# Missing Value Imputation Summary:

Missing data in untargeted MS-based metabolomics data occur for various reasons. First, it is possible that the molecules are truly absent from the sample, a situation that may occur e.g. for drug metabolites that only appear in a subset of people taking that medication. On the other hand, there are several technical reasons that could result in missing values, including: (i) instrument sensitivity thresholds, below which concentrations of a specific metabolite might not be detectable in a sample (i.e., below the limit of detection, LOD); (ii) matrix effects that impede the quantification of a metabolite in a sample through other co-eluting compounds and ion suppression; (iii) declining separation ability of the chromatographic column and increasing contamination of the MS instrument; and (iv) limitations in computational processing of spectra, such as poor selection and alignment of the spectral peaks across samples. A widely used and flexible class of missing data strategies is imputation, which involves the replacement of missing values by reasonable substitute values. The most commonly used imputation approaches for metabolomics data assume that missing data occur because they are below the limit of detection (left-censoring, a variant of MNAR).[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6153696/].

### Input Summary:

Input Dataset: report -> e.csv

Output Datasets and Files: report -> Missing Value Imputation -> result\_summary.csv -- e.csv

**- Defination of Missing Values:** . What in your dataset can be treated as missing values. "Empty cells" is where there is no value in the excel sheet. "Characters" is where the value is a character but not a number. "Zeros" is when the value is 0. "Negative Values" is where the value is less than 0. Or you can also cusomize a threshold to determine the missing value by selecting the "Values Less Than...". Any value in the excel sheet less than the VALUES LESS THAN will be treated as missing values.

**- Values Less Than:** . The the "values less than..." is selected, any value less than 500 will be treated as missing value.

**- Remove Compounds with Too Many Missing Values:** When the Remove compounds... is checked. The compounds with more than 50% of missing values will be eliminated before imputation.

**- Missing Value Imputation Method (with a disturb):**

replace by half minimum: replace the missing values with random samples from a normal distribution with half-minimum of non-missing value as mean and 1/10 of half-minimum of non-missing value as standard deviation for each compound.

replace by 1/10 minimum: replace the missing values with random samples from a normal distribution with 1/10-minimum of non-missing value as mean and 1/10 of 1/10-minimum of non-missing value as standard deviation for each compound.

replace by minimum: replace the missing values with random samples from a normal distribution with minimum of non-missing value as mean and 1/10 of minimum of non-missing value as standard deviation for each compound.

replace by mean: replace the missing values with random samples from a normal distribution with mean of non-missing value as mean and 1/10 of mean of non-missing value as standard deviation for each compound.

replace by median: replace the missing values with random samples from a normal distribution with median of non-missing value as mean and 1/10 of median of non-missing value as standard deviation for each compound.

imputed by kNN: replace the missing values by k Nearest Neighbors Imputation method. The original kNN imputation was developed for high-dimensional microarray gene expression data (n<p, n is the number of samples, and p is the number of variables). For each gene with missing values, this method finds the k nearest genes using Euclidean metric and imputes missing elements by averaging those non-missing elements of its neighbors. In metabolomics studies, we applied kNN to find k nearest samples instead and imputed the missing elements.

imputed by kNN: replace the missing values by k Nearest Neighbors Imputation method. The original kNN imputation was developed for high-dimensional microarray gene expression data (n<p, n is the number of samples, and p is the number of variables). For each gene with missing values, this method finds the k nearest genes using Euclidean metric and imputes missing elements by averaging those non-missing elements of its neighbors. In metabolomics studies, we applied kNN to find k nearest samples instead and imputed the missing elements.

Bayesian PCA missing value estimation (BPCA): see [https://rdrr.io/bioc/pcaMethods/man/bpca.html]

probabilistic PCA (PPCA): see [https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5169998]

Single Value Decomposition imputation (SVD): SVD imputation will initialize all missing elements with zero then estimate them as a linear combination of the k most significant eigen-variables iteratively until reaches certain convergence threshold. In metabolomics data, we scaled and centralized the data matrix first and then applied this imputation approach with the number of PCs setting to five.

missForest (Imputation with Random Forest): This imputation method applies random forest, a powerful machine learning algorithm, to build a prediction model by setting particular target variable with non-missing values as the outcome and other variables as predictors, then to predict the target variable with missing values iteratively.

QRILC (Quantile Regression Imputation of Left-Censored data): QRILC imputation was specifically designed for left-censored data, data missing caused by lower than LOQ. This method imputes missing elements with randomly drawing from a truncated distribution estimated by a quantile regression. A beforehand log-transformation was conducted to improve the imputation accuracy.

