Empirical Bayes Analysis of Covariance

Jacob M. Maronge

PhD Student
Department of Statistics
University of Wisconsin-Madison
https://jmmmaronge.github.io
jmmmaronge@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).
Differential Expression

C_1 \quad C_2
1 \quad 2 \ldots \quad n_1 \quad 1 \quad 2 \ldots \quad n_2

\begin{align*}
g_1 \\
g_2 \\
\vdots \\
g_p
\end{align*}

Differential Co-Expression
CTNNA1 is not DE.

NRTN is not DE.

CTNNA1 and NRTN are Differentially Co-Expressed. (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

• \( p(x|\Sigma_1) \sim N_p(\bar{\mu} = \bar{0}, \Sigma_1), p(y|\Sigma_2) \sim N_p(\bar{\mu} = \bar{0}, \Sigma_2) \)

• 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,

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

\[
\Rightarrow P(H_0|x, y) = \frac{\pi_0 p_0(x,y)}{\pi_0 p_0(x,y) + (1-\pi_0)p_0(x)p_0(y)}.
\]
Hierarchical Mixture Model (Continued)

• This results in a predictive distribution,

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

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

\[
\Rightarrow x \sim T_{n,p}(m - p + 1, J_0, I_n, \psi)
\]
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.
\[
\Sigma_1, \Sigma_2
\]

\[
S_1, S_2
\]

\[
S = X^T X
\]

\[
P(H_0 | Data)
\]
Simulation Setup

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

• Condition 2
  Two possibilities for $\Sigma_2$
  Generate data, $y \sim N_p(\bar{\mu} = \tilde{0}, \Sigma_2)$.

  Takes the same value as $\Sigma_1$ with some prob. $\xi$
  Random draw from $IW_p$ dist., which is independent from the dist. of $\Sigma_1$, with prob. $1 - \xi$

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

\[ \zeta = 0.6 \]
Future Work

• Implement EM algorithm to estimate $\pi_0$, $\psi$, and $m$ from data.
• Explore effects of different prior distributions on $\Sigma$.
• 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

