CHALMERS
UNIVERSITY OF TECHNOLOGY

Tackling Missing Values in Mass Spectrometry-based Proteomics Data

Master's thesis in Engineering Mathematics and Computational Science

LOUISE LEONARD

Data Sciences & Quantitative Biology
AstraZeneca
Gothenburg, Sweden 2021

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2021

MASTER'S THESIS 2021

Tackling Missing Values in Mass Spectrometry-based Proteomics Data

LOUISE LEONARD

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021

Tackling Missing Values in Mass Spectrometry-based Proteomics Data
LOUISE LEONARD

Supervisor:
Natalie van Zuydam, Data Sciences & Quantitative Biology, AstraZeneca

Examiner:
Erik Kristiansson, Mathematical Sciences, Chalmers University of Technology

Master's Thesis 2021
Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

Tackling Missing Values in Mass Spectrometry-based Proteomics Data
LOUISE LEONARD
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

In the development of therapeutics, analysis of differentially abundant proteins (DAPs) using mass spectrometry (MS) is essential. However, MS-based data suffers from high rates of missing values that severely complicate downstream analyses. Various imputation methods have been proposed to deal with the missing data, but there is no standard protocol for selecting a method. Here we have comprehensively evaluated common methods, to develop a best practice for imputation to inform downstream statistical analyses of MS proteomics data. We compared the performance of five imputation methods in their application to values missing completely at random and missing not at random introduced into data from the Cancer Cell Line Encyclopedia, and data simulated from a multivariate mixed-effects model respectively. Performance was measured in true positive rate (TPR) and false positive (FPR) of detected DAPs ($P_{\text{adj}} \leq 0.05$, est. $\log_2$ fold-change > 1, and an accuracy metric $[Q] < 10^3$). The FPR was below 5% for all methods under all conditions tested. If less than 10% of the data was missing, imputation did not increase the TPR compared to removing missing values. For 30% missingness irrespective of data or missingness type, the TPR was below 80%; and for 50% missingness the TPR was 25-75% depending on imputation method. Since the FPR was controlled, no artefacts were introduced by any methods under any circumstances. For large proportions of missingness (50%), we recommend imputation with Principal Component Analysis imputation if the sample size is large ($n > 50$). With small sample sizes ($n = 10$) or small proportions of missingness (10%), imputation is advised against.

KEYWORDS: imputation, missing data, mass spectrometry, multivariate mixed-effects models, differential abundance, proteomics.

Acknowledgements

I would like to thank AstraZeneca for the opportunity of writing this thesis at the company in collaboration with Chalmers University of Technology.

I would like to express my gratitude to my supervisor Natalie van Zuydam for her contribution of extensive knowledge, support, and feedback. Thank you for the continuous encouragement and commitment to bring my work to a higher level. Your assistance has been invaluable for the completion of this thesis.

I would like to thank my academic supervisor and examiner Erik Kristiansson. I truly appreciate your advice and guidance during this time.

I also wish to extend my gratitude to Steven Novick, for contributing with his expertise, constructive feedback and valuable advisement. Moreover, I would like to thank Peter Konings, Tom Marlow, Esha Mohamed and Avijit Singh, for warmly welcoming me to the team, and for the support I have received throughout my work.

My gratitude further extends to the members of the Quantitative Biology department at AstraZeneca. I thankfully acknowledge the contribution and support from Andrew Jarnuczak, Bairu Zhang, Stephanie Ashenden, Tiago Jose Paschoal Sobreira, Ian Barret and Claus Bendtsen.

Finally, I thank my family and friends for their support during this process.

Louise Leonard, Gothenburg, March 1, 2021

Contents

List of Abbreviations

CCLE	Cancer cell line encyclopedia
DAP	Differentially abundant protein
FDR	False discovery rate
FPR	False positive rate
kNN	k-Nearest neighbours
LMM	Linear mixed-effects model
MAR	Missing at random
MCAR	Missing completely at random
MICE	Multivariate imputation by chained equations
MNAR	Missing not at random
MS	Mass spectrometry
PCA	Principal component analysis
PEM	Penalized expectation maximum algorithm
RF	Random forest
TPR	True positive rate

Glossary

comparative statistical analysis analyses where two or more datasets are compared to determine their (in)consistency with one another.
complete data a dataset with no missing values.

differentially abundant proteins proteins that are significantly different in their abundance in two groups of samples.

ignorable missingness a property where inference based on approaches that ignore the missingness mechanism do not introduce bias.
imputation the process of substituting missing values with (imputed) values inferred from available information in the data.
incomplete data a dataset with some values missing.

mass spectrometry a technique used to quantitate proteins or other metabolites.
missing at random missingness mechanism that depends only on observed data.
missing completely at random missingness mechanism that does not depend on values in the data.
missing not at random missingness mechanism that depends on unobserved values.
missingness the manner in which data are missing from a dataset.
missingness mechanism the underlying probability distribution of missing values.

protein abundance the quantity or amount of a protein in a given sample.
proteomics the scientific field of large-scale determination of gene and cellular function directly at protein level.

