Reproducible Analysis of High-Throughput Experiments

Ying Liu

Department of Biostatistics, Columbia University

Summer Intern at Research and CMC Biostats, Sanofi, Boston

August 26, 2015

Outline

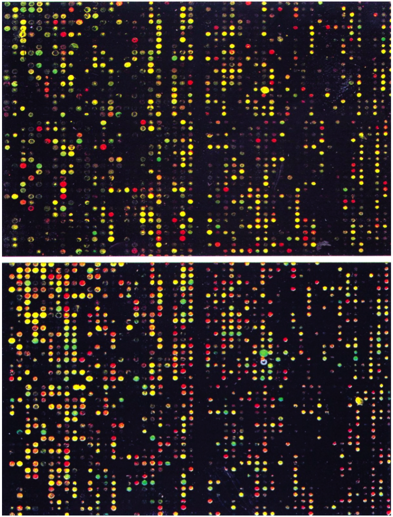
1. Introduction
   - Study Goal
   - Multiple Testing and Meta Analysis
   - Copula Mixture Model
2. Proposed Method: A Bayesian Classification Approach
   - Mixed Effect Model for Data Generation
   - Proposed Method: Mixture Gaussian Model for T-statistics
3. Simulation Results
4. Real Data Analysis

OUTLINE

1. Introduction
   - Study Goal
   - Multiple Testing and Meta Analysis
   - Copula Mixture Model
2. Proposed Method: A Bayesian Classification Approach
   - Mixed Effect Model for Data Generation
   - Proposed Method: Mixture Gaussian Model for T-statistics
3. Simulation Results
4. Real Data Analysis

MOTIVATION: HIGH-THROUGHPUT ARRAY ANALYSIS

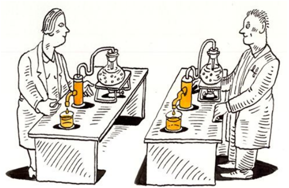
Finding differentially expressed Genes in normal and disease biological processes in high-throughput array data.



PROBLEM TO SOLVE: REPRODUCIBILITY

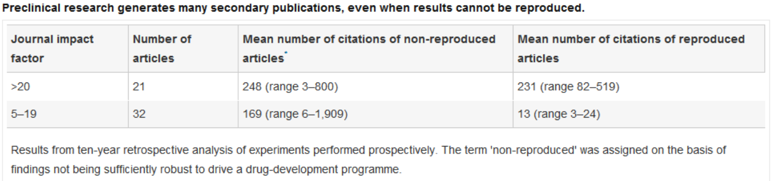
Finding Genes that are differentially expressed in 2 Studies for the same disease.

- Comparable Array Type
- Two human studies
- Animal and Human studies: finding the genes that differentially expressed in both animal and human.



REPRODUCIBILITY

- Reproducibility is one of the main principles of the scientific method.
- Reproducibility is the ability of an entire experiment or study to be duplicated.
- *"By repeating the same experiment over and over again, the certainty of fact will emerge."- Robert Boyle*
- Robust and reliable findings across studies.
- Begley, C. G.; Ellis, L. M. (2012). "Drug development: Raise standards for preclinical cancer research". Nature 2012



Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.

MULTIPLE TESTING SOLUTIONS

- H_0: There is no association between gene and disease.
- Test H_0 for tens of thousands of genes.
- need to consider Multiple testing

Hypothesis	Accept Null	Reject Null
Null True	TN	FP
Alternative True	FN	TP

FWER

- 'Family-wise error rate': the probability of rejecting at least one of the true H_0
- FWER$= P(FP \geq 1)$
- Example: Bonferroni
- Limitations: too stringent for high throughput data.

SINGLE ARRAY MULTIPLE TESTING SOLUTIONS

Hypothesis	Accept Null	Reject Null
Null True	TN	FP
Alternative True	FN	TP

- FDR 'false discovery rate': the proportion of incorrect rejections out of all the rejections.
- FDR$= E\left(\frac{FP}{max((FP+TP),1)}\right)$
- Benjamini Hochberg Procedure (1995): controling the FDR $\leq \alpha$.
- $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ are ordered p-values. Let $k = \max\{1 \leq i \leq m : P_{(i)} \leq i\alpha/m\}$.
- Reject hypothesis corresponding to $P_{(i)}$ for $i = 1, \ldots, k$.

