Statistical Issues on Genomic Composite Biomarker Classifiers

Jen-pei Liu[1,2] and Shein-Chung Chow[3]

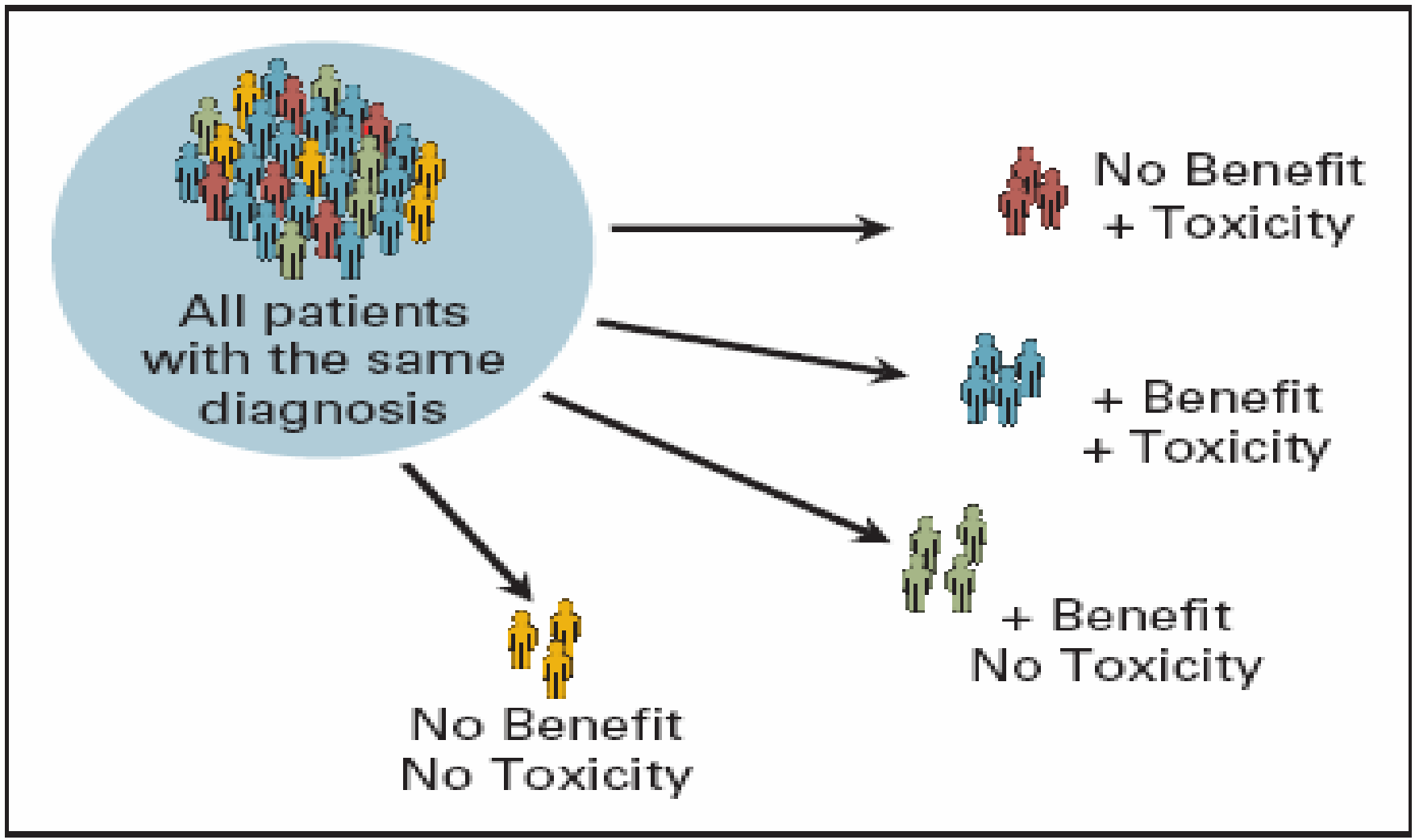
[1] Division of Biometry, Department of Agronomy
National Taiwan University, Taipei, Taiwan
[2] Division of Biostatistics and Bioinformatics
National Health Research Institutes, Zhunan, Taiwan
[3] Department of Biostatistics and Bioinformatics
Duke University School of Medicine
Durham, NC 27706, U.S.A