List of Figures

List of Tables

1

Introduction

Proteomics is the scientific field of large-scale determination of gene and cellular function directly at protein level. It covers the exploration of protein composition, structure, and activity by identification and quantitation of proteins. Mass spectrometry (MS) is a widely used technique in quantitative proteomics, offering advantages such as high throughput and coverage. Its ability to rapidly quantitate and determine the abundance of thousands of proteins has made it one of the most important tools in proteomics [1].

MS data is commonly used for comparison of protein abundance profiles between treatment groups. In such studies, comparative statistical analysis is carried out with the aim to identify proteins that have significantly different relative abundance in the two treatment groups, so called differentially abundant proteins (DAPs). In the development of therapeutics that degrade specific target proteins, this type of MS-based studies plays an essential role. In the early stages of the drug development pipeline, crucial steps are validation of target engagement and profiling of any toxicity related to off-target effects of the drug candidate. Misclassification at this stage may have large effects downstream in the pipeline. In later stages of the drug development, it is important to determine that therapeutics show specific degradation of the target protein and to quantify any-off target effects in response to treatment [2].

One crucial problem associated with MS-based data is the high proportion of missing values, where the abundances of some proteins in some samples are not detected. The missing data may have a huge impact on the ability to detect DAPs, as simply ignoring or removing variables with missing values severely decrease the statistical power. In addition, the missing values may bias the downstream analysis. In order to understand whether a compound treatment has been successful, it is therefore important to identify and handle any existing missingness in the data.

Currently there is no standard protocol for pre-processing of MS data in order to control for the effects of missing values [3, 4]. Several imputation methods have been suggested, where the missing values are substituted with values inferred from available data. A number imputation methods have been evaluated in terms of imputation accuracy in MS data [5, 6, 7], but to our knowledge there is no comprehensive evaluation on how these methods affect the ability to detect DAPs.

1.1 Aim & objectives

The aim of this thesis is to evaluate the performance of common imputation methods in terms of the ability to detect differentially abundant proteins in MS data. The evaluation is intended to establish suggestions on how missing values should be tackled in future comparative MS-based studies, considering the type of missingness, the sample size and the proportion of missing data.

Specifically, the issues that are investigated are:

- How do the different types of missing values affect the subsequent analysis of MS-based proteomics data?
- Do currently used imputation methods adequately address the effect of missing data on subsequent data analysis?
- In what ways do sample size, type of missingness and amount of missing data affect the performance of the imputation methods?
- Given these findings, can suggestions be established regarding under what circumstances and by which method imputation should be performed?

To answer these questions, comparative proteomics analyses were conducted, using publicly available proteomics data as well as simulated data. Missing values were introduced at different proportions by two different missingness mechanisms, allowing for comparison of different types of missingness. The missing entries were then imputed using a set of imputation strategies. Finally, comparative differential analyses were conducted for complete, incomplete and imputed data and the performance for each distinct dataset was evaluated.

1.2 Scope

In this thesis we have limited the evaluation on strategies for dealing with missing data to imputation. Other existing strategies, such as using a zero-inflated distribution to model the abundance data, have not been considered. We have selected five commonly used imputation methods that have been shown to yield high imputation accuracy (i.e. low root mean squared error between imputed values and unobserved value) and comprehensively evaluate the effect these methods have on the ability to detect DAPs [5, 6, 7]. The selected methods are; k-Nearest Neighbors (kNN) imputation, Random Forest (RF) imputation, Multivariate Imputation by Chained Equations (MICE), Penalized Expectation Maximum (PEM) imputation and Principal Component Analysis (PCA) imputation.

2

Background

2.1 Mass spectrometry-based proteomics

Mass spectrometry-based proteomics experiments begin with protein extraction from tissues or cells cultures. The extracted proteins are then digested into peptides by enzymatic- or chemical reactions. Mixtures of peptides may then be separated by liquid chromatography (LC) in order to ensure that only a small proportion of peptides in a sample are introduced to the mass spectrometer at a given time. This allows for greater precision in the detection of peptides [8]. By definition, the mass spectrometer consists of three main parts; an ion source, a mass analyzer, and a detector (Figure 2.1). For each part there are several existing methods using various techniques. In the first step, peptides are ionized and evaporated by the ion source. The peptide-ions are then sorted and separated according to their mass-to-charge (m/z) ratio by the mass analyzer. The detector registers the ion-intensity at each m/z-ratio. The output of MS-experiments is reported as m/z-spectra, with peaks at m/z-ratios corresponding to peptides or fragments in the sample under analysis. From peptide intensities the relative protein abundances can be derived, by pairing peptides to their origin protein. Finally, a dataset of relative protein abundances for the sample(s) under analysis is obtained. Usually the number of proteins are several thousand [1].