COMBINING P-VALUE FROM MULTIPLE STUDIES: REVIEW OF META ANALYSIS METHODS

Fisher's Combined Probability Test:

- p-values from study 1 and 2 for gene g: p_1, p_2
- For each gene g, $\chi^2_g = -2(\log p_1 + \log p_2)$.
- H_0: no association in both studies
- The combined p-value is $P(\chi^2_g > \chi^2(4))$

A similar approach: Stouffer's Z score:

$$Z = \frac{\sum_{i=1}^{K} \Phi^{-1}(1-p_{ki})}{\sqrt{K}} \sim \mathcal{N}(0,1)$$

DIFFERENCE BETWEEN REPRODUCIBILITY ANALYSIS AND META ANALYSIS

'Replicability Analysis for Genome-wide Association Studies', Ruth Heller and Daniel Yekutieli *AOAS* 2014

- Meta-analysis (as in Fisher's combined t-test): discover the associations that are present in at least one study
- Replicability (reproducibility) analysis: discover replicated associations in both studies

MULTI-DIMENSIONAL REPRODUCIBLE MULTIPLE TESTING PROBLEMS

Partial Conjunction Approach (Benjamini & Heller (2008,2009))

- Rejecting Null $H_0^{2/2}(g)$: there is significant replicated associations in both studies
- Partial conjunction null $H_0^{2/2}(g)$: there is no association in ANY one of the two studies.
- When p-value across studies are independent, partial conjunction Fisher p-values $p^{2/2}(g) = P(\chi_2^2 \geq -2 \log p_{\max}(g))$
- The Stouffer's Z score in this case is $\Phi^{-1}(1 - p_{\max})$, same p-value as in Fisher's.
- Use the BH procedure on the partial conjunction p-values

Summarize: Study Goal

Finding differentially expressed gene patterns among patients and control in multiple studies.

- Find and prioritize genes that are both
  - Differentially expressed between patients and controls
  - Reproducible across studies
- Need to consider
  - Multiple Testing Problem
  - Replicability Analysis

COPULA MIXTURE MODEL: AN EXISTING APPROACH

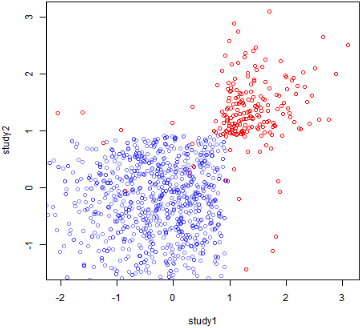
- 'Measuring Reproducibility of High-throughput Experiments' AOAS Quanhua Li, et.al. 2011.
- (X_1, X_2) could be pair of significance score: for example p-values
- The empirical marginal distribution function $\hat{F}_{1i} = \frac{1}{1+n}\sum_j I(X_{1j} \leq X_{1i})$, $\hat{F}_{2i} = \frac{1}{1+n}\sum_j I(X_{2j} \leq X_{2i})$, satisfies uniform distribution.
- $(Z_{1i}, Z_{2i}) = (\phi^{-1}(\hat{F}_{1i}), \phi^{-1}(\hat{F}_{2i}))$ follows a mixture normal distribution.

COPULA MIXTURE MODEL

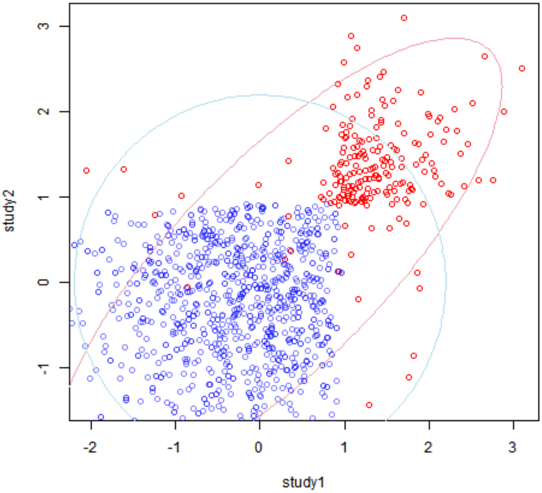
$$p(Z_1, Z_2) = \pi_1 \mathcal{N}((a, b); \Sigma_1) + \pi_2 \mathcal{N}((0, 0), \Sigma_2)$$