At
2006 Joint Statistical Meeting
Seattle, Washington, U.S.A.
August, 2006

Outlines

- Introduction
- Selection of Genes and Representation
- Distribution
- Agreement and Reproducibility
- Estimation of Treatment Effects in Targeted Clinical Trials
- Discussion and Summary

Introduction



Introduction

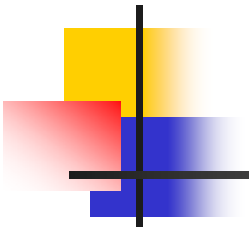
- Post HGP (Human Genome Project) Era
- Pharmacogentics and Pharmacogenomics
- Biochip Products
- Target Clinical Trials
- Personalized Medicine
- Diagnosis and Individualized Treatment
  - Drug-diagnostic Co-development (FDA, 2005)

Introduction

- TAILORx (JCO, 2006)
- Patients
  - 10,000 patients with early-stage breast cancer
  - ER and/or PR+ and Her2/new – and not spread to lymph nodes
- Diagnostic Device
  - Likelihood of distant recurrence
  - Based on 21 prospectively selected genes in paraffin-embedded tumor tissue
  - RT-PCR assay
  - Recurrence score from 0 to 100

Introduction

- TAILORx (JCO, 2006)
- Treatment
  - Recurrence scores
    - >25: chemotherapy + hormonal therapy
    - <11: hormonal therapy
    - 11 – 25: randomization
      - Standardized combination chemotherapy + adjuvant hormonal therapy
      - Adjuvant hormonal therapy
- Follow-up: 10 years additional 20 years after initial therapy



Introduction

- Issues in GCB classifiers
  - Selection of genes
  - Functional form for the overall representation of expression levels
  - Distribution of GCB classifiers and determination of thresholds
  - Evaluation of agreement and reproducibility for GCB classifiers
  - Estimation of treatment effects in targeted clinical trials based on the diagnostic results from GCB classifiers

Selection of Genes and Representation

- Differentially expression genes for diagnosis of molecular targets
  - Current hypothesis

    H_o: $\mu_{Ti} - \mu_{Ci} = 0$ and H_a: $\mu_{Ti} - \mu_{Ci} \neq 0$, i=1,…,G
  - Statistical significance based on hypothesis of difference does not take into consideration of magnitude of levels of differentially expressed genes and their biological significance
  - Fold change does not take into consideration of statistical significance

Selection of Genes and Representation

- Statistical hypothesis for identification of differentially expression genes should take into consideration biological significance

H_o: $\mu_{Ti} - \mu_{Ci} \geq \delta_{Li}$ and $\mu_{Ti} - \mu_{Ci} \leq \delta_{Ui}$

Vs.

H_A: $\mu_{Ti} - \mu_{Ci} < \delta_{Li}$ and $\mu_{Ti} - \mu_{Ci} > \delta_{Ui}$

i=1, …, G; and assume that each gene has its own differentially expressed level

Selection of Genes and Representation

A two one-sided test

$$T_L = \frac{\overline{Y}_{Ti} - \overline{Y}_{Ci} - \delta_{Li}}{\sqrt{s_i^2 \left(\frac{1}{n_{Ti}} + \frac{1}{n_{Ci}}\right)}}$$

and

$$T_L = \frac{\overline{Y}_{Ti} - \overline{Y}_{Ci} - \delta_{Ui}}{\sqrt{s_i^2 \left(\frac{1}{n_{Ti}} + \frac{1}{n_{Ci}}\right)}}$$

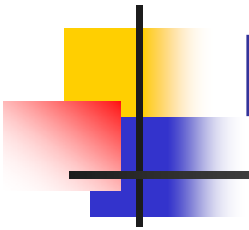
Gene i is claimed to be differentially expressed at the α significance level if $T > t(\alpha, n_{Ti} + n_{Ci} - 2)$ or $T < t(\alpha, n_{Ti} + n_{Ci} - 2)$, $i = 1,...,G$

