DRAGON analytic expressions

Kate

2022-05-30

Goal

Equation 8 of the main manuscript decomposes the quantity

$$R = E\left[||\hat{\Sigma} - \Sigma||_F^2\right]$$

into terms which can be efficiently calculated analytically, yielding estimates for optimal regularization parameters λ_1 and λ_2 of DRAGON in the 2-omic case. Here, we derive this equation.

Notation

Let $\Sigma = \{\sigma_{ij}\}$ be the true population covariance and let $S = \{s_{ij}\}$ be sample covariance, i.e., the usual maximum likelihood estimator for Σ. Next, let $\hat{\Sigma} = \{\hat{\sigma}_{ij}\}$ be the covariance shrinkage estimator defined in the main manuscript, i.e.:

$$\hat{\Sigma} = \{\hat{\sigma}_{ij}\} = \begin{pmatrix} (1-\lambda_1)S^{(1,1)} & \sqrt{1-\lambda_1}\sqrt{1-\lambda_2}S^{(1,2)} \\ \sqrt{1-\lambda_2}\sqrt{1-\lambda_1}S^{(2,1)} & (1-\lambda_2)S^{(2,2)} \end{pmatrix} + \begin{pmatrix} \lambda_1 \mathrm{diag}(S^{(1,1)}) & 0 \\ 0 & \lambda_2 \mathrm{diag}(S^{(2,2)})) \end{pmatrix}$$

where we have divided S into blocks consisting of $S^{(1,1)}$, the sample covariance of omics layer 1; $S^{(2,2)}$, the sample covariance of omics layer 2; and $S^{(1,2)} = t(S^{(2,1)})$, the sample covariance between omics layer 1 and omics layer 2.

Derivation

We derive an expression for R in terms of λ_1 and λ_2 by considering three cases:

(i) i, j such that σ_{ij} represents the within-omic covariance of two variables i and j, both in omics layer 1
(ii) i, j such that σ_{ij} represents the within-omic covariance of two variables i and j, both in omics layer 2
(iii) i, j such that σ_{ij} represents the between-omic covariance of variable i, which is in omics layer 1, and variable j, which is in omics layer 2.

In deriving these steps, we partition R into the sum of four components: $R^{(1,1)}$, obtained from $\Sigma^{(1,1)}$ and $S^{(1,1)}$; $R^{(2,2)}$, obtained from $\Sigma^{(2,2)}$ and $S^{(2,2)}$; $R^{(1,2)}$, obtained from $\Sigma^{(1,2)}$ and $S^{(1,2)}$; and $R^{(2,1)}$, obtained from $\Sigma^{(2,1)}$ and $S^{(2,1)}$. From these, we can write the final equation

$$\begin{aligned} R &= R^{(1,1)} + R^{(2,2)} + R^{(1,2)} + R^{(2,1)} \\ &= R^{(1,1)} + R^{(2,2)} + 2R^{(1,2)} \end{aligned}$$

where the last line arises due to the symmetric nature of the covariance matrix and the Frobenius norm.

Case (i): both variables in omics layer 1

In this case, we have

$$
\begin{aligned}
R^{(1,1)} = E\left[||\hat{\Sigma}^{(1,1)} - \Sigma^{(1,1)}||_F^2\right] &= \sum_{i,j} E\left[(\hat{\sigma}_{ij} - \sigma_{ij})^2\right] \\
&= \sum_{i=j} E\left[\left((1-\lambda_1)s_{ij} + \lambda_1 s_{ij} - \sigma_{ij}\right)^2\right] + \sum_{i\neq j} E\left[\left((1-\lambda_1)s_{ij} - \sigma_{ij}\right)^2\right] \\
&= \sum_{i=j} E\left[\left(s_{ij} - \sigma_{ij}\right)^2\right] + \sum_{i\neq j} E\left[\left((1-\lambda_1)s_{ij} - \sigma_{ij}\right)^2\right]
\end{aligned}
$$

Next, we use the standard identity:

$$
\begin{aligned}
Var[U] &= E[U^2] - (E[U])^2 \\
E[U^2] &= Var[U] + (E[U])^2
\end{aligned}
$$

First, when $i \neq j$, let $U = s_{ij} - \sigma_{ij}$. Then we have

$$
\begin{aligned}
R^{(1,1)} &= \sum_{i=j} Var[s_{ij} - \sigma_{ij}] + (E[s_{ij} - \sigma_{ij}])^2 + \sum_{i\neq j} E\left[\left((1-\lambda_1)s_{ij} - \sigma_{ij}\right)^2\right] \\
&= \sum_{i=j} Var[s_{ij}] + \sum_{i\neq j} E\left[\left((1-\lambda_1)s_{ij} - \sigma_{ij}\right)^2\right]
\end{aligned}
$$

where the latter line arises because s_{ij} is an unbiased estimator of σ_{ij}. Next, we consider the term when $i \neq j$, letting $U = (1-\lambda_1)s_{ij} - \sigma_{ij}$.

