

DE_j: differential expression for gene j (μ_{j1} and μ_{j2})

Fold change: $\rho_j = \mu_{j1} / \mu_{j2}$

Instead of computing average of fold-change estimates from pairs of arrays or by taking the ratio of average expression from each group, EB estimates of ρ_j are obtained by specifying a probability distribution on μ_{jk} and then computing some measure $\hat{\rho}_j$ (central tendency) of the posterior distribution of ρ_j (given the data).

"Significance" of Differential Expression

- Classical one-gene-at-a-time tests treats gene effects as fixed and separate.
- Here, some unknown fraction p of the genes are DE and the remainder are EE, and the state for each gene is considered to be a matter of chance.

EE gene j $\rightarrow$ $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jI})$

$$f_0(\mathbf{x}_j) = \int \left(\prod_{i=1}^{I} f_{obs}(x_{ij} \mid \boldsymbol{\mu}) \right) \pi(\boldsymbol{\mu}) d\boldsymbol{\mu}$$ (predictive distribution)

(average over possible gene effects μ_j)

DE gene j $\rightarrow$ $\mathbf{x}_j = (\mathbf{x}_{j1}, \mathbf{x}_{j2})$

$$f_1(\mathbf{x}_j) = f_0(\mathbf{x}_{j1}) f_0(\mathbf{x}_{j2})$$

(both continuous mixing over the unknown values of μ_j and discrete over two patterns (DE_j and EE_j) for each gene)

☞ Marginal distribution of the data: $p f_1(\mathbf{x}_j) + (1-p) f_0(\mathbf{x}_j)$

☞ Posterior probability of differential expression (Bayes' rule):

$$\frac{p f_1(\mathbf{x}_j)}{p f_1(\mathbf{x}_j) + (1-p) f_0(\mathbf{x}_j)}$$

➡ Multiple Conditions

- 2 conditions → 2 patterns (EE_j and DE_j)
- 3 conditions → 5 possible patterns
  - (1) equivalent expression across 3 conditions
  - (3) altered expression in just one condition
  - (1) distinct expression in each condition
- 4 conditions → 15 different patterns
  - Patterns: grouping or clustering of conditions
  - Extra information might reduce the number of patterns to be considered

- $(m + 1)$ distinct patterns for $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jI})$
- Here, the marginal distribution of the data becomes: $\sum_{k=0}^{m} p_k f_k(\mathbf{x}_j)$ (mixing proportions)

☞ Posterior probability of expression pattern k: $\Pr(k \mid \mathbf{x}_j) \propto p_k f_k(\mathbf{x}_j)$

product of $f_0(.)$ contributions of the different groups or conditions

- $\Pr(k|\mathbf{x}_j)$, $k = 0, 1, \ldots, m$, can be used to identify genes with altered expression in at least one condition, to order genes within conditions, or to classify genes into distinct expression patterns.
- More generic concepts: pattern of increasing (decreasing) means over different conditions (e.g. time course experiment)

➡ The Gamma-Gamma Model

☞ Choice of observation component and mean component, each depending on a few additional parameters θ (to be estimated from the data)

Observation component: Gamma distribution

(shape parameter $\alpha > 0$ and scale parameter $\lambda = \alpha/\mu_j \rightarrow$ mean value μ_j)

$$f_{obs}(x \mid \mu_j) = \frac{\lambda^{\alpha} x^{\alpha-1} \exp\{-\lambda x\}}{\Gamma(\alpha)}, \quad x > 0$$

Note: $CV = 1/\sqrt{\alpha}$ constant across genes j.

