1 Introduction

- Differential expression analysis ⇒ Differential gene co-expression network
- Introduce the concept of scale-free topology and hub genes
- Goal: model the connectivity of one gene (especially hub genes) in terms of the covariates of interest
- Explain the difficulty of low sample size and weak signals
- In laymen language, explain how accruing information across genes (degree analysis) can solve this problem
- literature review: modeling heterokedasticity, vs liquid association
- organization

2 Methods

2.1 Framework

We assume data $\boldsymbol{y_i} \in \mathbb{R}^K$, $i = 1, \cdots, n$, are independent, and we define $Y = \{y_{ik}\}_{i=1,k=1}^{n,K} \in \mathbb{R}^{n \times K}$. The K variables have a non-trivial dependence structure. For example, $\boldsymbol{y_i}$ is the gene expression levels measured at K genes for individual i, and the dependence structure demonstrates the gene co-expression network. We define a covariate matrix $X = \{x_{ip}\}_{i=1,p=1}^{n,P} \in \mathbb{R}^{n \times P}$ with P features to study the effect of X on the variance structure of Y. The first column of X is $\mathbf{1}$ to account for the intercept. For example, the feature vector $\boldsymbol{x_i} \in \mathbb{R}^P$ holds ancestry information of individual i, for instance, a scalar value denoting the proportion of African ancestry in an African American genome ($P = 2$), the first few principal components of the genotypes, or local ancestry information at multiple loci.

We assume that $\boldsymbol{y_i}$ follows a K-dimensional multivariate normal distribution where the diagonal entries of the covariance matrix are scaled to be 1.

$$\boldsymbol{y_i} = \boldsymbol{x_i}^T B + \boldsymbol{\epsilon_i}, \boldsymbol{\epsilon_i} \sim \mathcal{N}_K(\mathbf{0}, \Sigma(\boldsymbol{x_i})) \tag{1}$$

$$\Sigma(\boldsymbol{x_i}) = \{\rho_{k_1k_2}(\boldsymbol{x_i})\}_{k_1k_2=1}^K, \quad \rho_{k_1k_2}(\boldsymbol{x_i}) = 1 \text{ if } k_1 = k_2$$

where $B \in \mathbb{R}^{P \times K}$ measures the effect size of X on the mean of Y, and the variance Σ, also a function of X, is essentially a correlation matrix with scaled variance 1 on diagonals and correlation ρ for off-diagonals. We further assume that each column of X except the intercept has mean 0 and sum of squares n. In other words, $\sum_{i=1}^n x_{ip} = 0$ and $\sum_{i=1}^n x_{ip}^2 = n$ for $p = 2, \cdots, P$.

We first focus on two variables and model their correlation, ρ, in terms of X. Then we combine information across multiple pairs of genes to make inference on each gene's connectivity across the entire network.

2.2 Test Statistic

Without loss of generality, we focus on variables $k_1 = 1$ and $k_2 = 2$ to model their correlation with respect to X. We assume that Fisher-transformed correlation $\rho_{12}(\boldsymbol{x})$ is linearly associated with the covariates of interest with effect size $\boldsymbol{\alpha_{12}}$.

$$\frac{1}{2} log \left(\frac{1 + \rho_{12}(\boldsymbol{x_i})}{1 - \rho_{12}(\boldsymbol{x_i})} \right) = \frac{1}{2} \boldsymbol{x_i}^T \boldsymbol{\alpha_{12}}, \quad \boldsymbol{\alpha_{12}} \in \mathbb{R}^P \tag{2}$$

Then we define a new random variable $W_{12} \in \mathbb{R}^n$ where $w_{i,12} = y_{i1} + y_{i2}$ by considering the sum of two variables. From (1), we write the following

$$w_{i,12} = \boldsymbol{x_i}^T \boldsymbol{\beta_{12}} + u_{i,12}, \qquad u_{i,12} \sim \mathcal{N}(0, \sigma^2_{i,12}) \tag{3}$$

$$\boldsymbol{\beta_{12}} = \boldsymbol{\beta_1} + \boldsymbol{\beta_2}, \qquad \sigma^2_{i,12} = 2 + 2\rho(\boldsymbol{x_i})$$

where $\boldsymbol{\beta_1}$ is the first column of matrix B. From (2), we can re-write $\sigma^2_{i,12}$ in terms of $\boldsymbol{\alpha_{12}}$ as follows, where we also define function h to indicate the form of heterokedasticity.

