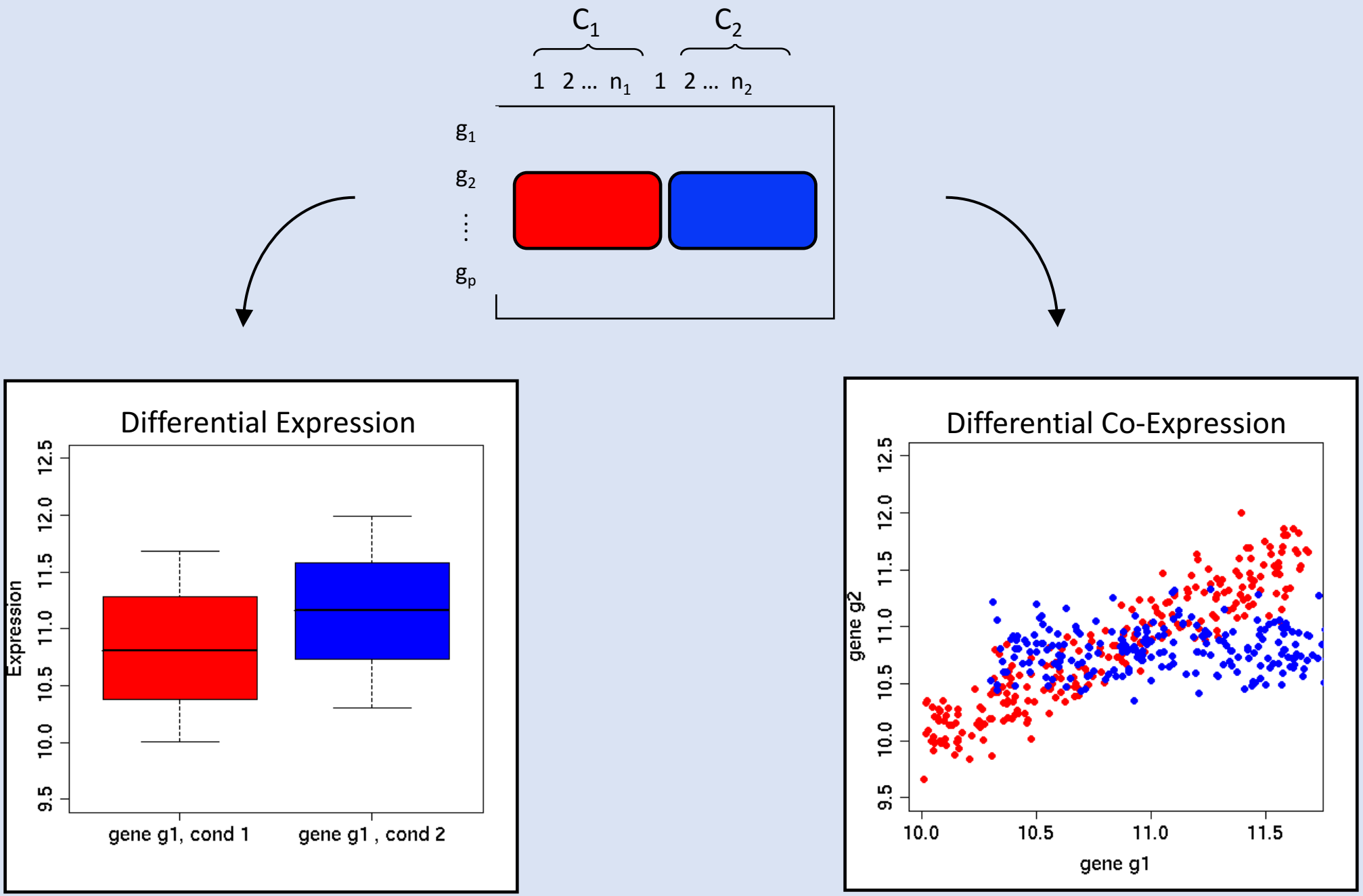
Empirical Bayes Analysis of Covariance

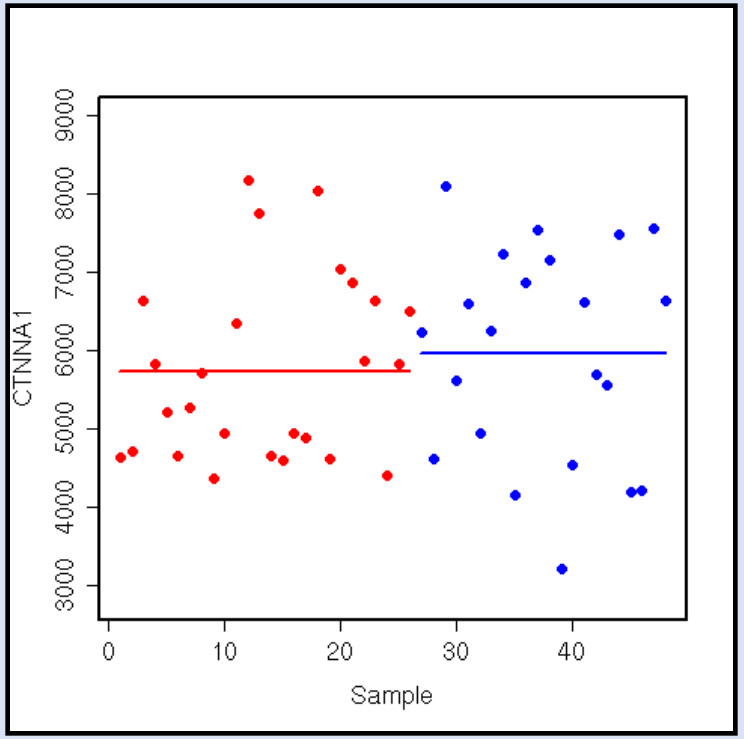
Jacob M. Maronge

PhD Student
Department of Statistics
University of Wisconsin-Madison
https://jmmaronge.github.io
jmmaronge@gmail.com

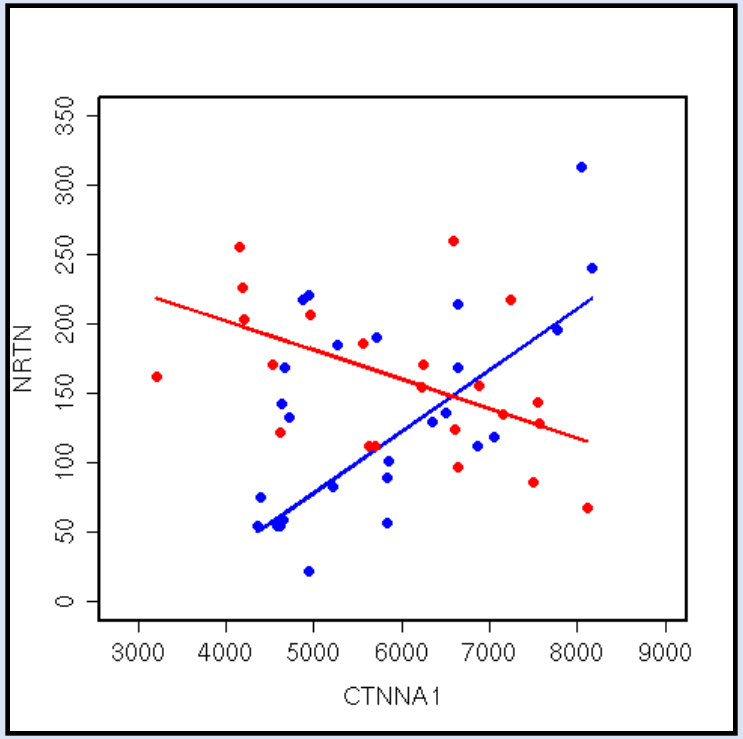
Motivation

- In many genetics experiments, it is common to attempt to identify genes that are differentially expressed (DE) across two conditions.
- Though this is extremely important, it does not characterize the only interesting changes in genetic expression across conditions (Dawson, de la Fuente).