Mean component: $\pi(\mu_j)$ ~ Inverse Gamma distribution

$$\alpha \text{ fixed} \;\rightarrow\; \lambda = \alpha/\mu_j \sim \text{Gamma}(\alpha_0, \nu)$$

shape α_0, scale ν

Parameters: $\theta = (\alpha, \alpha_0, \nu)$

$$f_0(x_1, x_2, \ldots, x_I) = C\frac{\left(\prod_{i=1}^{I} x_i\right)^{\alpha-1}}{\left(\gamma + \sum_{i=1}^{I} x_i\right)^{I\alpha+\alpha_0}} \quad , \text{ where } \quad C = \frac{\gamma^{\alpha_0}\Gamma(I\alpha+\alpha_0)}{\Gamma^I(\alpha)\Gamma(\alpha_0)}$$

➡ The Lognormal-Normal Model

common variance

$$\log(x) \sim \text{Normal}(\mu_j, \sigma^2)$$

Note: $CV = \sqrt{\exp(\sigma^2) - 1}$ (raw scale)

$$\mu_j \sim \text{Normal}(\mu_0, \tau_0^2)$$

$$f_0(\mathbf{x}) \sim MN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_n)$$

$$\boldsymbol{\mu}_0 = \begin{bmatrix} \mu_0 \\ \mu_0 \\ \vdots \\ \mu_0 \end{bmatrix} \qquad \boldsymbol{\Sigma}_n = \sigma^2\mathbf{I}_n + \tau_0^2\mathbf{M}_n \qquad \mathbf{M}_n = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

➡ Model Fitting

- Maximum (marginal) likelihood of θ

$$\begin{cases} GG: \theta = (\alpha, \alpha_0, \gamma) \\ LNN: \theta = (\mu_0, \sigma^2, \tau_0^2) \end{cases}$$

(EM Algorithm: any gene with at least one negative intensity value in one condition is omitted from this step of the calculation)

- With data x_j governed by a mixture as $\sum_{k=0}^{m} p_k f_k(\mathbf{x}_j)$, indicator variables (ϕ_{jl}) are defined as 1 if the expression pattern of gene j is pattern l, and 0 otherwise.
- The complete data log likelihood is:

$$l_C(\theta) = \sum_j \left\{ \sum_{k=0}^{m} \phi_{jk} [\log f_k(\mathbf{x}_j) + \log(p_k)] \right\}$$

- E-step: Using current estimate θ_0, replace ϕ_{jl} by $\hat{\phi}_{jl}$.
  ($\hat{\phi}_{jl}$: posterior probability of expression pattern i for gene j)
- M-step: Estimate p_k with the arithmetic mean $\hat{\phi}_{\cdot k}$.

☞ Initial values: $\alpha = (1/CV)^2$ (GG) and $\sigma^2 = \log(1 + CV^2)$ (LNN)

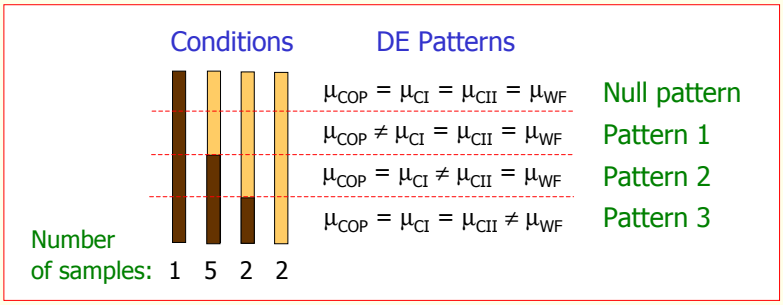
➡ Example:

- Mammary epithelial cells in rat (model of breast cancer)
- Affymetrix U34 chip (26,379 genes)
- I = 10 mRNA samples
- Four inbred strains (2 parental + 2 derived congenic lines)

☞ Parental
- Copenhagen (COP): almost completely resistant to certain carcinogens
- Wistar Furth (WF): highly susceptible

☞ Intermediate inbred lines
- CI and CII: homozygous for WF/WF or COP/COP in different (small) regions of the genome

After normalization/filtering:
J = 25,248 genes
(1131 outliers)



➡ Discussion:

- Classifying genes into expression patterns by the posterior probability is an optimal procedure in the context of the mixture model: it minimizes the expected number of errors.
- Interestingly, this goal is different from the goal in classical testing, which is to bound the type I error rate and then aim to maximize the power (Newton and Kendziorski, 2003).

