Bayesian statistical modeling reveals missing value mechanisms in label-free Mass Spectrometry-based proteomics experiments

Devon Kohler and Olga Vitek

Liquid chromatography coupled with Mass Spectrometry (LC-MS/MS)-based proteomics is the tool of choice for experimenters interested in quantitatively measuring the intensity of proteins in a sample. While broadly popular, technological and biological variations in the experiment can lead to missed measurements for proteins across MS-runs. These missing values create challenges in downstream analysis and can have a major impact on the protein-level conclusions derived from the experiment. Many strategies have been developed to deal with missing values; from simply removing feature’s that have an overwhelming number of missing values, to attempting to predict and impute the missing data. Missing value imputation is complicated by the differing underlying mechanisms which are the root cause of the missingness. The two main mechanisms that cause missingness are defined as Missing Completely at Random (MCAR), for example due to the stochastic nature of data acquisition, and Missing Not at Random (MNAR), mainly caused by the protein’s intensity being too low to measure in the MS run. Different imputation strategies exist which target each of these mechanisms. K-nearest neighbors (KNN) finds similar MS runs and utilizes them to impute the missing value. The accelerated failure time (AFT) model, widely used in MSstats, assumes values are MNAR and leverages run and feature level changes to impute the low value features. Other methods, such as MSImpute, use frequentist approaches to attempt to disentangle the MCAR vs MNAR mechanisms and try to determine a more accurate imputation compared to focusing on a single mechanism.

In this paper we propose a Bayesian modeling approach to imputing missing values in MS-based proteomics experiments. The approach allows us to encode the missing mechanisms directly into our model and leverage the observed values in the experiment to infer the probability the value is MCAR or MNAR. Once the reason for missingness is inferred, we can more accurately impute the value via a variety of existing methods, such as sampling from the inferred distribution of the protein in cases of MCAR or utilizing an AFT approach in cases of MNAR. A major advantage of the proposed approach over existing frequentist approaches is the ability to specify an informed prior probability over our variables, allowing researchers to leverage their extensive prior knowledge as input into the model. The method is implemented in the probabilistic programming Pyro, which utilizes stochastic variational inference to learn the parameters of the model given the priors. We compare the approach to popular existing imputation methods using both simulated and real-world experimental data and benchmark the performance of each method. Additionally, we assess the impact of the approach on differential analysis utilizing spike-in experimental datasets with known ground truth.