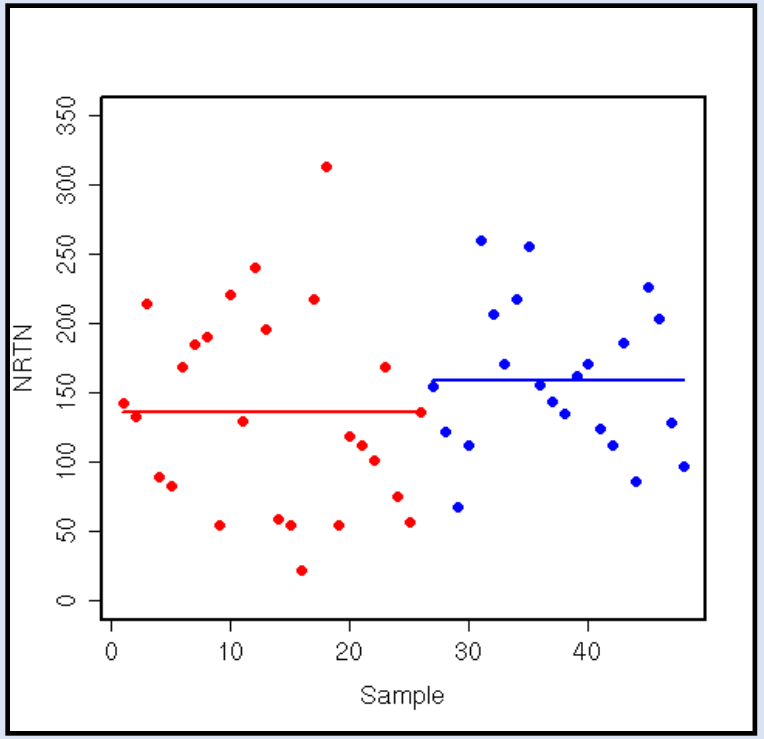

C1
C2
1 2 … n1 1 2 … n2
g1
g2
gp
Differential Expression
Expression
9.5
10.0
10.5
11.0
11.5
12.0
12.5
gene g1, cond 1
gene g1 , cond 2
Differential Co-Expression
gene g2
gene g1
10.0
10.5
11.0
11.5



CTNNA1 is not DE.



CTNNA1 and NRTN are Differentially Co-Expressed.



NRTN is not DE.

(Dawson and Kendziorski (2012), Biometrics)

Next steps

- There are many methods available for detecting differential co-expression.
- What if we want to look at the co-expression between gene groups of size > 2?
- Many online databases (e.g. KEGG, GO) suggest existence of gene networks.

Hierarchical Mixture Model

- $p(x|\Sigma_1) \sim N_p\left(\tilde{\mu} = \tilde{0}, \Sigma_1\right), p(y|\Sigma_2) \sim N_p\left(\tilde{\mu} = \tilde{0}, \Sigma_2\right)$
- Under H_0: $\Sigma_1 \sim IW_p(\psi, m)$ and $\Sigma_1 = \Sigma_2$.
- Under H_A: $\Sigma_1, \Sigma_2 \sim$ iid $IW_p(\psi, m)$.
- Mixture Distribution:

$$p(x, y) = \pi_0 p_0(x, y) + (1 - \pi_0) p_0(x) p_0(y),$$

where,

$$p_0(x, y) = \int_{\theta \in \Theta} p(x, y|\theta) p(\theta) d\theta,$$

$$\Longrightarrow P(H_0|x, y) = \frac{\pi_0 p_0(x,y)}{\pi_0 p_0(x,y) + (1-\pi_0) p_0(x) p_0(y)}.$$

Hierarchical Mixture Model (Continued)

- This results in a predictive distribution,

$$p_0(x) = \int_{\Sigma \in \Theta} p(x|\Sigma)p(\Sigma)d\Sigma$$

$$= \frac{\Gamma_p\left(\frac{n+m}{2}\right)}{\pi^{\frac{np}{2}}\Gamma\left(\frac{m}{2}\right)}|\psi|^{-\frac{n}{2}}|I_n + X\psi^{-1}X^T|^{-\frac{n+m}{2}}$$

$$\Longrightarrow x \sim T_{n,p}(m-p+1, J_0, I_n, \psi)$$

Why Use This Framework?