BAYESIAN MODELS

(Ibrahim et al., 2002)

➡ Data:

- Affymetrix Hu6800; 7,000+ probe sets (≈ 5,600 distinct genes)
- 14 individuals (samples): 4 normal + 10 endometrial cancer tissues
- Filtering/normalization: 3,214 genes (final data set)

A: absent
M: marginal
P: present

A, M and negative values were set to $c_0 = 20$ (arbitrary value)

non expressed genes

Expression level x = c_0 with probability p; $c_0 + y$ with probability (1 – p)

➡ Model settings:

$$y_{jig} \sim \text{log normal} \begin{cases} i = 1, 2, \ldots, n_j & \text{(sample)} \\ j = 1, 2 & \text{(tissue type)} \\ g = 1, 2, \ldots, G & \text{(gene)} \end{cases}$$

$$p(y_{jig} \mid \mu_{jg}, \sigma^2_{jg}) = (2\pi)^{-1/2} y_{jig}^{-1} \sigma_{jg}^{-1} \exp\left\{-\frac{1}{2\sigma^2_{jg}}[\log(y_{jig}) - \mu_{jg}]^2\right\}$$

$$\delta_{ijg} = \begin{cases} 1 & \text{if } x_{jig} = c_0 \\ 0 & \text{otherwise} \end{cases} \qquad p_{jg} = \Pr(x_{jig} = c_0) \equiv \Pr(\delta_{jig} = 1)$$

$$L(\boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\delta}) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \prod_{g=1}^{G} p_{jg}^{\delta_{jig}} (1 - p_{jg})^{1-\delta_{jig}} p(y_{jig} \mid \mu_{jg}, \sigma^2_{jg})^{1-\delta_{jig}}$$

- Differentially expressed genes:

$$\begin{aligned} \psi_{jg} &= E_{\delta,y}[c_0 \delta_{jig} + (1 - \delta_{jig})(c_0 + y_{jig}) \mid p_{jg}, \mu_{jg}, \sigma^2_{jg})] \\ &= c_0 p_{jg} + (1 - p_{jg})[c_0 + E(y_{jig} \mid \mu_{jg}, \sigma^2_{jg})] \\ &= c_0 p_{jg} + (1 - p_{jg})\left[c_0 + \exp\left\{\mu_{jg} + \sigma^2_{jg}/2\right\}\right] \end{aligned}$$

- Posterior distribution of $\xi_g = \psi_{2g} / \psi_{1g}$ for comparing expression level means between normal and cancer tissues.
  - ☞ posterior mean, standard deviation, quantiles, and probabilities such as $\Pr(\xi_g > 1 \mid \mathbf{x}, \boldsymbol{\delta})$.

- Priors: $\mu_{jg} \mid \sigma^2_{jg}, \mu_{j0} \sim N(\mu_{j0}, \tau_0 \sigma^2_{jg} / \bar{n}_j)$

$$\tau_0 > 0 \qquad \bar{n}_j = \frac{1}{G} \sum_{g=1}^{G} \left(n_j - \sum_{i=1}^{n_j} \delta_{jig} \right)$$

$$\mu_{j0} \sim N(m_{j0}, v^2_{j0})$$

$$\sigma^2_{jg} \sim \text{Inv.Gamma}(a_{j0}, b_{j0}) \begin{cases} a_{j0} \quad \text{fixed} \\ b_{j0} \sim \text{Gamma}(q_{j0}, t_{j0}) \quad \text{(or fixed)} \end{cases}$$