Selection of Genes and Representation

- The average type I error rate is controlled at the nominal level
- The power function is a parabola and symmetric and reaches the minimum at $(\delta_{Li} + \delta_{Ui})/2$
- Simulation studies was conducted to compare with current methods
  - Unadjusted, Bonferroni adjustment, fold changes only, fold changes with p-value < 0.04 (Affymetrix)

Selection of Genes and Representation

- Functional Form
  - Differentially expressed genes
    - Over-expressed in T and under-expressed in C
    - Under-expressed in T and over-expressed in C
  - Representation
    - Difference of expression levels of differentially expressed genes between the test and control
    - Ratio of expression levels of differentially expressed genes between the test and control



Distributions

- Genomic Composite Biomarker (GCB) Classifier
  - The number of differentially expressed gene for diagnosis of certain molecular targets is a random variables
  - The number of genes in GCB classifier is a random variable
  - The expression level for each selected gene in GCB classifier is also a random variable

Distributions

- Only consider the linear function
  - $X = w_1Y_1 + w_2Y_2 + \ldots + w_gX_g$,
  - g is the number of differentially expressed genes selected from a pool of a total of G genes
  - Y_i is the expression level of gene i based on log 2
  - w_i is the weight for gene i

Distributions

- Suppose all weights are equal and Yi are i.i.d. with mean μ and variance σ^2 and the number of differentially expressed gene follows a Poisson distribution with mean λ
- The distribution of X can be expressed by convolution

$$F(x) = \sum_{g=0}^{\infty} p_g . S^{g*}(x),$$

where the probability that the number of differentially expressed genes is g, and

$$S^{g*}(x) = P(\sum_{i=0}^{g} Y_i \le x)$$

Distributions

- $E(X) = \lambda\mu$
- $Var(X) = \lambda\sigma^2 + \lambda\mu^2$
- Asymptotic normality

$$\text{Let } \overline{Y} = \sum_{i=1}^{g} Y_i, \text{ and } s^2 = \sum_{i=1}^{g} (Y_i - \overline{Y})^2$$

An estimate of variance of X is given as

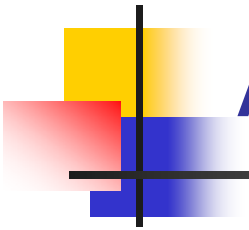
$$\widehat{Var}(X) = m\overline{Y} + ms^2$$

$$Z = \frac{X\text{-}E(X)}{\sqrt{\widehat{Var}(X)}} \rightarrow N(0,1)$$

Agreement and Reproducibility

- Only recognition and evaluation of agreement and reproducibility:
  - Dobbin et al. (Clinical Cancer Research, 2005)
  - Larkin et al. (Nature Methods, 2005)
  - Irizarry et al. (Nature Methods, 2005)
  - Members of the Toxicogenomic Research Consortium (Nature Methods, 2005)
  - Tan, et al. (Nucleic Acids Research, 2003)
  - Yauk, et al. (Nucleic Acids Research, 2004)

All used correlation coefficient for evaluation of agreement and reproducibility



Agreement and Reproducibility

- Correlation coefficient
  - A measure for association
  - Not a measure for similarity (or agreement)
- Euclidean distance
  - A measure for agreement
  - Not a measure for association

Agreement and Reproducibility

- Example

Case I		Case II		Case III	
X1	X2	X1	X2	X1	X2
1	1	1	2	1	4
2	2	2	4	2	8
3	3	3	6	3	12
4	4	4	8	4	16
$r=1, d^2=0$		$r=1, d^2=30$		$r=1, d^2=270$	

Agreement and Reproducibility

- Hypothesis of no correlation can not prove agreement not reproducibility
- With 5000 genes, a correlation of 0.05 is statistically significant from 0 at 1% level
- Use of concordance correlation coefficient for evaluation of agreement (Lin, 1989)

Agreement and Reproducibility

We want to evaluate agreement of expression levels of two replicates of genes,

i.e., $Y_{1i} = Y_{2i}$, i=1,...,G, and

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right]$$

Concordance Correlation Coefficient (CCC)

$$\rho = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

The sample estimate is given as

$$r = \frac{2s_{12}}{s_1^2 + s_2^2 + (\overline{Y}_1 - \overline{Y}_2)^2}$$

Agreement and Reproducibility

- Hypothesis for Agreement
- H_O: $\rho \leq \rho_a$ vs. H_O: $\rho > \rho_a$, where ρ_a is the minimal required level of agreement
- Reject H_O and conclude agreement if the lower limit of the (1-α)% asymptotic C.I. is greater than ρ_a
- Use the concept of generalized pivotal quantity (GPQ) to construct the exact C.I.