Here we assume there are two clusters of genes.

$$\Sigma_1 = \begin{bmatrix} \sigma_1^2 & \sigma_1^2\rho \\ \sigma_1^2\rho & \sigma_1^2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \pi_1 + \pi_2 = 1.$$



We can get the maximum likelihood estimation from EM algorithm similar to proposed method

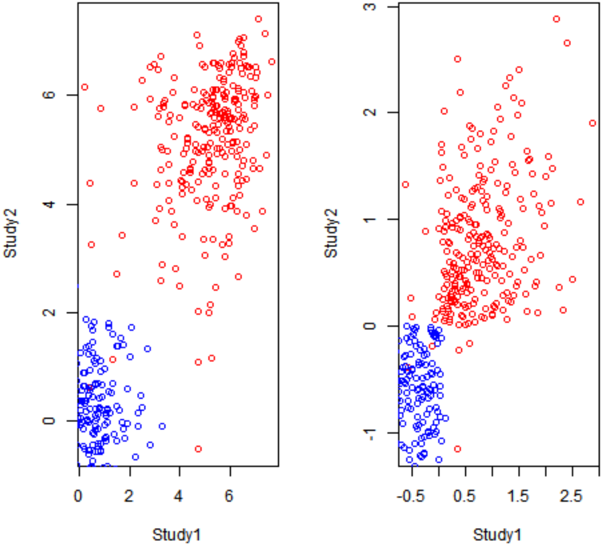


COPULA MIXTURE MODEL: PICK GENES

- The local irreproducible discovery rate: $idr(z_{i,1}, z_{i,2}) = P(i \in \text{Irreproducible group} | z_{i,1}, z_{i,2}) = \frac{\pi_2 f_2}{\pi_1 f_1 + \pi_2 f_2}$
- Rejection Region: $R_\gamma = \{(z_{i,1}, z_{i,2}) : idr(z_{i,1}, z_{i,2}) < \gamma\}$

COPULA MIXTURE MODEL: LIMITATIONS

- The significance scores omit the directions of effects. Pick genes with small p-value in both studies but different effect signs.
- The model assumes only small proportion of genes are significant; when a large proportion of p-values are small, the significant p-values could have medium/ large rank and cannot be differentiated from others.
- It could pick largest p-values (in the real data analysis)


Study2
6
4
2
0
0
2
4
6
Study1
Study2
3
2
1
0
-1
-0.5
0.5
1.5
2.5
Study1

ALTERNATIVE METHOD: RANK PROD

- *’A bioconductor package for detecting differentially expressed genes in meta-analysis.’* Hong, F., et.al. (2006) Bioinformatics.
- R package ’RankProd’
- Taking the FC ratio for each pair of measurements between patient and control group
- Taking the geometric average of FCs
- Permute the FCs w.r.t. gene ID and get the empirical distribution
- Choose genes by threshold of empirical pfp (percentage of false prediction)

Limitations :Heavy computational burden.

OUTLINE

1. Introduction
    - Study Goal
    - Multiple Testing and Meta Analysis
    - Copula Mixture Model
2. Proposed Method: A Bayesian Classification Approach
    - Mixed Effect Model for Data Generation
    - Proposed Method: Mixture Gaussian Model for T-statistics
3. Simulation Results
4. Real Data Analysis

Goal

- Bayesian Classification Approach
- Only pick genes with same directions of effect in both studies: t-statistics
- Can deal with all proportions of true associations: original scale (not empirical distribution)

Simpler model works better!!