- Flexibility – Can change the prior distribution and dimension easily.
- Increase Power – Since the estimates for the hyper-parameters will be eventually estimated from the complete data, when doing many tests we expect to see an increase in power because we can share information across covariance matrices.

$$\Sigma_1, \Sigma_2$$

	[,1]	[,2]	[,3]
[1,]	7.205584	1.561549	1.435980
[2,]	1.561549	1.114490	1.207751
[3,]	1.435980	1.207751	7.229193

	[,1]	[,2]	[,3]
[1,]	0.2607096	0.3007188	-0.2609899
[2,]	0.3007188	0.8611390	-0.7443316
[3,]	-0.2609899	-0.7443316	2.9603221

$$S_1, S_2$$

$$S = X^T X$$

	[,1]	[,2]	[,3]
[1,]	110.16150	17.35239	46.79744
[2,]	17.35239	14.96605	18.81225
[3,]	46.79744	18.81225	113.61027

	[,1]	[,2]	[,3]
[1,]	7.751668	10.01520	-9.880433
[2,]	10.015203	20.57942	-26.655482
[3,]	-9.880433	-26.65548	55.054883

0.008017213

$$P(H_0|Data)$$

Simulation Setup

- Condition 1
  Generate observation, $\Sigma_1 \sim IW_p(\psi, m)$.
  Use Σ_1 to generate data, $x \sim N_p(\tilde{\mu} = \tilde{0}, \Sigma_1)$.

- Condition 2
  Two possibilities for Σ_2

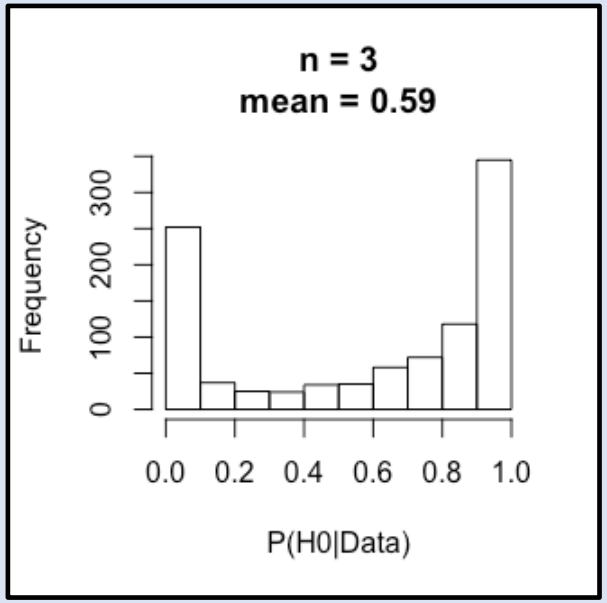
  Takes the same value as Σ_1 with some prob. ξ

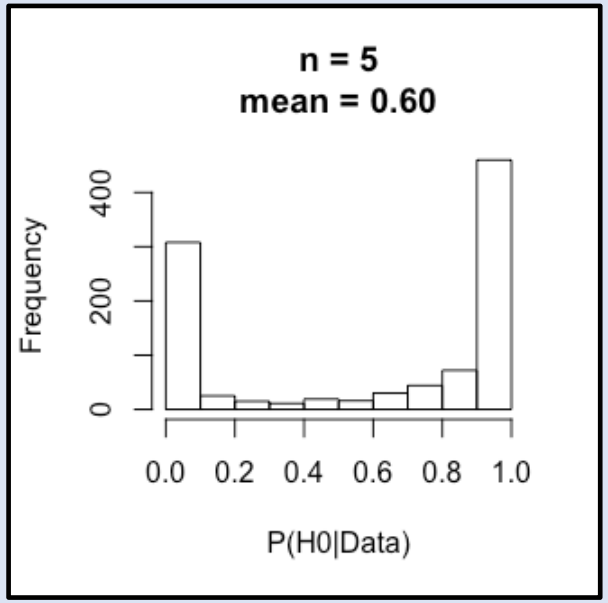
  Random draw from IW_p dist., which is independent from the dist. of Σ_1, with prob. $1 - \xi$