Agreement and Reproducibility

$$U_2 \sim \chi^2(G-1),\ U_{1.2} \sim \chi^2(G-2),$$

$Z_1 \sim N(0,1)$, and $Z_2 \sim N(0,1)$;

$U_2, U_{1.2}, Z_1, Z_2$ are independent

$$\text{Define } R_2 = \frac{s_2}{U_2},$$

$$R_{12} = \frac{s_{12}}{U_{12}} - [\sqrt{s_{1.2}s_2}\, x \frac{Z_1}{\sqrt{U_{1.2}}}\, x \frac{1}{U_{22}}],$$

$$R_1 = \frac{s_{1.2}}{U_{1.2}} + \frac{R_{12}^2}{R_2},$$

Agreement and Reproducibility

$$R_{\mu_1-\mu_2} = (\bar{y}_1 - \bar{y}_2) - Z_2\sqrt{\frac{R_1}{G} + \frac{R_2}{G} - \frac{2R_{12}}{G}},$$

$$s_{1.2} = s_1^2 - \frac{s_{12}^2}{s_2^2}.$$

A GPQ for CCC is given as

$$R = \frac{2R_{12}}{R_1 + R_2 + R_{\mu_1-\mu_2}^2}$$

A (1-α)% C.I. can be obtained by Monte Carlo method

Agreement and Reproducibility

- Reproducibility
  - Dobbin et al. (Clinical Cancer Research, 2005)
  - Reproducibility of 4 different labs based on Affymetrix Human Genome U133A array
  - Each of 12 tumor tissue was divided into 6 blocks
  - Each laboratory got one or two blocks

Agreement and Reproducibility

Site 1	Site 2	Site 3	Site 4
AABBCCD	AABBCDE	ABCCDDF	ABCDEEF
DEEFFGHI	FGGHHIIJ	GGHHIJK	FGHIIJJK
JKL	JKL	KLL	LL

The underlined sample was failed sample

Agreement and Reproducibility

- Source of variation
  - Between laboratory (α_i)
  - Between samples (β_j)
  - Imbalance
- Model: Two-way classification random-effects model without interaction
  - $Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$,
    $i = 1,\ldots,a; j = 1,\ldots,b; k=1,\ldots n_{ij}$;
    where μ is an unknown constant, $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\beta_j \sim N(0, \sigma_\beta^2)$, and $e_{ijk} \sim N(0, \sigma_e^2)$, and α_i, β_j, and e_{ijk} are mutually independent

Agreement and Reproducibility

- Parameters of Reproducibility

$$\rho_{\alpha} = \frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2 + \sigma_{\beta}^2 + \sigma_{e}^2}$$

- Hypothesis of Reproducibility

  H_O: $\rho_{\alpha} \leq \rho_{\alpha 0}$ vs. H_A: $\rho_{\alpha} > \rho_{\alpha 0}$,

  where $\rho_{\alpha 0}$ is the minimal required level for reproducibility

Agreement and Reproducibility

- Reject the null hypothesis if the upper limit of the (1-α)% C.I. for ρ_α is greater than $\rho_{\alpha 0}$
- For imbalanced case and small sample size, apply the method of generalized pivotal quantity (GPQ) to obtain the exact C.I.