MODELING GENE EXPRESSION DATA IN MULTIPLE STUDIES: ASSUMPTIONS FOR PROPOSED METHOD

For gene g, study i, group k (k=1 disease k=0 control), and observation j, $x_{gikj} \sim \mathcal{N}(\mu_{gik}, \sigma^2_{err})$ is gene abundance measurements after log transformation.

$$\mu_{gik} = \mu+\alpha_g+\beta_i+(\alpha\beta)_{gi}+\delta I(k=1)+\gamma_g I(k=1)+(\gamma\beta)_{gi} I(k=1)$$

- μ is overall mean
- α_g is main effect of gene g, β_i is main effect of study i, $(\alpha\beta)_{gi}$ is the gene-study interaction
- δ fixed effect of group difference
- γ_g is effect of gene on the group difference; $(\gamma\beta)_{gi}$ is the gene-study interaction of the group difference.

SIMULATION SETTINGS

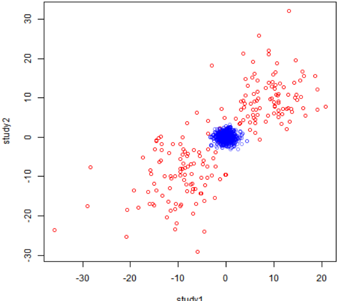
- We are most interested in the group difference between patient and control. $\mu_{gi1} - \mu_{gi0} = \gamma + (\gamma\beta)_{gi}$.
- Assume γ_g and $(\gamma\beta)_{gi}$ are normal distributed.
- Two groups:
  - Significant and Reproducible: $\gamma \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$ or $\mathcal{N}(-\mu_\gamma, 0.5)$ and $(\gamma\beta) \sim \mathcal{N}(0, \sigma^2)$
  - Irreproducible (non-significant): $\gamma = 0$ and $(\gamma\beta) = 0$.

DISTRIBUTION OF T-STATISTICS

$$T_{gi} = \frac{\bar{X}_{gi1} - \bar{X}_{gi0}}{\frac{1}{m}\sqrt{\hat{s}^2_{gi1} + \hat{s}^2_{gi0}}}$$

Where $X_{gi1j} = \alpha_g + \beta_i + (\alpha\beta)_{gi} + \gamma_g + (\gamma\beta)_{gi} + e_{gi1j}$ and $X_{gi0j} = \alpha_g + \beta_i + (\alpha\beta)_{gi} + e_{gi0j}$.

$\bar{X}_{gi1} - \bar{X}_{gi0} = \gamma_g + (\gamma\beta)_{gi} + \sum e_{gi1j}/m - \sum e_{gi0j}/m$



NORMAL APPROXIMATION FOR T-STATISTICS

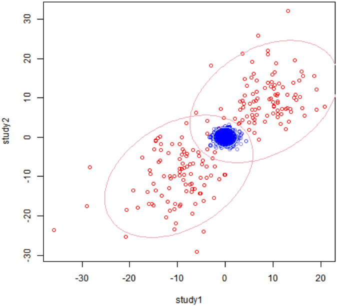
- Irreproducible group: $T \sim \mathrm{t}(2m-2)$, $E(T) = 0$, $Var(T) = 1.125$.
- Reproducible group, $T \sim \frac{\mathcal{N}(M,V)}{\sqrt{\chi^2(2m-2)/(2m-2)}}$ $M = \mathrm{effectsize}\sqrt{\frac{m}{2\sigma_e^2}}$, $V = (\sigma_{\gamma\beta}^2 + \sigma_\gamma^2 + 2\sigma_e^2/m)\frac{m}{2\sigma_e^2}$
- T statistics is approximately normal distribution when df is large. In our simulation setting, the mean is $M = 4\sqrt{5} \approx 8.94$, the variance is $V = 20(0.3 + s^2) \approx 3.3^2, 5.1^2$.
- The coefficients of our proposed method are $a = b = 8.94, c = d = -8.94$. $\sigma_1^2 = 3.3^2/5.1^2$, $\rho = 0.2/0.13$, $\sigma_2^2 = 1.125$

PROPOSED METHOD: GAUSSIAN MIXTURE MODEL

$$p(X, Y) = \pi_1\mathcal{N}((a, b); \Sigma_1) + \pi_2\mathcal{N}((c, d), \Sigma_2) + \pi_3\mathcal{N}((0, 0), \Sigma_3))$$

Here we assume there are three clusters of genes.