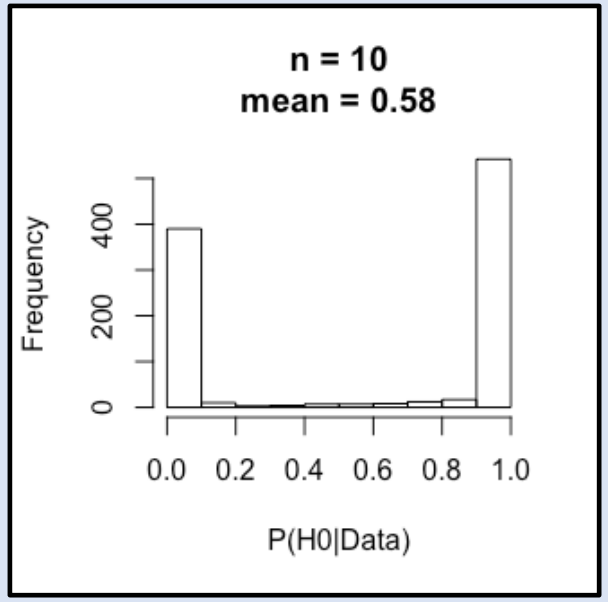
  Generate data, $y \sim N_p(\tilde{\mu} = \tilde{0}, \Sigma_2)$.

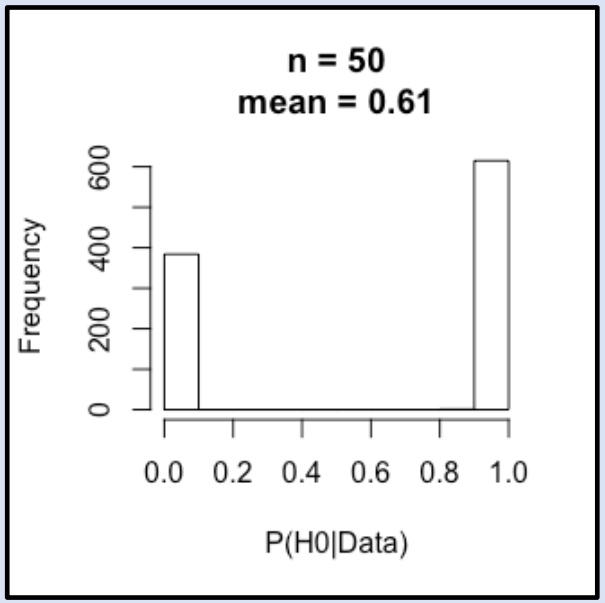
- Calculate $P(H_0 | x, y)$, repeat many times.

Results









$$\xi = 0.6$$

Future Work

- Implement EM algorithm to estimate π_0, ψ, and m from data.
- Explore effects of different prior distributions on Σ.
- Apply to genomics dataset using predefined networks (e.g. KEGG, GO)
- Possible application to Diffusion Tensor Imaging (DTI).

Conclusions

- There is a need for powerful and flexible methods for detecting differences in gene networks across conditions.
- The hierarchical mixture model framework is a flexible way to do so (can change prior, easy to add more conditions, etc.).
- It is possible to implement in R!

Special Thanks

- Prof. Michael Newton

References

- Dawson, John A, and Christina Kendziorski. 2012. “An Empirical Bayesian Approach for Identifying Differential Coexpression in High-Throughput Experiments.” Biometrics 68 (2): 455–65.
- de la Fuente, A. 2010. “From ‘differential expression’ to ‘differential networking’ - identification of dysfunctional regulatory networks in diseases.” Trends in Genetics, 26(7):326–333.
- Kendziorski, C.M., M.A. Newton, H. Lan, and M.N. Gould. 2003. “On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.” Statistics in Medicine, 22: 3899-3914