### Result Summary:

The defination of missing values are empty cells, characters. In total, there are 1113 missing values in the dataset. Compounds with more than 50% missing values were excluded. In total, there were 21 compounds excluded. Then the missing values in the dataset were imputed by the method of replace by half minimum. For more details, please see Table 1.

Table Explanation.

- index: the index of compounds, mainly for sorting the table.

- label: compound labels.

- empty.cells: the number of empty cell for each compound.

- characters: the number of character values for each compound.

- total: the total number of missing values for each compound.

- excluded: if Yes, the compound is excluded.

| label | empty.cells | characters | total | excluded |
| --- | --- | --- | --- | --- |
| Ceramide (d34:2) | 40 | 0 | 40 | Yes |
| 2-(2H-Benzotriazol-2-yl)-4,6-bis(1-methyl-1-phenylethyl)phenol | 40 | 0 | 40 | Yes |
| LPE (17:1) | 38 | 0 | 38 | Yes |
| SM (d30:1) | 38 | 0 | 38 | Yes |
| TG (54:3); TG(16:0/18:1/20:2); | 38 | 0 | 38 | Yes |
| Ser-Tyr-Lys | 38 | 0 | 38 | Yes |
| SM (d32:0) | 37 | 0 | 37 | Yes |
| 3-Chlorotyrosine | 37 | 0 | 37 | Yes |
| TG (56:1) | 36 | 0 | 36 | Yes |
| SM (d44:2) | 33 | 0 | 33 | Yes |

Table 1: several compounds with most missing values.

# Student t-test Summary:

Two Student t-tests show which compounds have the power to differentiate the different two-groups in the data set. The two sample t-test is applied to one metabolite at a time (i.e., a univariate analysis) to determine whether the mean values of the two groups are different. The null hypothesis for the test is H0 : mu\_group1 = mu\_group2, and the alternative hypothesis is Ha : mu\_group1 != mu\_group2. If the p-value for the test is smaller than a cutoff value, typically 0.05, the null hypothesis is rejected. If the p-value is large, there is no significant difference between the mean values for the two groups, indicating the metabolite has little power to separate them.

The 0.05 cutoff value is often used when the t-test for a metabolite is examined individually, without considering the tests for other metabolites. A multiple comparison procedure can be employed, in which a smaller cutoff value is used, to control the overall error caused by using all the t-tests together.

### Input Summary:

Input Dataset: report -> Missing Value Imputation -> e.csv

Output Datasets and Files: report -> Missing Value Imputation -> Student t-test -> student\_t\_test.csv

**- Treatment Group:** Gender. The student t-test will be performed on each compound to detect those significantly altered by Gender.

**- Variance Equality Assumption:** FALSE. If TRUE, the equality of variance is assumed, and the tests are Student's t-test, otherwise Welch t-test.

**- Correct the False Discovery Rate:** Benjamini -- Hochberg (1995). When conducting multiple tests, the rate of incorrectly reject a null hypothesis will increase. FDR-controlling procedures are designed to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections). The suggested method for metabolomics is the Benjamini -- Hochberg procedure. For more information, please visit https://en.wikipedia.org/wiki/False\_discovery\_rate.

### Result Summary:

Welch t-test (not assuming equal variance) was performed on each compound to test if the mean average of Female different from Male. Out of 725 compounds, 384 are significant with p-value less than 0.05. To control the false discovery rate (FDR), the Benjamini -- Hochberg (1995) procedure was used. After FDR correction, 350 compounds are still significant. See Table 2 for more detail.

Table Explanation.

- index: the index of compounds, mainly for sorting the table.

- label: compound labels.

- p\_values: p-values from t-tests.

- p\_values\_adjusted: p-values adjusted by the FDR correction procedure.

| label | p\_values | p\_values\_adjusted |
| --- | --- | --- |
| LysoPC (19:0); | 1.404057e-15 | 7.070086e-13 |
| PC (38:2) | 2.746181e-15 | 7.070086e-13 |
| PC (35:3); PC(17:1/18:2); | 2.925553e-15 | 7.070086e-13 |
| SM (d39:1); SM(d14:0/25:1); | 4.910669e-15 | 8.900588e-13 |
| SM (d40:2) B | 2.356200e-14 | 3.416490e-12 |
| SM (d40:1) | 6.071633e-14 | 7.336557e-12 |
| SM (d39:1) | 7.750593e-14 | 8.027400e-12 |
| PI (36:3); PI(18:1/18:2); | 1.050852e-13 | 9.523344e-12 |
| PC (37:1); PC(16:0/21:1); | 1.250757e-13 | 1.007554e-11 |
| PC (42:6); PC(20:0/22:6); | 4.313461e-13 | 3.127259e-11 |

Table 2: most significant compounds (i.e. small p-values)