Transition matrix

October 8, 2015

For two matrices describing bipartite networks, $\mathbf{A}$ and $\mathbf{B}$, of identical dimensions $n \times m$, we want some $m \times m$ matrix, $\mathbf{T}$ with columns τ_i, such that $D(\mathbf{AT} - \mathbf{B})$ is minimized "in some way".

We may frame this as a set of m independent regression problems, where m is the number of transcription factors and also the column rank of $\mathbf{A}, \mathbf{B}, \mathbf{T}$. For a column in $\mathbf{B}$, $\mathbf{b}_i$, we note that a corresponding column in $\mathbf{T}$, τ_i represents the solution to

$$E[\mathbf{b}_i] = \tau_{i1}\mathbf{a}_{1i} + \tau_{i2}\mathbf{a}_{2i}$$

or alternatively expressed

$$\begin{bmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \\ \vdots \\ \mathbf{b}_{in} \end{bmatrix} = \tau_{i1} \begin{bmatrix} \mathbf{a}_{11} \\ \mathbf{a}_{21} \\ \vdots \\ \mathbf{a}_{n1} \end{bmatrix} + \tau_{i2} \begin{bmatrix} \mathbf{a}_{12} \\ \mathbf{a}_{22} \\ \vdots \\ \mathbf{a}_{n2} \end{bmatrix} + \dots \tau_{in} \begin{bmatrix} \mathbf{a}_{1n} \\ \mathbf{a}_{2n} \\ \vdots \\ \mathbf{a}_{nn} \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in} \end{bmatrix}$$

$E[e_{ij}] = 0$

This can be solved with the normal equations,

$$\begin{aligned} \tau_i &= \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}_i \\ \mathbf{T} &= [\tau_1, \tau_2, \dots, \tau_n] \end{aligned}$$

Which produces the least square estimate. I.e. loss function

$$L(\mathbf{T}) = \sum_{gene=1}^{N} ||\mathbf{B}_{gene} - \mathbf{A}_{gene}\mathbf{T}||^2$$

is minimized.

This is unbiased and minimizes mean squared error, which is great, but, maybe we can do better. We have many coefficients to measure, most of which are probably not important and there is high colinearity. Perhaps we would rather select a small subset of important off-diagonal points and set all others to zero (or otherwise penalize them). A number of methods are available to explore this. We can express the transition matrix $\mathbf{T}$ as the set of results of m coefficients from m independent LASSO calculations.

Efficiency of estimation

Let $\mathbf{x}_p$ be a Gaussian p-vector representing a sample of gene expression data containing q transcription factors and $p - q$ non-transcription factor genes.

$$\mathbf{x}_p \sim N(\mu, \Sigma)$$

where μ is the p-vector of mean gene expression values and Σ is the $p \times p$ variance-covariance matrix. In this scenario,Σ may be regarded as a combination of two variance-covariance sources- (1) biological signal, (2) biological noise and (2) technical noise.

In investigating gene regulation, many network inference methods are constructed for the estimation of the $p \times q$ subset of Σ pertaining to the effect of the q TFs on the p genes. In identifying drivers of state transitions, we seek to focus on the $q \times q$ matrix of TF-TF effects. We show that our method vastly ourperforms commonly used network inference methods in estimating these specific effects.

Consider a state change between two experimental conditions, A and B, characterized by an alteration of size δ to the biological signal component of the TF-TF variance-covariance matrix at point $\Sigma_{i,j}$ where i and j are indices for two TFs in Σ.

