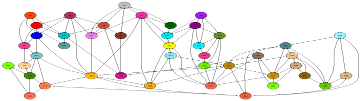
Method Selection and Graphical Network: Applications to Gene Expression Data

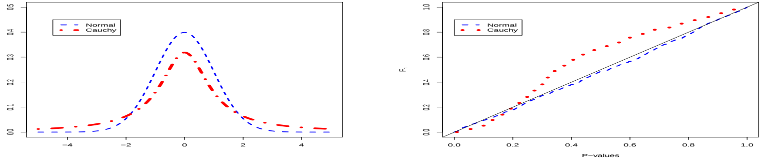
Demba Fofana, PhD
University of Texas Rio Grande Valley

Introduction

- Problem: How to perform a large number of tests using method M_1 or M_2 and adjust for multiple testing.
- When an assumption A is valid M_1 has more power than M_2 and when A does not hold M_2 reveals to be more powerful than M_1.
- And also take into account Graphical Network that exists among entities.
- Solution: Hybrid-Network assesses Assumption Validity and takes into account Graphical Network.



Motivations & Description



(a) Statistics

(b) P-value CDF

Theorem (Hybrid P-values)

Suppose there are two different procedures M_1 and M_2 that can be used to test the null hypothesis, say $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Let P_1 be the p-value obtained if the method M_1 is used for testing the null hypothesis H_0, and P_2 be the p-value if the method M_2 is used instead. Let P be defined by

$$P = \begin{cases} P_1, & \text{if } M_1 \\ P_2, & \text{if } M_2. \end{cases}$$

Then P is uniformly distributed under the null hypothesis H_0.

Motivations & Description

Proof.

Under the null hypothesis (H_0) (of primary interest, gene is expressed say), both P_1 and P_2 are uniformly distributed $[0;1]$.

$$\begin{aligned}
\mathbb{P}(P < p \mid H_0) &= \mathbb{P}\left\{(P < p) \cap [M_1 \cup M_2] \mid H_0\right\} \\
&= \mathbb{P}\left\{(P < p) \cap M_1 \mid H_0\right\} + \mathbb{P}\left\{(P < p) \cap M_2 \mid H_0\right\} \\
&= \mathbb{P}(P < p \mid M_1, H_0)\mathbb{P}(M_1 \mid H_0) + \\
&\quad \mathbb{P}(P < p \mid M_2, H_0)\mathbb{P}(M_2 \mid H_0) \\
&= \mathbb{P}(P_1 < p \mid H_0)\mathbb{P}(M_1 \mid H_0) + \\
&\quad \mathbb{P}((P_2 < p) \mid H_0)\mathbb{P}(M_2 \mid H_0) \\
&= p\mathbb{P}(M_1 \mid H_0) + p\mathbb{P}(M_2 \mid H_0) \\
&= p\mathbb{P}(M_1 \mid H_0) + p(1 - \mathbb{P}(M_1 \mid H_0)) \\
&= p.
\end{aligned}$$

□

edd

Methodology

- In a spatial normal mixture model,

$$f(z_g) = \pi_{g0} f_o(z_g) + \pi_{g1} f_1(z_g), \tag{1}$$

  where $z_g = \Phi^{-1}(1 - P_g)$ and π_{gs} are gene-specific prior probabilities.

- The prior probabilities, π_{gs}, based on gene network, are related to two latent Markov random fields $\mathbf{x}_s = \{x_{gs}; g = 1, \cdots, G\}$, $s = 0, 1$ by:

$$P(T_g = s) = \pi_{gs} = \frac{exp(x_{gs})}{exp(x_{g0}) + exp(x_{g1})}, \tag{2}$$

  $T_g \equiv 1$ if gene g is expressed and $T_g \equiv 0$ if not expressed.

- The distribution of each spatial latent variable x_{gs} conditional on $x_{-gs} = \{x_{ks}; k \neq g\}$ depends only on its direct neighbors,

$$x_{gs} \mid x_{-gs} \sim N(\frac{1}{m_g} \sum_{l \in \delta_g} x_{ls}, \frac{\sigma_s^2}{m_g}) \tag{3}$$

  where δ_g is the set of indices for the neighbors of gene g, and m_g is the corresponding number of neighbors.

Results: Simulations

- To compare the Hybrid-Network method with other methods we conducted simulation studies designed to mimic testing situations that might arise in real world situations. We conducted standard two-group comparison studies (treatment vs control), k-group comparison (ANOVA), and regression analysis.
- The description of the setup is as follows:
  1) There are two groups of sample size varying from 5, 10, 25, 50.
  2) The number of genes with the normal distribution, $N(\mu, 1)$, is 30, $\mu = 0$ for the null hypothesis and $\mu = 1$ for the alternative, and the number of genes with the Log-normal distribution, $Log - normal(\mu, 1)$, with $\mu = 0$ in some cases and $\mu = 1$ in other cases, is 14.
  3) A graphical network is built among genes with 212 number of neighbors.