Agreement and Reproducibility

Application of GPQ to obtain the (1-α)% C.I. for ρ_α

The generalized pivotal quantity for ρ_α is given as

$$T(Y,y,\xi)=\frac{A}{A+B+C},$$

where $$A=\max[0,(\frac{1}{b})(\frac{ss(\alpha)}{U_A}-\lambda_{max}(\mathbf{D}_0)\frac{ss_2(e)}{U_E})],$$

$$B=\max[0,(\frac{1}{a})(\frac{ss(\beta)}{U_B}-\lambda_{max}(\mathbf{D}_0)\frac{ss_2(e)}{U_E})], \text{ and}$$

$$C=\lambda_{max}(\mathbf{D}_0)\frac{ss_2(e)}{U_E},$$

Agreement and Reproducibility

$U_A \sim \chi^2(a-1),\ U_B \sim \chi^2(b-1),\ U_E \sim \chi^2(n-ab-a-b+2),$

P is a abxab orthogonal matrix with the first row being

$\mathbf{1}'/ab$ such that

$$\mathbf{P}(\mathbf{I}_a \otimes \frac{1}{b}\mathbf{J}_b)\mathbf{P}' = \mathrm{diag}(1, \mathbf{I}_{a-1}, \mathbf{0}, \mathbf{0}) \text{ and}$$

$$\mathbf{P}(\frac{1}{b}\mathbf{J}_b \otimes \mathbf{I}_a)\mathbf{P}' = \mathrm{diag}(1, \mathbf{0}, \mathbf{I}_{a-1}, \mathbf{0});$$

$\mathbf{D}_0$ is a (a+b-2)x(a+b-2) submatrix of **P** corresponding

to $\mathbf{z}_\alpha$ and $\mathbf{z}_\beta$, the orthogonal transformation of vector of cell means;

$\lambda_{max}(\mathbf{D}_0)$ is the largest eigenvalue of $\mathbf{D}_0$

Agreement and Reproducibility

$$ss_0(e) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij.})^2 = \mathbf{y}'\mathbf{H}\mathbf{y},$$

$$\mathbf{H} = \mathbf{I}_n - \text{diag}\{\mathbf{J}_{n_{ij}} / n_{ij}\}$$

There exists an orthogonal matrix such that

$$\mathbf{H} = \mathbf{G}\text{diag}(\mathbf{I}_{n\text{-}ab}, \mathbf{0})\mathbf{G}'$$

$$= (\mathbf{G}_1 : \mathbf{G}_2 : \mathbf{G}_3)\text{diag}(\mathbf{I}_{a+b-2}, \mathbf{I}_{n\text{-}ab\text{-}a\text{-}b+2}, \mathbf{0})(\mathbf{G}_1 : \mathbf{G}_2 : \mathbf{G}_3)'$$

$$= \mathbf{G}_1\mathbf{G}_1' + \mathbf{G}_2\mathbf{G}_2';$$

$$ss_0(e) = ss_1(e) + ss_2(e) = \mathbf{y}'\mathbf{G}_1\mathbf{G}_1'\mathbf{y} + \mathbf{y}'\mathbf{G}_2\mathbf{G}_2'\mathbf{y};$$

$$\mathbf{w} = (\mathbf{w}_\alpha', \mathbf{w}_\beta') = (\mathbf{z}_\alpha', \mathbf{z}_\beta')' + [\lambda_{\max}(\mathbf{D}_0)\mathbf{I}_{a+b-2} - \mathbf{D}_0]^{1/2}\mathbf{G}_1'\mathbf{y};$$

$$ss(\alpha) = \mathbf{w}_\alpha'\mathbf{w}_\alpha \text{ and } ss(\beta) = \mathbf{w}_\beta'\mathbf{w}_\beta$$

Treatment Effects for Targeted Clinical Trial

- Enrichment design for targeted clinical trials
  - Patients with positive diagnosis for the molecular targets were randomized into the test drug or control group
  - The diagnostic device for the molecular targets is not a perfect device
  - Patients with positive diagnosis may not have the molecular targets – false positive rate (FPR)

Treatment Effects for Targeted Clinical Trial

- Enrichment design for targeted clinical trials
  - The objective of targeted clinical trials is to estimate the treatment effect of the test drug
  - Due to the FPR, the observed mean difference between the test and control groups under-estimate the true treatment effect