$$
\begin{aligned}
R^{(1,1)} &= \sum_{i=j} Var[s_{ij}] + \sum_{i\neq j} Var[(1-\lambda_1)s_{ij} - \sigma_{ij}] + (E[(1-\lambda_1)s_{ij} - \sigma_{ij}])^2 \\
&= \sum_{i=j} Var[s_{ij}] + \sum_{i\neq j} (1-\lambda_1)^2 Var[s_{ij}] + ((1-\lambda_1)E[s_{ij}] - \sigma_{ij})^2 \\
&= \sum_{i=j} Var[s_{ij}] + \sum_{i\neq j} (1-\lambda_1)^2 Var[s_{ij}] + \lambda_1^2 (E[s_{ij}])^2 \\
&= \sum_{i=j} Var[s_{ij}] + \sum_{i\neq j} Var[s_{ij}] - 2\lambda_1 Var[s_{ij}] + \lambda_1^2 Var[s_{ij}] + \lambda_1^2 (E[s_{ij}])^2 \\
&= \sum_{i,j} Var[s_{ij}] + \sum_{i\neq j} -2\lambda_1 Var[s_{ij}] + \lambda_1^2 E[s_{ij}^2]
\end{aligned}
$$

Case (ii): both omics variables in layer 2

In this case, we can follow the analogy to the argument above to arrive at

$$
R^{(2,2)} = \sum_{i,j} Var[s_{ij}] + \sum_{i\neq j} -2\lambda_2 Var[s_{ij}] + \lambda_2^2 E[s_{ij}^2]
$$

Case (iii): variable i is in omics layer 1 and variable j is in omics layer 2.

In this case, we derive $R^{(1,2)}$, noting that $R^{(2,1)} = R^{(1,2)}$ by symmetry.

$$\begin{aligned}
R^{(1,2)} &= E\left[||\hat{\Sigma}^{(1,2)} - \Sigma^{(1,2)}||_F^2\right] \\
&= \sum_{i,j} E\left[(\hat{\sigma}_{ij} - \sigma_{ij})^2\right] \\
&= \sum_{i,j} Var[\hat{\sigma}_{ij} - \sigma_{ij}] + (E[\hat{\sigma}_{ij} - \sigma_{ij}])^2 \\
&= \sum_{i,j} (1-\lambda_1)(1-\lambda_2)Var[s_{ij}] + (E[\hat{\sigma}_{ij} - s_{ij} + s_{ij} - \sigma_{ij}])^2 \\
&= \sum_{i,j} (1-\lambda_1)(1-\lambda_2)Var[s_{ij}] + (E[\sqrt{1-\lambda_1}\sqrt{1-\lambda_2}s_{ij} - s_{ij}] + E[s_{ij} - \sigma_{ij}])^2 \\
&= \sum_{i,j} (1-\lambda_1)(1-\lambda_2)Var[s_{ij}] + (\sqrt{1-\lambda_1}\sqrt{1-\lambda_2} - 1)^2(E[s_{ij}])^2
\end{aligned}$$

Now the expression is in terms of the variance and expectation of the sample covariance, which we can easily approximate with moment estimators. We continue manipulating the expression:

$$\begin{aligned}
R^{(1,2)} &= \sum_{i,j} \{1-\lambda_1-\lambda_2+\lambda_1\lambda_2\} Var[s_{ij}] + \left\{(1-\lambda_1)(1-\lambda_2) - 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2} + 1\right\}(E[s_{ij}])^2 \\
&= \sum_{i,j} \{1-\lambda_1-\lambda_2+\lambda_1\lambda_2\} Var[s_{ij}] + \left\{2-\lambda_1-\lambda_2+\lambda_1\lambda_2 - 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2}\right\}(E[s_{ij}])^2
\end{aligned}$$

Next, we again apply the identity $(E[U])^2 = E[U^2] - Var[U]$:

$$\begin{aligned}
&= \sum_{i,j} \{1-\lambda_1-\lambda_2+\lambda_1\lambda_2\} Var[s_{ij}] + \left\{2-\lambda_1-\lambda_2+\lambda_1\lambda_2 - 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2}\right\}(E[s_{ij}^2] - Var[s_{ij}]) \\
&= \sum_{i,j} \left\{1-\lambda_1-\lambda_2+\lambda_1\lambda_2 - (2-\lambda_1-\lambda_2+\lambda_1\lambda_2 - 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2})\right\} Var[s_{ij}] \\
&+ \left\{2-\lambda_1-\lambda_2+\lambda_1\lambda_2 - 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2}\right\} E[s_{ij}^2] \\
&= \sum_{i,j} \left\{-1 + 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2}\right\} Var[s_{ij}] + \left\{2-\lambda_1-\lambda_2+\lambda_1\lambda_2 - 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2}\right\} E[s_{ij}^2] \\
&= -\sum_{i,j} Var[s_{ij}] + 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2}\sum_{i,j}(Var[s_{ij}] - E[s_{ij}^2]) + (2-\lambda_1-\lambda_2+\lambda_1\lambda_2)\sum_{i,j} E[s_{ij}^2])
\end{aligned}$$