$$\sigma^2_{i,12} = 2 + 2\rho_{12}(\boldsymbol{x_i}) = 4\left(1 - \frac{1}{e^{\boldsymbol{x_i}^T \boldsymbol{\alpha_{12}}} + 1}\right) = h(\boldsymbol{x_i}^T \boldsymbol{\alpha_{12}}) \tag{4}$$

We partition $\boldsymbol{\alpha_{12}}$ into $\begin{bmatrix}\alpha^0_{12} & \boldsymbol{\alpha^1_{12}}\end{bmatrix}^T$ where $\alpha_{0,12}$ is a scalar corresponding to the intercept term. Then our null hypothesis is

$$H_0^{(12)} : \boldsymbol{\alpha^1_{12}} = \boldsymbol{0}. \tag{5}$$

If H_0 is true, $\sigma^2_{i,12}$ is constant across $i = 1, \cdots, n$, and $\sigma^2_{i,12} = h(x_i^T \boldsymbol{\alpha_{12}}) = h(\alpha^0_{12}) = \sigma^2_{12}$. In other words, the variance of Y does not depend on X.

The Lagrange Multiplier (LM) test, proposed by Breusch and Pagan (1979) [1], or equivalently efficient scores criterion by Rao (1948) [11], is a popular choice to test for a heterokedasticity in a linear model. One practical advantage is that the test statistic does not depend on the specific form of heterokedasticity. It is an attractive property under our context too, since it can easily incorporate the Fisher transformation and function h. Another advantage is that it only requires parameter estimation assuming that the null hypothesis is true, and it uses restricted maximum likelihood, i.e., it obtains the maximum log likelihood of (3) subject to the restriction of null hypothesis (5). So, in order to compute the restricted maximum likelihood estimate, we first define the OLS residuals of model (3) as $\hat{u}^2_{i,12}$ and estimated residual variance under the null as $\hat{\sigma}^2_{12} = \frac{1}{n}\sum_{i=1}^n \hat{u}^2_{i,12}$. Then, the Lagrange Multiplier test statistic q_{12} for the variable pair (1,2) is, as defined in Breusch and Pagan (1979),

$$q_{12} = \frac{1}{2n}\left(\sum_{i=1}^n \boldsymbol{x_i}\left(\frac{\hat{u}^2_{i,12}}{\hat{\sigma}^2_{12}} - 1\right)\right)^T \left(\sum_{i=1}^n \boldsymbol{x_i}\boldsymbol{x_i}^T\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{x_i}\left(\frac{\hat{u}^2_{i,12}}{\hat{\sigma}^2_{12}} - 1\right)\right) \tag{6}$$

Note that our null hypothesis tests for all covariates at the same time, so we can orthogonalize X to make $\sum_{i=1}^n \boldsymbol{x_i}\boldsymbol{x_i}^T$ an identity matrix, and the inference is not affected. Let $\tilde{X}$ be the orthogonalized covariate matrix, and $\tilde{x}_{ip}$ be the corresponding entries with $\sum_{i=1}^n \tilde{x}_{ip} = 0$ and $\sum_{i=1}^n \tilde{x}_{ip} = n$. Then (6) can be alternatively written as follows, where we define $r_{12,p}$

$$q_{12} = \sum_{p=1}^P \left(\frac{1}{\sqrt{2n}}\sum_{i=1}^n \tilde{x}_{ip}\left(\frac{\hat{u}^2_{i,12}}{\hat{\sigma}^2_{12}} - 1\right)\right)^2 = \sum_{p=1}^P {r_{12,p}}^2 \tag{7}$$

When $p = 1$, $r_{12,p} = 0$ due to the construction of $\hat{u}^2_{12}$ and $\hat{\sigma}^2_{12}$. When $p \neq 1$, due to CLT,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{x}_{ip}\left(\hat{u}^2_{i,12} - \hat{\sigma}^2\right)\right)$$

converges in distribution to normal with mean $E(\tilde{x}_{ip}(\hat{u}^2_{i,12} - \hat{\sigma}^2)) = E(x_{ip})E(\hat{u}^2_{i,12} - \hat{\sigma}^2) = 0$ and $\mathrm{Var}(\tilde{x}_{ip}(\hat{u}^2_{i,12} - \hat{\sigma}^2)) = E(\tilde{x}^2_{ip})E(\hat{u}^2_{i,12} - \hat{\sigma}^2)^2 = 2\sigma^4$ from the fact that the sample variance converges asymptotically to the population variance. Also, $\frac{1}{\hat{\sigma}^2}$ converges in probability to $\frac{1}{\sigma^2}$, and therefore,