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} \sigma_1^2 & \sigma_1^2\rho \\ \sigma_1^2\rho & \sigma_1^2 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \pi_1 + \pi_2 + \pi_3 = 1.$$



ALTERNATIVE VIEW: LATENT VARIABLES

- Latent variables (Z_1, Z_2, Z_3): $(1, 0, 0)$ first class; $(0, 1, 0)$ second class; $(0, 0, 1)$ third class.
- $P(Z_k = 1) = \pi_k$
  $P(x, y) = \sum_z P(x, y|z)P(z) = \sum_k \pi_k \mathcal{N}((x, y)|\mu_k, \Sigma_k)$
- Responsibility of class k for observation (x, y) is
  $\gamma(z_k) = P(z_k = 1|(x, y)) = \frac{\pi_k \mathcal{N}((x,y)|\mu_k,\Sigma_k)}{\sum_k \pi_k \mathcal{N}((x,y)|\mu_k,\Sigma_k)}$
- The posterior expectation of gene g belongs to category k

EM ALGORITHM FOR MAXIMUM LIKELIHOOD ESTIMATIONS

- Initialize $\{a, b, c, d, \sigma_1^2, \sigma_2^2, \rho, \pi_1, \pi_2, \pi_3\}$ and evaluate log-likelihood
- Expectation-step: Evaluate responsibilities $\gamma_i(z_k) = p(z_k = 1|(x_i, y_i))$ using the current parameters. (The posterior expectation of gene g belongs to category k)
- Maximization-step: Update parameters: maximum likelihood estimators given the current responsibilities
- Evaluate log-likelihood $\log P(X, Y|\pi, \mu, \Sigma)$ and check for convergence (go to E-step)

DETAILS MAXIMIZATION-STEP: UPDATING PARAMETER

$$a^{new} = \frac{\sum_i \gamma_i(z_1) x_i}{\sum_i \gamma_i(z_1)}, b^{new} = \frac{\sum_i \gamma_i(z_1) y_i}{\sum_i \gamma_i(z_1)},$$

$$c^{new} = \frac{\sum_i \gamma_i(z_2) x_i}{\sum_i \gamma_i(z_2)}, d^{new} = \frac{\sum_i \gamma_i(z_2) y_i}{\sum_i \gamma_i(z_2)}, \pi_k = \sum_i \gamma_i(z_k)/N$$

$$\sigma_1^2 = \frac{\sum_i (\gamma_i(z_1)((x_i - a^{new})^2 + (y_i - b^{new})^2) + \gamma_i(z_2)((x_i - c^{new})^2 + (y_i - d^{new})^2))}{2\sum_i \gamma_i(z_1) + 2\sum_i \gamma_i(z_2)}$$

$$\rho = \frac{\sum_i \gamma_i(z_1)(x_i - a^{new})(y_i - b^{new}) + \gamma_i(z_2)(x_i - c^{new})(y_i - d^{new})}{\sigma_1^2(2\sum_i \gamma_i(z_1) + 2\sum_i \gamma_i(z_2))}$$

$$\sigma_2^2 = \frac{\sum_i \gamma_i(z_3)(x_i^2 + y^2)}{2\sum_i \gamma_i(z_3)}$$

Averages weighted by the responsibility.

Classification Rules

- Classify according to the final responsibilities $\gamma_i(z_k)$'s: the posterior probability of gene i belongs to category k.
- The Rejection Region depending on λ is

$$R_\lambda = \{(x_i, y_i) : \gamma_i(z_1) + \gamma_i(z_2) > \lambda\}$$

Rejecting genes with large posterior prob. belong to class 1 and 2.

- $\lambda = 0.85, 0.9, 0.95$, more stringent as λ increases.
- Alternative: If the class 1 or 2 has the largest posterior prob.

$$R_{\max} = \{(x_i, y_i) : \max_k \gamma_i(z_k) = \gamma_i(z_2) \vee \gamma_i(z_1)\}$$

Measurement of False Discovery

Irreproducible Discovery Rate: the rate of irreproducible genes in rejection region

$$P(\text{gene i is irreproducible}|i \in R_\gamma) = \frac{P(\text{gene i is irreproducible}, i \in R_\gamma)}{P(i \in R_\gamma)}$$

$$= \frac{FP}{TP + FP}.$$

This can be computed empirically.