Combining cases (i), (ii), and (iii)

Our final expression for R is

$$
\begin{aligned}
R &= R^{(1,1)} + R^{(2,2)} + 2R^{(1,2)} \\
&= \sum_{i,j} Var[s_{ij}] + \sum_{i \neq j} -2\lambda_1 Var[s_{ij}] + \lambda_1^2 E[s_{ij}^2] \\
&+ \sum_{i,j} Var[s_{ij}] + \sum_{i \neq j} -2\lambda_2 Var[s_{ij}] + \lambda_2^2 E[s_{ij}^2] \\
&+ 2 * \left\{ -\sum_{i,j} Var[s_{ij}] + 2\sqrt{1-\lambda_1}\sqrt{1-\lambda_2} \sum_{i,j} (Var[s_{ij}] - E[s_{ij}^2]) + (2 - \lambda_1 - \lambda_2 + \lambda_1\lambda_2) \sum_{i,j} E[s_{ij}^2]) \right\} \\
&= \sum_{i \neq j} -2\lambda_1 Var[s_{ij}] + \lambda_1^2 E[s_{ij}^2] + -2\lambda_2 Var[s_{ij}] + \lambda_2^2 E[s_{ij}^2] \\
&+ 4\sqrt{1-\lambda_1}\sqrt{1-\lambda_2} \sum_{i,j} (Var[s_{ij}] - E[s_{ij}^2]) \\
&+ 2 * (2 - \lambda_1 - \lambda_2 + \lambda_1\lambda_2) \sum_{i,j} E[s_{ij}^2]
\end{aligned}
$$

Combining like terms for functions of λ_1 and λ_2, we can write

$$
\begin{aligned}
R &= \sum_{i \neq j} -2\lambda_1 Var[s_{ij}] + \lambda_1^2 E[s_{ij}^2] + -2\lambda_2 Var[s_{ij}] + \lambda_2^2 E[s_{ij}^2] \\
&+ 4\sqrt{1-\lambda_1}\sqrt{1-\lambda_2} \sum_{i,j} (Var[s_{ij}] - E[s_{ij}^2]) \\
&+ 2 * (2 - \lambda_1 - \lambda_2 + \lambda_1\lambda_2) \sum_{i,j} E[s_{ij}^2] \\
&= 4 \sum_{i,j} E[s_{ij}^2 \\
&+ \lambda_1 \left\{ -2 \sum_{i \neq j} Var[s_{ij}] - 2 \sum_{i,j} E[s_{ij}^2] \right\} \\
&+ \lambda_2 \left\{ -2 \sum_{i \neq j} Var[s_{ij}] - 2 \sum_{i,j} E[s_{ij}^2] \right\} \\
&+ \lambda_1^2 \sum_{i \neq j} E[s_{ij}^2] \\
&+ \lambda_2^2 \sum_{i \neq j} E[s_{ij}^2] \\
&+ \lambda_1\lambda_2 \left\{ 2 \sum_{i \neq j} E[s_{ij}^2] \right\} \\
&+ \sqrt{1-\lambda_1}\sqrt{1-\lambda_2} \left\{ 4 \sum_{i,j} (Var[s_{ij}] - E[s_{ij}^2]) \right\}
\end{aligned}
$$

This expression is of the form

$$
R = \text{const.} + \lambda_1 T_1^{(1)} + \lambda_2 T_1^{(2)} + \lambda_1^2 T_2^{(1)} + \lambda_1^2 T_2^{(2)} + \lambda_1\lambda_2 T_3 + \sqrt{1-\lambda_1}\sqrt{1-\lambda_2} T_4
$$

where the first term is a constant with respect to λ_1 and λ_2 and the remaining terms T are defined as:

$$T_1^{(k)} = -2\left(\sum_{i\neq j} Var[s_{ij}] + \sum_{i,j} E[s_{ij}^2]\right); k = 1, 2$$

$$T_2^{(k)} = \sum_{i\neq j} E[s_{ij}^2]; k = 1, 2$$

$$T_3 = 2\sum_{i,j} E[s_{ij}^2]$$

$$T_4 = 4\sum_{i,j}(Var[s_{ij}] - E[s_{ij}^2])$$

Remaining to do

These sums are only over a subset of the i and j, that correspond to e.g. omics 1 and omics 2. We need to add some notation to reflect this.