$$r_{12,p} \to N(0,1)$$

independently for all $p = 2, \cdots, K$. It confirms the result of Breusch and Pagan (1979) that

$$q_{12} = \sum_{p=2}^{P} r_{12,p}^2 \to \chi_{P-1}^2.$$

2.3 Small Sample Correction

Although the introduced test statistic has nice asymptotic properties, in many applications, the sample size is not large enough to safely assume the asymptotic properties. Specifically, the LM statistic has its error in the order of n^{-1} [7], and many Monte Carlo experiments showed that the test rejects the null hypothesis less frequently than indicated by its nominal size [8] [5] [6]. In response, Harris (1985) [7] used Edgeworth expansion to obtain the distribution and moment generating function to order n^{-1} of the LM test statistic. Building on this expansion, Honda (1986) [8] and Cribari-Neto and Ferrari [3] [4] proposed corrections to the critical value or to the test statistic that allows better inference even when the sample size is small while preserving the asymptotic properties.

Honda (1988) provided a closed-form formula to adjust the critical value in the order of $O(n^{-1})$ to correct the size of the test. This adjustment only depends on the covariate, sample size, and the degrees of freedom, but not on the data. The proposed adjustment is a cubic function with respect to C_γ, the critical value at the level of type I error γ, i.e. $P(\chi_{P-1}^2 \geq C_\gamma) = \gamma$, and we refer to this cubic function as g defined as follows.

$$\begin{aligned} g(C_\gamma) &= C_\gamma + C_\gamma \left(\frac{A_3 - A_2 + A_1}{12n(P-1)} \right) + C_\gamma^2 \left(\frac{A_2 - 2A_3}{(12n(P-1)(P+1)} \right) + C_\gamma^3 \left(\frac{A_3}{12n(P-1)(P+1)(P+3)} \right) \quad (8) \\ &= C_\gamma + \tilde{g}(C_\gamma), \end{aligned}$$

where the scalars A_1, A_2, and A_3 follow the notation of Honda (1988) directly.

One of the desirable properties of g would be monotonicity, because regardless of sample size, we would like to maintain the same ordering of the strength of evidence against the null. This turns out to be almost always true in practice. The derivative of $g(C)$ is

$$g'(C_\gamma) = \frac{A_3}{12n(P-1)} \left(\frac{A_3 - A_2 + A_1 + 12n(P-1)}{A_3} + \frac{2(A_2 - 2A_3)}{(P+1)A_3} C_\gamma + \frac{3}{(P+1)(P+3)} C_\gamma^2 \right)$$

A_3 is proven to be strictly positive by definition [3], and we can solve the above quadratic equation to see in which case the derivative is positive. In other words, we can study when the following discriminant is complex.

$$\sqrt{\left(\frac{2(A_2 - 2A_3)^2}{(P+1)A_3} \right)^2 - 4 \cdot \frac{3A_3(A_3 - A_2 + A_1)}{(P+1)(P+3)A_3} - 4 \cdot \frac{3 \cdot 12n(P-1)}{(P+1)(P+3)}}$$

The first two terms inside the square root are $O(1)$ and the last term is $O(n)$, so we can see that the discriminant becomes complex quickly as n increases. Also when the covariates are simulated from normal distribution, $g'(C)$ was always positive unless $n < 3$.

Similar argument is offered in Cribari (1995) [3]. Based on the correction of the critical value in (8), Cribari (1995) proposes to subtract the correction $\tilde{g}(C_\gamma)$ so that

$$P(q_{12} \geq g(C_\gamma)) = P(q_{12} \geq C_\gamma + \tilde{g}(C_\gamma)) = P(q_{12} - \tilde{g}(C_\gamma) \geq C_\gamma).$$

This treats the correction as de-biasing instead of changing the overall shape of the distribution. Although this adjustment of the test statistic corrects the size of the test at a given threshold, it prevents the next, more sophisticated analysis of the test statistic which we introduce in the next section.

Instead, we aim to adjust the test statistic so that the overall shape of null distribution is closer to χ^2_{P-1}. We assume a large enough sample size for monotonicity of g and define the inverse function of g to propose the new adjusted LM statistic $\tilde{q}_{12} = g^{-1}(q_{12})$

$$\gamma = P(\chi^2_{p-1} \geq C_\gamma) = P(q_{12} \geq g(C_\gamma)) = P(g^{-1}(q_{12}) \geq C_\gamma)$$

Our final LM test statistic $\tilde{q}_{12}$ is the real solution to the following equation

$$q_{12} - g(C_\gamma) = 0$$

which is guaranteed to be unique by the monotonicity of g. The cubic equation can be solved both analytically and numerically efficiently given the covariate X.