Using a univariate coexpression calculation (the basis for Pearson and WGCNA estimates), the estimated variance of our estimate of δ can be calculated:

$$for\ -\rho_A < \delta < \rho_A, \delta + \rho_A \leq 1$$

$$\begin{aligned}
Var\left(\hat{\rho}_{i,j,A} - \hat{\rho}_{i,j,B}\right) &= Var\left(\hat{\delta}_{cor}\right) \\
& \quad Var\left(\hat{\rho}_{i,j,A}\right) + Var\left(\hat{\rho}_{i,j,B}\right) \\
&= \frac{1-\rho^2_{i,j,A}}{n_A - 2} + \frac{1-\rho^2_{i,j,B}}{n_B - 2} \\
&= \frac{1}{n_A - 2} + \frac{1}{n_B - 2} - \frac{\rho^2_{i,j,A}}{n_A - 2} - \frac{\rho^2_{i,j,B}}{n_B - 2}
\end{aligned}$$

Meanwhile, in condition B the new correlation of TF_i, denoted cor^* with some gene, $gene_k$ $k \in 1, 2 \ldots p$ becomes

$$cor^*\left(TF_i, gene_k\right) = cor\left(TF_i, gene_k\right) + \delta cor\left(TF_j, gene_k\right)$$

The variance of our estimate using the transition matrix can be expressed as follows:

$$\begin{aligned}
Var\left(TM_{i,j}\right) &= Var\left(\hat{\delta}_{TM}\right) \\
&= \frac{\left(\frac{1}{p}\right)\sum_{k=1}^{p} vVar\left(\hat{\rho}_{i,k,A} - \hat{\rho}_{i,k,B}\right)}{\sum_{k=1}^{p}\left(\rho_{j,k} - \bar{\rho}_j\right)^2} \\
&= \frac{\left(\frac{1}{p}\right)\sum_{k=1}^{p}\left[Var\left(\hat{\rho}_{i,k,A}\right) + Var\left(\hat{\rho}_{i,k,B}\right)\right]}{\sum_{k=1}^{p}\left(\rho_{j,k} - \bar{\rho}_i\right)^2} \\
&\leq \frac{\left(\frac{1}{p}\right)\sum_{k=1}^{p}\left[\frac{1}{n_A-2} + \frac{1}{n_B-2}\right]}{\sum_{k=1}^{p}\left(\rho_{j,k} - \bar{\rho}_i\right)^2} \\
&\leq \frac{\frac{1}{n_A-2} + \frac{1}{n_B-2}}{\sum_{k=1}^{p}\left(\rho_{j,k} - \bar{\rho}_i\right)^2} \\
&\leq \frac{Var\left(\hat{\delta}_{cor}\right) + \frac{\rho^2_{i,j,A}}{n_A-2} + \frac{\rho^2_{i,j,B}}{n_B-2}}{\sum_{k=1}^{p}\left(\rho_{j,k} - \bar{\rho}_i\right)^2} \\
&\leq Var\left(\hat{\delta}_{cor}\right) + \frac{Var\left(\hat{\delta}_{cor}\right)\left(1 - \sum_{k=1}^{p}\left(\rho_{j,k} - \bar{\rho}_i\right)^2\right) + \frac{\rho^2_{i,j,A}}{n_A-2} + \frac{\rho^2_{i,j,B}}{n_B-2}}{\sum_{k=1}^{p}\left(\rho_{j,k} - \bar{\rho}_i\right)^2}
\end{aligned}$$

So we have that $Var\left(TM_{i,j}\right) < Var\left(\hat{\delta}_{cor}\right)$ when

$$Var\left(\hat{\delta}_{cor}\right)\left(1 - \sum_{k=1}^{p}\left(\rho_{j,k} - \bar{\rho}_i\right)^2\right) < \frac{\rho^2_{i,j,A}}{n_A - 2} + \frac{\rho^2_{i,j,B}}{n_B - 2}$$

Since each term except $\left(1 - \sum_{k=1}^{p} \left(\rho_{j,k} - \bar{\rho}_i\right)^2\right)$ is strictly non-negative, we see that this inequality holds when

$$\sum_{k=1}^{p} \left(\rho_{j,k} - \bar{\rho}_i\right)^2 < 1$$

Thus, we have a more efficient estimator of δ when

$$p > \frac{1}{Var\left(\rho_{j,k}\right)}$$

In practice, we typically have a large number of genes, p, so that our transition matrix estimator will expected to be dramatically more efficient than the commonly used Pearson or WGCNA estimators.