- Prior correlation between μ_{jg}:

$$\begin{bmatrix} \mu_{jg} \\ \mu_{jg'} \end{bmatrix} \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) \rightarrow \boldsymbol{\mu}^* = \begin{bmatrix} m_{j0} \\ m_{j0} \end{bmatrix} \qquad \boldsymbol{\Sigma}^* = \begin{bmatrix} \frac{\tau_0 \sigma^2_{jg}}{\bar{n}_j} + v^2_{j0} & v^2_{j0} \\ v^2_{j0} & \frac{\tau_0 \sigma^2_{jg'}}{\bar{n}_j} + v^2_{j0} \end{bmatrix}$$

$$\text{corr}(\mu_{jg}, \mu_{jg'} \mid \sigma^2_{jg}, \sigma^2_{jg'}, v_{j0}) = \frac{v^2_{j0}}{\left\{ [\tau_0 \sigma^2_{jg} / \bar{n}_j + v^2_{j0}][\tau_0 \sigma^2_{jg'} / \bar{n}_j + v^2_{j0}] \right\}^{1/2}}$$

Note: If $v^2_{j0} \rightarrow \infty$ or $\bar{n}_j \rightarrow \infty \Rightarrow$ corr $\rightarrow 1$

$$e_{jg} = \text{logit}(p_{jg}) = \log\left(\frac{p_{jg}}{1 - p_{jg}} \right) \sim N(u_{j0}, k_{j0} w^2_{j0})$$

$$u_{j0} \sim N(\hat{u}_{j0}, h_{j0} w^2_{j0})$$

$$(k_{j0}, w_{j0}, h_{j0} : \text{ specified hyperparameters})$$

- Data-based guide values for hyperparameters:

$$m_{j0} = \frac{1}{N_j} \sum_{i=1}^{n_j} \sum_{g=1}^{G} (1 - \delta_{jig}) \log(y_{jig}) \quad \text{where} \quad N_j = \sum_{i=1}^{n_j} \sum_{g=1}^{G} (1 - \delta_{jig})$$

(gene sample mean on the natural logarithm scale for tissue type j)

$$v_{j0}^2 = \eta_{j0} MSG_j \quad \text{where} \quad MSG_j = \frac{1}{G-1}\sum_{g=1}^{G} n_{jg}(m_{jg0} - m_{j0})^2$$

$$\text{and} \quad m_{jg0} = \frac{\sum_{i=1}^{n_j}(1-\delta_{jig})\log(y_{jig})}{\sum_{i=1}^{n_j}(1-\delta_{jig})} \qquad n_{jg} = n_j - \sum_{i=1}^{n_j}\delta_{jig}$$

$$t_{j0}^{-1} = d_{j0} MSE_j \quad \text{where} \quad MSE_j = \frac{1}{N_j - G}\sum_{g=1}^{G}\sum_{i=1}^{n_j}(1-\delta_{jig})(\log(y_{jig}) - m_{jg0})^2$$

- Gene selection:

① Compute the posterior distributions of all ξ_g and $\gamma_g = Pr(\xi_g > 1/D)$.

② Select a threshold value γ_0 (possible values γ_0 = .7, .8, .9, and .95).

③ Once a set of genes is declared different, set the mean parameters for the tissue types to be unequal for that gene in the model.

④ Create several submodels using different values of γ_0 and using the step 3 and evaluate them (Bayesian criterion).

⑤ Select the best-fitting model.

REFERENCES

Ibrahim, J. G., Chen, M. H., Gray, R. J. (2002) Bayesian models for gene expression with DNA microarray data. Journal of the American Statistical Association 97: 88-99.

Newton, M. A. et al. (2001) On differential variability of expression ratios: Improving statistical inference about gene expression changes. Journal of Computational Biology 8: 37-52.

Newton, M. A, Kendziorski, C. (2003) Parametric Empirical Bayes Methods for Microarrays. In Parmigiani, G; E. S. Garrett; R. A. Irizarry and S. L. Zeger (Eds). The Analysis of Gene Expression Data: Methods and Software. Springer, New York. p. 254-271.

Pan, W., Lin, J., Lee, C. T. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biology 3(5):research0022.1-0022.10.