In analogy to False Discovery Rate: the rate of picking the irreproducible genes.

OUTLINE

1. Introduction
   - Study Goal
   - Multiple Testing and Meta Analysis
   - Copula Mixture Model
2. Proposed Method: A Bayesian Classification Approach
   - Mixed Effect Model for Data Generation
   - Proposed Method: Mixture Gaussian Model for T-statistics
3. Simulation Results
4. Real Data Analysis

SIMULATION SETTINGS

- Considering 2 studies; each study has $m = 10$ subjects in each group (patient/control)
- Overall mean $\mu = 0$, group difference $\delta = 0$;
- Gene main effect $\alpha_g \sim \mathcal{N}(0, 1)$, study effect $\beta = 0.1$;
- Gene study interaction effect $(\alpha\beta)_g \sim \mathcal{N}(0, 0.5)$
- Gene effect on group difference γ
- Gene-study interaction of the group difference $(\gamma\beta)$
- Two groups:
  - Significant and Reproducible: $\gamma \sim \mathcal{N}(2, 0.5)$ or $\mathcal{N}(-2, 0.5)$ and $(\gamma\beta) \sim \mathcal{N}(0, s^2)$, $s \in (0.5, 1)$;
  - Irreproducible (non-significant): $\gamma = 0$ and $(\gamma\beta) = 0$.

MEASUREMENTS FOR COMPARISON OF METHODS

Methods to Compare: (subset presented):

- Proposed methods: $\gamma = 0.9, 0.95$, and pick by max.
- Copula Mixture Model. $\gamma = 0.7, 0.9$
- Benjaminni & Heller (2008,2009) with Fisher combined p-values, FDR $\alpha = 0.05$
- RankProd: with empirical FDR 0.05

Measurements of Performance:

- Misclassification Rate: $(FP + FN)/p$
- Irreproducible Identification Rate (False Discovery Rate): $FP/(TP + FP)$

SELECTED RESULT

$p = 5000$,$\sigma_{\gamma\beta} = 1$, based on 50 replications.

Table: Misclassification Rate

Irrepro.	0.9	0.95	max	copula0.7	copula0.9	fisher0.05	RP0.05
60%	0.012	0.013	0.012	0.595	0.578	0.070	0.054
80%	0.008	0.008	0.008	0.053	0.176	0.041	0.015
90%	0.004	0.004	0.005	0.019	0.073	0.023	0.005
95%	0.002	0.002	0.003	0.008	0.017	0.013	0.002
99%	0.001	0.001	0.001	0.002	0.002	0.003	0.001

Table: False Discovery Rate

Irrepro.	0.9	0.95	max	copula0.7	copula0.9	fisher0.05	RP0.05
60%	0.003	0.002	0.010	0.600	0.596	0.001	0.000
80%	0.005	0.003	0.014	0.074	0.031	0.000	0.000
90%	0.006	0.004	0.019	0.008	0.003	0.000	0.000
95%	0.010	0.006	0.029	0.005	0.000	0.000	0.002
99%	0.025	0.017	0.075	0.041	0.003	0.000	0.019

SELECTED RESULT

$p = 5000, \sigma_{\gamma\beta} = 0.5$

Table: Misclassification Rate

Irrepro.	0.9	0.95	max	copula0.7	copula0.9	fisher0.05	RP0.05
60%	0.005	0.006	0.006	0.589	0.568	0.020	0.028
80%	0.004	0.004	0.004	0.028	0.166	0.013	0.006
90%	0.002	0.002	0.003	0.008	0.058	0.008	0.002
95%	0.001	0.001	0.002	0.003	0.011	0.005	0.001
99%	0.001	0.000	0.001	0.001	0.001	0.002	0.000