2.4 Extension to Overall Connectivity

In the previous section, we proposed the LM test statistic q_{12} that tests a pair of variables 1 and 2 to measure the evidence that their correlation changes with respect to the covariate X. Then we made a small sample correction to obtain $\tilde{q}_{12}$ that closely follows χ^2_{P-1} distribution even when sample size is small. As a natural extension, we can repeat the procedure for all variable pairs k_1 and k_2 to obtain $\tilde{q}_{k_1k_2}$. In this section, we propose a way to combine the test statistics to test a new "global" null hypothesis with improved statistical power.

We define the global null hypothesis for variable 1 as follows by extending (5),

$$\boldsymbol{H_0^{(1)}} : \boldsymbol{\alpha^1_{12}} = \boldsymbol{\alpha^1_{13}} = \cdots = \boldsymbol{\alpha^1_{1K}} = 0, \tag{9}$$

where the superscript in $\boldsymbol{H_0^{(1)}}$ indicates that the null hypothesis applies to variable 1, while the superscript in $\boldsymbol{\alpha^1_{1k}}$ indicates the partition of coefficient vector $\boldsymbol{\alpha}$ as in (5). Under $\boldsymbol{H_0^{(1)}}$, no other variables' correlation with variable 1 changes across the different values of X.

Combining the test statistics is a double-edged sword; the procedure can accrue relevant evidence to improve the statistical power, or it can accumulate noise to do the exact opposite. Therefore, we must carefully decide how to combine the test statistics based on the alternative hypothesis we would like to leverage against, and the alternative hypothesis must be constructed reflecting our prior knowledge about the network structure. Chen (2012) discusses two ways to construct the alternative hypothesis. One way, called a sparse alternative, is to test whether only a small number among all tests have non-zero effects while all other tests are null. Another way is to test if at least one test has a non-zero effect size. Chen (2012) focuses on the sparse alternative and proposes the exponential-combination framework [2]. Here, we do not assume that our alternative is "sparse", and we propose a simpler linear combination of the test statistics

$$d_1 = \tilde{q}_{12} + \tilde{q}_{13} + \cdots + \tilde{q}_{1K} = \sum_{k=2}^{K} \tilde{q}_{1k}. \tag{10}$$

We believe combining the test statistics like (10) improves the statistical power of tests for any network whose structure is similar to scale-free topology [9], i.e. where the "hot spot" variables or "hub" variables are connected to a lot of other nodes forming cliques or modules. This is therefore appropriate to apply to the gene co-expression network. We know that transcription factors regulate the gene expression of multiple genes, and if one transcription factor varies with respect to the covariate, the transcriptions of those genes regulated by that transcription factor are likely to be correlated with the covariate as well. The effect sizes for each gene pair may be too small to be detected, but combining them by simple addition like in (10) can form a stronger signal.

In order to test the significance of d_1 against the null hypothesis (9), we need the null distribution of d_1, which is non-trivial because each $\tilde{q}_{1k}$ is correlated to one another although they separately follow χ^2_{P-1}. Therefore, in this section, we acquire a closed-form covariance structure of r defined in (7). First, we begin with a multivariate central limit theorem to write the following in terms of r

$$\boldsymbol{r_{1,p}} = \begin{bmatrix} r_{12,p} \\ r_{13,p} \\ \cdots \\ r_{1K,p} \end{bmatrix} \rightarrow N_{K-1}(\mathbf{0}, H_1), \qquad \forall p = 2, \cdots, P \tag{11}$$

H_1 is a $(K-1) \times (K-1)$ matrix where $(k_1 - 1, k_2 - 1)$th element is $\eta_{1k_1,1k_2}$ for $k_1, k_2 = 2, \cdots K$. From (7), it is easy to see that H_1 has 1 at the diagonals. Also, we use the 4th moments of 3-dimensional multivariate normal distribution and the fact that $\hat{\sigma}_{k_1k_2} \rightarrow \sigma^2_{k_1k_2} = 2\rho_{k_1k_2} + 2$ in probability to show that $\eta_{1k_1,1k_2}$ converges in probability to

$$\frac{6 + 8\rho_{1k_1} + 8\rho_{1k_2} + 4\rho_{k_1k_2} + 2\rho^2_{1k_1} + 2\rho^2_{1k_2} + 2\rho^2_{k_1k_2} + 8\rho_{1k_1}\rho_{1k_2} + 4\rho_{1k_1}\rho_{k_1k_2} + 4\rho_{1k_2}\rho_{k_1k_2}}{8(\rho_{1k_1} + 1)(\rho_{1k_2} + 1)} - \frac{1}{2}. \tag{12}$$

