How original UTMOST works:

For each gene, cut the entire data into five parts. Let’s say, ABCDE.

1. Hyperparameter tuning (5-fold):

1.1 In the first fold, taking ABC and D as the training and tuning set (for early stop), respectively. The testing set E will not be used in here.

1.2 Single tissue prediction model was trained by elastic net for each tissue independently (5-fold cross-validation in ABC).

1.3 A joint cross-tissue group-LASSO prediction model was built for each hyperparameter pairs (lambda1 and lambda2, five values for each lambda. i.e. 25 combinations). The range of lambda was copied from the single tissue lambda setting. In the optimization step, the parameters (beta) were initialized from the weights of the single tissue model with lowest cross-validation error. The optimization will stop if the new training error (error in ABC) or the new tuning error (error in D) was higher than the old ones from the previous step. Tuning error will be saved for each lambda pairs. The best lambda pair (for this fold) was chosen according to the average tuning error across all the tissues.

1.4 Same procedure will be conducted by taking BCD, CDE, DEA, EAB as training set and taking E, A, B, C as tuning set, respectively. After the 5-fold training, 5 best lambda combinations were generated.

2. Training the model using entire data.

2.1 Single tissue elastic net model was trained by 5-fold cross-validation using entire data.

2.2 The joint cross-tissue group-LASSO was performed by applying each of the five best lambda pairs. The optimization was initialized by the parameters (beta) from the single tissue model (see 2.1). The iteration will stop when the new training error was greater than the old ones, or the error is very small between the two steps. The final lambda pair will be picked up by taking the one with lowest average error across all the tissues in the entire data. The model with parameters (beta) from the last iteration (not the last but two) will be considered as the final model for downstream analysis. (not sure why, the training error should be higher in the last iteration than the last but two.)

3. Prediction accuracy evaluation

The prediction quality was measured by applying the final model to the entire data. Gene with r>0.1 and P<0.05 will be considered as imputable gene (iGene).

Mod 1 UTMOST:

For each gene, cut the entire data into five parts. Again, ABCDE.

1. Hyperparameter initialization by single tissue elastic net.

Prediction models were trained by 5-fold cross-validation elastic net using the entire data for each tissue independently. The range of lambdas for cross-tissue prediction was set up matching the range of single tissue lambda. So, 25 lambda pairs were generated.

2. Hyperparameter tuning and model training (5-fold):

2.1 In the first fold, taking ABCD and E as the training and tuning set (for early stop), respectively.

2.2 Single tissue prediction model was trained using elastic net for each tissue independently (5-fold cross-validation in ABCD).

2.3 The joint model was trained using each of the 25 lambda pairs in the training set (ABCD). The optimization was initialized by single tissue weights generated in the current fold. The optimization will stop if the new training error (error in ABCD) or the new tuning error (error in E) was higher than the old ones from the previous step. Tuning error and the predicted expression level in E (25 models for each tissue) will be saved for each lambda pairs.

2.4 Same procedure will be conducted by taking BCDE, CDEA, DEAB, EABC as training set and taking A, B, C, D as tuning set, respectively. After the 5-fold training, one of the 25 lambda pairs will be picked up as the best lambda pair according to the average tuning error across the five folds.

3. Training the model using entire data.

The joint training was performed by applying the best lambda combination. The optimization was initialized by the parameters (beta) from the single tissue model using entire data (step 1). The iteration will stop when the new training error was greater than the old ones, or the error is very small between the two steps.

4. Prediction accuracy evaluation

The imputation accuracy was estimated by measuring the correlation of the observed expression level and the predicted expression level calculated in tuning set. i.e. The predicted expression levels were generated using five different models with same hyperparameters (the best lambda pair). Gene with r>0.1 and P<0.05 will be considered as imputable gene (iGene).

Mod 2 UTMOST:

For each gene, cut the entire data into five parts. Once more, ABCDE.

1. Hyperparameter initialization by single tissue elastic net.

Prediction models were trained by 5-fold cross-validation elastic net using the entire data for each tissue independently. The range of lambdas for cross-tissue prediction was set up matching the range of single tissue lambda. So, 25 lambda pairs were generated.

2. Hyperparameter tuning and model training (5-fold):

2.1 In the first fold, taking ABC and D as the training and tuning set (for early stop), respectively. E will be the test set.

2.2 Single tissue prediction model was trained using elastic net for each tissue independently (5-fold cross-validation in ABC).

2.3 The joint model was trained using each of the 25 lambda pairs in the training set (ABC). The optimization was initialized by single tissue weights generated in the current fold. The optimization will stop if the new training error (error in ABC) or the new tuning error (error in D) was higher than the old ones from the previous step. Tuning error and the predicted expression level in the test set E (25 models for each tissue) will be saved for each lambda pairs.

2.4 Same procedure will be conducted by taking BCD, CDE, DEA, EAB as training set and taking E, A, B, C as tuning set and taking A, B, C, D as testing set, respectively. After the 5-fold training, one of the 25 lambda pairs will be picked up as the best lambda pair according to the average error in the test set across the five folds.

3. Training the model using entire data.

The joint training was performed by applying the best lambda combination. The optimization was initialized by the parameters (beta) from the single tissue model using entire data (step 1). The iteration will stop when the new training error was greater than the old ones, or the error is very small between the two steps.

4. Prediction accuracy evaluation

The imputation accuracy was estimated by measuring the correlation of the observed expression level and the predicted expression level calculated in the test set. i.e. The predicted expression levels were generated using five different models with same hyperparameters (the best lambda pair). Gene with r>0.1 and P<0.05 will be considered as imputable gene (iGene).