Table: False Discovery Rate

Irrepro.	0.9	0.95	max	copula0.7	copula0.9	fisher0.05	RP0.05
60%	0.002	0.001	0.007	0.596	0.589	0.001	0.000
80%	0.003	0.002	0.010	0.057	0.024	0.000	0.000
90%	0.005	0.003	0.016	0.008	0.002	0.000	0.000
95%	0.008	0.006	0.026	0.008	0.001	0.000	0.002
99%	0.045	0.030	0.102	0.034	0.004	0.000	0.022

Outline

1. Introduction
   - Study Goal
   - Multiple Testing and Meta Analysis
   - Copula Mixture Model
2. Proposed Method: A Bayesian Classification Approach
   - Mixed Effect Model for Data Generation
   - Proposed Method: Mixture Gaussian Model for T-statistics
3. Simulation Results
4. Real Data Analysis

REAL DATA ANALYSIS

- Data from Gene Expression Omnibus

  Study 1: GSE28042 (Herazo-Maya JD,et al.Sci Transl Med 2013)

  Study 2: GSE33566 (Yang IV, et al. PLoS One 2012)
- Idiopathic Pulmonary Fibrosis vs Healthy Control
- Gene expression profiles of peripheral blood RNA
- Study 1: monocytes(PBMC); Study 2: whole blood cell (PBC)
- There are 17708 common genes in both studies.
- 75 patients and 16 controls for study 1; 93 patients and 30 controls for study 2.

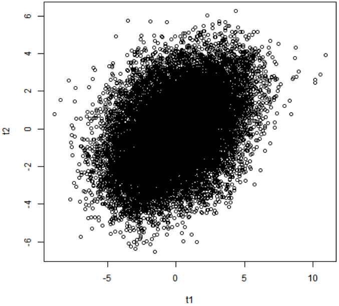
RELAX THE MODEL ASSUMPTION:

T statistics are not in the same scale.

$$p(X,Y) = \pi_1\mathcal{N}((a,b);\Sigma_1) + \pi_2\mathcal{N}((c,d),\Sigma_2) + \pi_3\mathcal{N}((0,0),\Sigma_3))$$

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} \sigma_{11}^2 & \sigma_{11}\sigma_{12}\rho \\ \sigma_{11}\sigma_{12}\rho & \sigma_{12}^2 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} \sigma_{21}^2 & 0 \\ 0 & \sigma_{22}^2 \end{bmatrix},$$

$$\pi_1 + \pi_2 + \pi_3 = 1.$$



PROPOSED METHOD

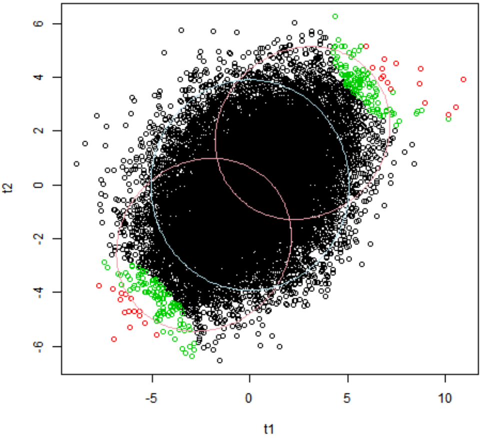


Figure: T-statistics of Proposed Method, green: picked by ldr 95%; picked by ldr 99%

Table: Top 10 genes picked by proposed method

	Proposed	t stat study 1	t stat study 2
1	FOS	10.92	3.94
2	BACH2	-7	-5.74
3	DUSP1	8.74	4.33
4	LRRN3	-6.13	-5.31
5	FCAR	7.47	4.53
6	NACC2	8.68	3.76
7	ABLIM1	-6.43	-4.74
8	ASGR2	6.85	4.67
9	GPR18	-6.21	-4.68
10	SRGN	10.57	2.9

COPULA MIXTURE : T STATISTICS

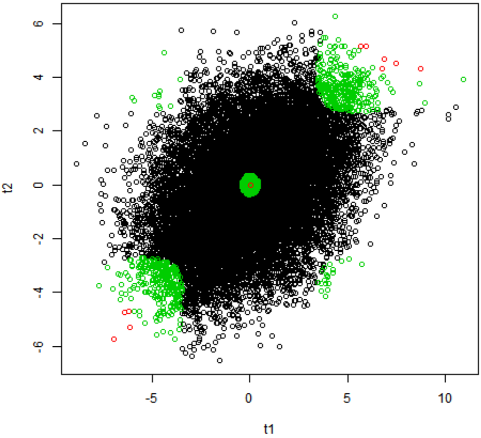


Figure: T-statistics of Copula Mixture Model, green: picked by ldr 50%; picked by ldr 70%