Then, d_1 can be written as the sum of L2 norm of $\boldsymbol{r_{1,p}}$ with a known distribution,

$$d_1 = \sum_{p=2}^{P} \|\boldsymbol{r_{1,p}}\|_2^2 = \sum_{p=2}^{P}\sum_{k=2}^{K} r^2_{1k,p}, \tag{13}$$

and we can derive the distribution of d_1 as well. Let $H_1 = U_1 \Lambda_1 U_1^T$ be the eigen-decomposition of the covariance matrix H_1 in (11), where the diagonal matrix Λ has eigenvalues $\lambda_{12}, \cdots, \lambda_{1K}$ in a decreasing order. Then, we can next consider the transformation $\boldsymbol{r^*_{1,p}} = \boldsymbol{U r_{1,p}}$ that follows normal distribution with diagonal covariance matrix $N_{K-1}(\mathbf{0}, \Lambda_1)$. Note that $\|\boldsymbol{r_{1,p}}\|_2^2 = \|\boldsymbol{r^*_{1,p}}\|_2^2$ due to the orthogonal invariance of L2 norm. Then,

$$\begin{aligned} d_1 &= \sum_{p=2}^{P} {r^*_{12,p}}^2 + \cdots + \sum_{p=2}^{P} {r^*_{1K,p}}^2 \\ &\sum_{p=2}^{P} {r^*_{1k,p}}^2 \sim \Gamma\left(\frac{P-1}{2}, \frac{\lambda_{1k}}{2}\right), \qquad k = 2, \cdots, K \end{aligned}$$

Assuming that we know the true, symmetric, positive definite H, we can acquire positive λ_{1k} for $k = 2, \cdots, K$, and we have expressed the null distribution of d_1 as the sum of distributions of independent gamma variables. We can computationally simulate this null distribution easily. Alternatively, Moschopoulos (1985) [10] provides another interpretation by expressing the cumulative distribution in a form of infinite sum.

Estimation of H: K $<$ n

In (12), we define the element-wise mapping $\phi : \Sigma \rightarrow H$. It is clear from the construction of H that if we can estimate a well-conditioned, symmetric, positive definite correlation matrix $\hat{\Sigma}$, $\phi(\hat{\Sigma})$ is also symmetric and positive definite. As $n \rightarrow \infty$, $\hat{\Sigma}$ converges to Σ, so we can easily acquire the null distribution of d_1.

Estimation of H: K $>$ n

Permutation test.. there are ways to estimate Σ but can't get asymptotic result with good enough error rate. Or maybe show that it works using simulation with our given data? In order to use permutation test, we must show (or justify) that each subject is exchangeable. (Do some data analysis to show that each individual is independent)

References

[1] Trevor S Breusch and Adrian R Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294, 1979.

[2] Lin S Chen, Li Hsu, Eric R Gamazon, Nancy J Cox, and Dan L Nicolae. An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics*, 91(6):977–986, 2012.

[3] Francisco Cribari-Neto and Silvia LP Ferrari. An improved lagrange multiplier test for heteroskedasticity. *Communications in Statistics-Simulation and Computation*, 24(1):31–44, 1995.

[4] Francisco Cribari-Neto and Silvia LP Ferrari. Monotonic improved critical values for two $\chi 2$ asymptotic criteria. *Economics Letters*, 71(3):307–316, 2001.

[5] Leslie G Godfrey. Testing for multiplicative heteroskedasticity. *Journal of Econometrics*, 8(2):227–236, 1978.

[6] WE Griffiths and K Surekha. A monte carlo evaluation of the power of some tests for heteroscedasticity. *Journal of Econometrics*, 31(2):219–231, 1986.

[7] P Harris. An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika*, 72(3):653–659, 1985.

[8] Yuzo Honda. A size correction to the lagrange multiplier test for heteroskedasticity. *Journal of Econometrics*, 38(3):375–386, 1988.

[9] Steve Horvath and Jun Dong. Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*, 4(8):e1000117, 2008.

[10] Peter G Moschopoulos. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathssematics*, 37(1):541–544, 1985.

[11] C Radhakrishna Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44-1, pages 50–57. Cambridge University Press, 1948.