Results: Simulations

Table: 2−Group Comparison: Specificities

Sample size (n_i)	T-test sp	Rank Sum-test sp	Hybrid-Network-test sp
5	0.571726	0.557244	0.575314
10	0.689223	0.69797	0.716146
25	0.884244	0.918197	0.921273
50	0.9839	0.994575	0.994575

sp ≡ specificity

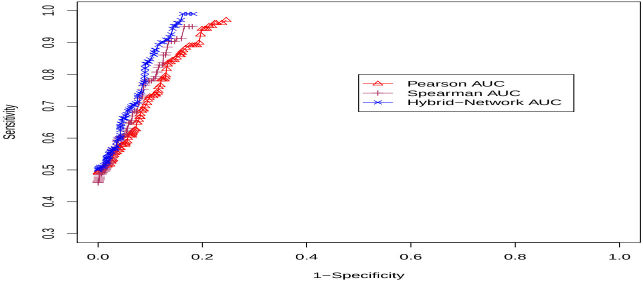
Table: 3−Group Comparison: Specificities

Sample size (n_i)	F-test sp	H-test sp	Hybrid-Network test sp
5	0.579557	0.57232	0.585729
10	0.668287	0.668287	0.684932
25	0.89141	0.918197	0.929054
50	0.92437	0.9839	0.985663

sp ≡ specificity

Results: Simulations

- The description of the setup is as follows:
  - The sample size is 25 and the cutoff point, τ, is varied.
  - The number of genes with the normal distribution , $N(\mu, 1)$, is 30, $\mu = 0$ for the null hypothesis and $\mu = 1$ for the alternative, and the number of genes with the Log-normal distribution, $Log - normal(\mu, 1)$, with $\mu = 0$ in some cases and $\mu = 1$ in other cases, is 14.
  - A graphical network is built among genes with 212 number of neighbors.



Results: Application to Tumor Data

- Tumor is cancer disease that occurs in 2 distinct anatomic regions:
- We use Affymetrix arrays to compare expression across the 2 groups.
- A graphical network is provided.
- We develop a Hybrid-Network test procedure using t-test, Rank Sum, Shapiro-Wilk tests, and CAR (Conditional Autoregressive Priors).

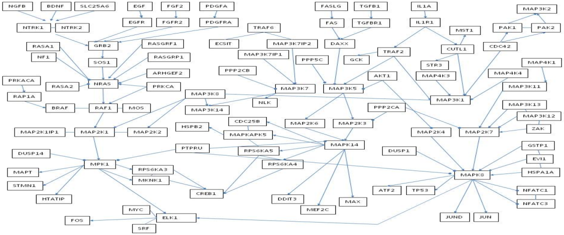
Table: Human Ependymoma Microarray Data

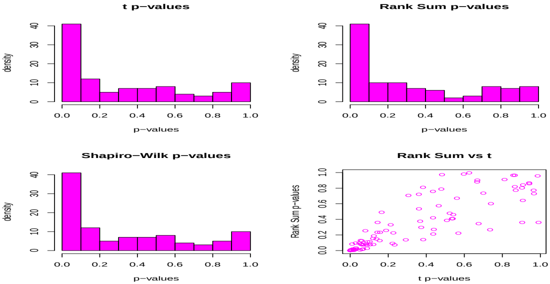
Genes	Gr1	Gr1	···	Gr2	Gr2	···
AKT1	12.48167	11.75317	···	10.95536	11.51737	···
ARHGEF2	14.99632	13.81004	···	13.45263	14.02982	···
ATF2	12.93096	13.14289	···	13.44182	12.72238	···
BDNF	3.392317	4.542258	···	4.716991	5.738768	···
BRAF	9.111918	10.3433	···	10.07682	9.107217	···
CDC25B	10.33114	11.04207	···	11.7139	11.76408	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮

This shows the human ependymoma expression data: genes as gene annotation, groups (Gr1 and Gr2) as sample annotation and real values as gene expression levels.

edd

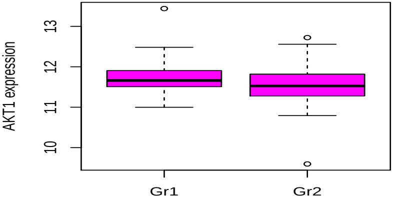
Results: Application to Tumor Data



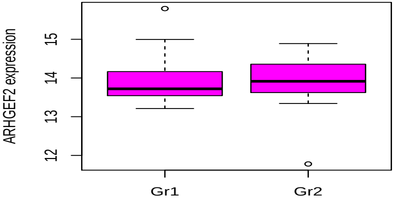


edd

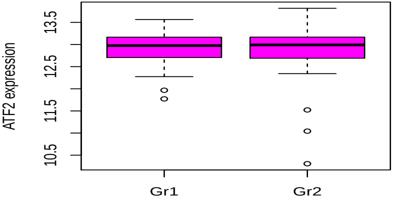
Results: Application to Tumor Data



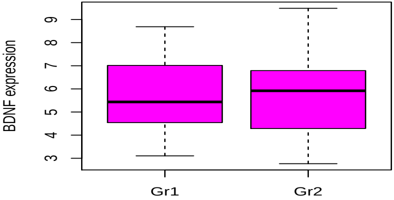
t = 0.846 ; rs = 0.962 ; hybN = 0.615 ; Shp = 0.002