Treatment Effects for Targeted Clinical Trial

- Enrichment design for targeted clinical trials
  - The expected value of mean difference

$$\mathrm{E}(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}}) = \mathrm{PPV}(\mu_{+\mathrm{T}} - \mu_{+\mathrm{C}}) + \mathrm{FP}(\mu_{-\mathrm{T}} - \mu_{-\mathrm{C}})$$

  - Under-estimation of the true treatment effect becomes more severe as the prevalence rate of molecular targets decreases

Treatment Effects under Targeted Clinical Trials

- The true status of molecular targets for each patient is not available
- The positive predictive value can be estimated from the clinical validation trials for the medical diagnostic device
- Under normal assumption, the EM algorithm can be applied to estimate the true treatment effect of the test drug for the patients with molecular targets by assuming the PPV is a constant

Treatment Effects under Targeted Clinical Trials

Let Y_{ij} be the observation of patient j receiving treatment i; i=T,C; j=1,...,n_i

Let π_{ij} be the indicator for molecular target

$\pi_{ij} = 1(0)$ if patient j in treatment group i has the molecular target

$$\pi_{ij} \overset{i.i.d}{\sim} \text{bin(PPV)}$$

Treatment Effects under Targeted Clinical Trials

Under normal assumption, given a value π_{ij}, for treatment i, y_{ij} has a normal distribution with density

$$f(y_{ij}1\pi_{ij}) \propto \frac{1}{\sigma}\exp\{-\frac{1}{\sigma^2}[\pi_{ij}(y_{ij}-\mu_{+i})^2+(1-\pi_{ij})(y_{ij}-\mu_{-i})^2] \quad (a)$$

The conditional probability π_{ij} given y_{ij}

$$P(\pi_{ij}1y_{ij}) = 1/\{1+\exp[(\mu_{+i}-\mu_{-i})(\mu_{+i}+\mu_{-i}-2y_{ij})/2\sigma^2]\} \quad (b)$$

The log-likelihood

$$l=n\log\sigma\{\sum_{j=1}^{n_i}[\pi_{ij}(y_{ij}-\mu_{+i})^2+(1-\pi_{ij})(y_{ij}-\mu_{-i})^2]\}/2\sigma^2 \quad (c)$$

Treatment Effects under Targeted Clinical Trials

Procedure

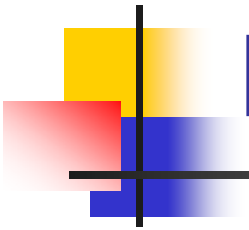
(1) E step: substitute "current" estimates of μ_{+T}, μ_{-T}, $\mu_{+C}, \mu_{-C}, \sigma^2$ into (b) to obtain provisional values for the expectation of π_{ij}

(2) M Step: Obtain the MLE of μ_{+T}, μ_{-T}, $\mu_{+C}, \mu_{-C}, \sigma^2$after replacing π_{ij} in (c)

(3) Repeat (1) and (2) until the estimates of $\mu_{+T} - \mu_{+C}$ and $\mu_{-T} - \mu_{-C}$ stabilize

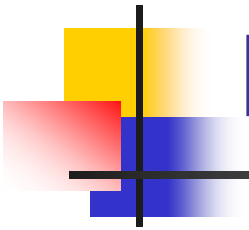
Discussion and Summary

- Issues in evaluation of quality and utility of GCB classifiers
- Identification of differentially expressed genes
- Distribution of GCB classifiers
- Agreement and reproducibility
- Estimation of treatment effects for targeted clinical trials



Discussion and Summary

- Identification of differentially expressed genes
  - Determination of thresholds
  - Definition of type I error
  - Sample size estimation
  - Optimal functional form for an overall summary representation



Discussion and Summary

- Distribution of GCB classifiers
  - Correlated expression levels among genes
  - Expression levels are not identically distributed
  - Estimation of weights
  - Exact distribution
  - Determination of decision thresholds and evaluation of their systematic bias
  - Masking effects of individual genes

Discussion and Summary

- Agreement and Reproducibility
  - Determination of minimal required threshold
  - Correlation of expression levels among genes
- Estimation of Treatment Effects
  - Two or more molecular targets for different pathways
  - Variability of the estimated positive predicted values
  - Extension to binary and survival endpoints