Reproducibility in the Simulation

This file contains instructions for reproducing the results and figures in the simulation. We assume the working directory has been appropriately set.

**Step 1: "Step1\_Data\_generation.R"**

Generate the data in the simulation with 10 replications. The outputs of this step are " 20201205\_Sigma.RData"--"20201214\_Sigma.RData" /, which will be used in the following simulation and comparation step.

In the " 20201205\_Sigma.RData"--" 20201214\_Sigma.RData" (10 ".RData" files):

* Sigma.k: True cell type gene expression covariance matrix, the dimension of which is (G,G,K), where G is the number of genes and K is the number of cell types.
* Corr.true: True gene expression correlation matrix, the dimension of which is (G,G,n).

In the "20201205\_Sim\_network.RData"--"20201214\_Sim\_network.RData" (10 ".RData" files):

* X: Gene expression matrix
* cell.info: Cell information matrix with n (number of cells) rows and 3 columns, the first column of which is the cell type of cells, the last two columns corresponds to X coordinates and Y coordinates of cells.

**Step 2: "Step2\_Simulation.R"**

Based on the output of Step 1, we applied two-step algorithm, CTS, CSN-joint and CSN-separate on the ten replication datasets, and got results of the four methods. Notice that CSN-joint and CSN-separate were performed based on the matlab codes, which is available at <https://github.com/wys8c764/CSN>. The outputs of Step 2 are "methods\_error.RData", "CSN\_result\_ct\_integrate.RData", "WGCNA\_20201205.RData"--"WGCNA\_20201214.RData", CSN\_result.RData", "CTS\_20201205.RData"--"CTS\_20201214.RData",

"Result\_20201205\_0.1\_70.RData"-- Result\_20201214\_0.1\_70.RData".

In the "methods\_diff.RData":

* sum.diff.ct, a vector of errors in each cell of method CTS.
* sum.diff.sim, a vector of errors in each cell of two-step algorithm.
* sum.diff. csn, a vector of errors in each cell of CSN-joint.
* sum.diff.csnct, a vector of errors in each cell of CSN-separate.
* sum.diff.wgcna, a vector of errors in each cell of WGCNA.

In the "CSN\_result.RData":

* cs.Sgm, a list contains 10 elements, each of which is an array of gene co-expression matrix, the dimension of which is (G,G,n). It’s an estimation of CSN-joint method for 10 replications.

In the "CSN\_result\_ct\_integrate.RData":

* cs.Sgm.celltype, a list contains 10 elements, each of which is an array of gene co-expression matrix, the dimension of which is (G,G,n). It’s an estimation of CSN-separate method for 10 replications.

In the "Result\_20201205\_0.1\_70.RData"-- Result\_20201214\_0.1\_70.RData":

* Sparse.Corr, estimated Sparse gene correlation matrix by two-step algorithm, the dimension of which is (G,G,n).

In the "CTS\_20201205.RData"--"CTS\_20201214.RData":

* ct.Corr, estimated Sparse gene correlation matrix by CTS, the dimension of which is (G,G,n).

"WGCNA\_20201205.RData"--"WGCNA\_20201214.RData",

* wgcna.Corr, estimated Sparse gene correlation matrix by WGCNA, the dimension of which is (G,G,n).

**Step 3: "Step3\_Figure\_Table.R"**

Draw figure 2 and table 1 using the output of step 1, step 2 and other ".RData" files, such as "time\_violin.RData", "ROC\_curve.RData" and "Prediction.RData".

In "time\_violin.RData":

* time.ctn, 10 dimension vector of computation time of CTS.
* time.wgcna, 10 dimension vector of computation time of WGCNA
* time.sim, 10 dimension vector of computation time of two-step algorithm.
* time.csn, 10 dimension vector of computation time of CSN-joint.
* time.csnct, 10 dimension vector of computation time of CSN-separate.

In "ROC\_curve.RData":

* total.FPR.sim, a vector of FPRs of two-step algorithm.
* total\_FPR\_ct, a vector of FPRs of CTS.
* total\_FPR\_csn, a vector of FPRs of CSN-joint.
* total\_FPR\_csnct, a vector of FPRs of CSN-separate.
* total\_FPR\_csnct, a vector of FPRs of WGCNA.
* total\_TPR\_csnct, a vector of FPRs of WGCNA.
* total.TPR.sim, a vector of TPRs of two-step algorithm.
* total\_TPR\_ct, a vector of TPRs of CTS.
* total\_TPR\_csn, a vector of TPRs of CSN-joint.
* total\_TPR\_csnct, a vector of TPRs of CSN-separate.

In "Prediction.RData":

* est.miss, estimated gene co-expression matrix by two-step algorithm of the missing cells, and the dimension of which is (G,G,50).
* Corr.miss, the true gene co-expression matrix of missing cells, and the dimension of which is (G,G,50).

**"ROC\_Data.R"**

Generate "ROC\_curve.RData" if readers are interested.

**"prediction.R"**

Based on the results of Step 2, generate missing cells and predict gene co-expression networks via two-step algorithm.