COPULA MIXTURE: Z SCORES

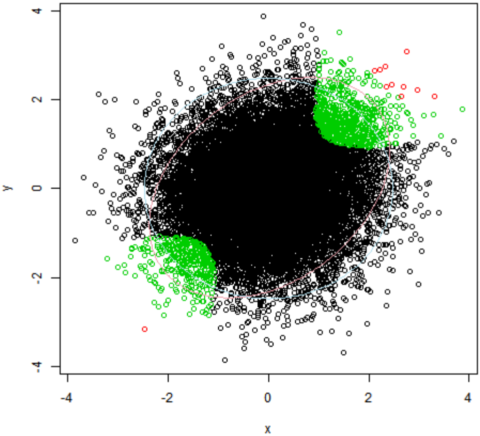


Figure: Z-score of Copula Mixture Model, green: picked by ldr 50%; picked by ldr 70%

Table: Top 10 genes picked by Copula Mixture Model

	Copula	t stat study 1	t stat study 2
1	BACH2	-7	-5.74
2	LOC92659	0.02	0
3	FCAR	7.47	4.53
4	LRRN3	-6.13	-5.31
5	DUSP1	8.74	4.33
6	ASGR2	6.85	4.67
7	PHF21A	5.95	5.18
8	ABLIM1	-6.43	-4.74
9	SLC16A3	5.68	5.17
10	GPR18	-6.21	-4.68

Fisher BH

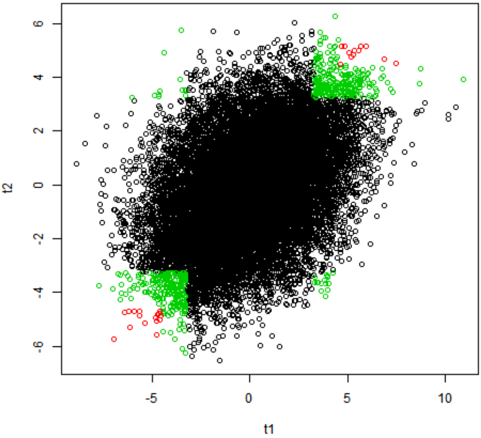


Figure: Z-score of Fisher BH Method, green: picked by FDR 95%; picked by FDR 99%

Table: Top 10 genes picked by Fisher and BH

	Fisher	t stat study 1	t stat study 2
1	BACH2	-7	-5.74
2	LRRN3	-6.13	-5.31
3	PHF21A	5.95	5.18
4	SLC16A3	5.68	5.17
5	TXK	-5.37	-5.12
6	TIMP2	5.54	4.99
7	QSOX1	5.37	4.99
8	NLRX1	5.05	4.93
9	LOC439949	-5.67	-4.85
10	PTPRE	5.31	4.84

Table: Top 10 genes picked

	Proposed	Copula	Fisher
1	FOS	BACH2	BACH2
2	BACH2	LOC92659	LRRN3
3	DUSP1	FCAR	PHF21A
4	LRRN3	LRRN3	SLC16A3
5	FCAR	DUSP1	TXK
6	NACC2	ASGR2	TIMP2
7	ABLIM1	PHF21A	QSOX1
8	ASGR2	ABLIM1	NLRX1
9	GPR18	SLC16A3	LOC439949
10	SRGN	GPR18	PTPRE

CONCLUSIONS AND DISCUSSION

Conclusion:

- We developed a Bayesian Classification for the t-statistics.
- Superior performance is demonstrated in simulation studies and real data analysis.

Discussion:

- The method can be extended to multiple (>2) study case using Mixture Multivariate Normal Distribution
- Can be easily extended to adjust for covariates: age, gender, etc. Use the T statistics in the regression model for the disease group indicator variable.

Thank You!