t = 0.447 ; rs = 0.81 ; hybN = 0.067 ; Shp = 0



t = 0.5 ; rs = 0.5 ; hybN = 0.5 ; Shp = 0



t = 0.359 ; rs = 0.74 ; hybN = 0.099 ; Shp = 0.02

Discussions

- Assumptions and graphical network profoundly impact the validity of an analysis.
- Assumptions are not routinely evaluated in multiple testing applications (Gene expression data analysis) because they entail adding new layers of multiplicity.
- Hybrid-network that incorporates both assumptions and graphical network shows good performances in simulations and in real data.
- Writing an R Package that considers assumptions and graphical network into the analysis of gene expressions data.

References

- Bioconductor: HybridMTest
- Comput Stat Data Anal. 53(5): 1604-1612.
- J Roy Statist Soc Ser B (Methodological) 57:289-300.
- Spatial and Spatio-temporal Epidemiology 2 (2011) 79-89.

Appendix

```
model
{
for (i in 1 : N) {
z[i] ∼ dnorm(muR[i],tauR[i]) #z-score
muR[i]< −mu[T[i]]
tauR[i]< −tau[T[i]]
#logistic
pi[i,1]< −exp(X1[i])/(exp(X1[i])+exp(X2[i]))
pi[i,2]< −exp(X2[i])/(exp(X1[i])+exp(X2[i]))
T[i]∼dcat(pi[i,1:2])
T1[i]< −equals(T[i],1)
T2[i]< −equals(T[i],2)
}
#Random Fields specification
X1[1:N]∼car.normal(adj[],weights[],num[],tau[1])
X2[1:N]∼car.normal(adj[],weights[],num[],tau[2])
#Weights Specification
for(k in 1:sumNumNeigh){weights[k]< −1}
#Priors specification(precision for MRF)
#Prior: means of normal mixture components
mu[1]∼dnorm(0,1.0E-6)
mu[2]∼dnorm(0,1.0E-6) #I(0.0,) #add I(,0.0)?
#Priors:precision/variance of normal mixture component
tau[1]∼dgamma(0.1,0.1)
tau[2]∼dgamma(0.1,0.1)
}
```

Appendix

```
source(”http : //bioconductor.org/biocLite.R”)
biocLite(”RBGL”)
library(”graph”)
myNodes¡-c(”G1”,”G2”,”G3”,”G4”,”G5”,”G6”,”G7”,”G8”,”G9”,”G10”,
”G11”,”G12”,”G13”,”G14”,”G15”,”G16”,”G17”,”G18”,”G19”,”G20”,
”G21”,”G22”,”G23”,”G24”,”G25”,”G26”,”G27”,”G28”,”G29”,”G30”,
”G31”,”G32”,”G33”,”G34”,”G35”,”G36”,”G37”,”G38”,”G39”,”G40”,
”G41”,”G42”,”G43”,”G44”)
myEdges< −list(G1=list(edges=c(”G17”,”G12”,”G9”,”G8”,”G4”)),
G2=list(edges=c(”G14”,”G13”,”G10”,”G7”)),
G3=list(edges=c(”G32”,”G17”,”G15”,”G11”,”G8”,”G6”)),
G4=list(edges=c(”G33”,”G32”,”G17”,”G16”,”G14”,”G12”,”G1”)),
G44=list(edges=c(”G41”,”G32”,”G31”,”G26”,”G25”,”G22”)))
g< −new(”graphNEL”,nodes=myNodes,edgeL=myEdges, edgemode=”directed”)
library(”Rgraphviz”)
library(”RBGL”)
cc< −connectedComp(g)
colors< −c(”gray”,”purple”,”maroon”,”maroon2”,”orangered”,
”red”, ”darkmagenta”,”tomato3”,”tomato4”,”olivedrab”,
”blue”, ”darkgreen”,”turquoise1”,”turquoise2”,”turquoise3”,
”yellow”, ”violet”,”violetred”,”violetred1”,”violetred2”,
”cadetblue”,”cadetblue1”,”cadetblue2”,”cadetblue3”, ”cadetblue4”,
”burlywood”,”burlywood1”,”burlywood2”,”burlywood3”,”burlywood4”,
”darkgoldenrod”,”darkgoldenrod1”,”darkgoldenrod2”,”darkgoldenrod3”,”darkgoldenrod4”,
”chartreuse”,”chartreuse1”,”chartreuse2”,”chartreuse3”,”chartreuse4”,
”coral”, ”coral1”, ”coral2”,”tomato2”,listlen=(cc))
names(colors)< −unlist(cc)
plot(g, nodeAttrs=list(fillcolor=colors))
```

Thank You All !!!