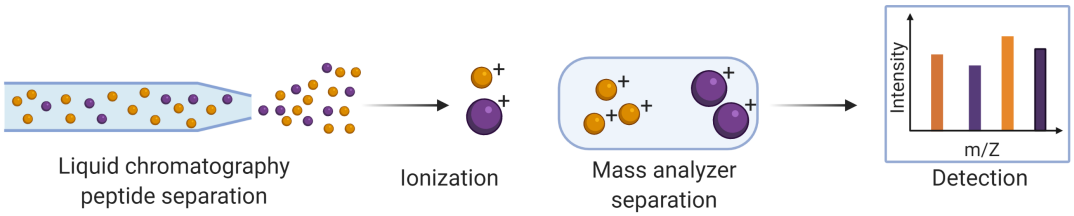


Figure 2.1: Illustration of the main steps of LC-MS experiments. Figure created with BioRender.com.

To increase detection accuracy, mass analyzers are often coupled together in tandem mass spectrometry (MS/MS). In such a system, the peptide-ions from one mass analyzer are selected, fragmented and introduced to a second mass spectrometer. The second mass analyzer in turn separates and sorts the ion-fragments, before they are detected. The

additional fragmentation step increases the ability to distinguish ions that have very similar m/z-ratios in stand-alone mass spectrometers [1].

The number of samples under analysis for each experimental run of MS depends on whether the experiment is label-free or labeled. In a label-free setup, each sample corresponding to an individual or tissue is analyzed in a separate MS experiment. Several samples can be combined into one dataset, but they are not analyzed simultaneously. For labeled experiments on the other hand, it is possible to run multiple samples simultaneously in a batch-processed manner. The labeling is carried out by attaching isobaric mass tags to the peptides. The tags are molecules with equal mass, and each sample is supplied with its own chemically distinct tag. Thereby samples can be distinguished from each other, even when mixed together. With this technique, one experimental batch usually includes 4-11 samples, which greatly improves throughput of the experiments. However, batch-processing also results in considerable technical variations in the output data (i.e. batch-effects) that should be accounted for in the analysis (Figure 2.2) [9].

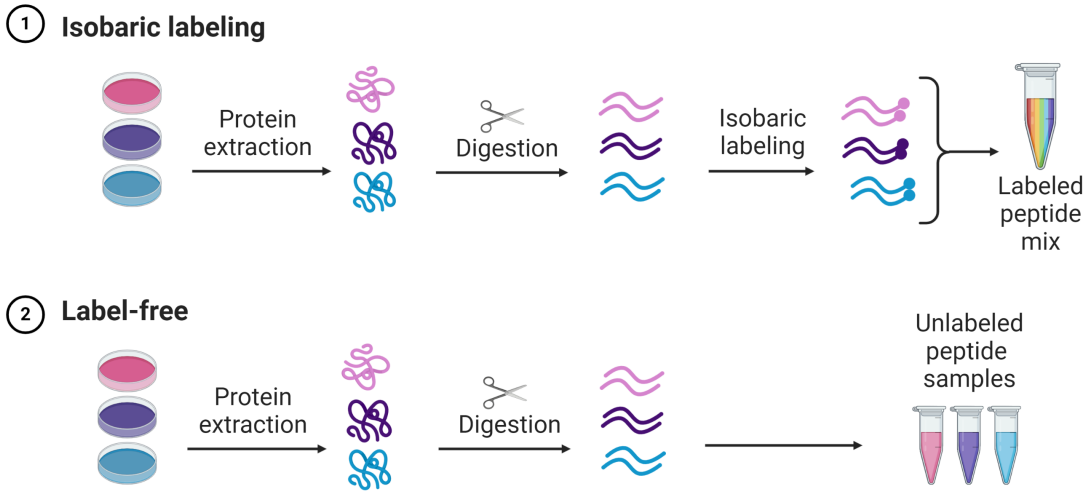


Figure 2.2: Illustration of sample preparation for isobaric labeled and label-free MS experiments. Figure created with BioRender.com.

Although the many advantages of mass spectrometry, MS-based data often suffers from a large proportion of missing values. The missing values arise from many events, including biochemical, analytical, and bioinformatical processes. For example, missing values can be generated due to suboptimal sample preparation, possibly occurring both in the protein extraction and in the LC-step. Other technical reasons for missing values are miscleavage of peptides, incomplete ionization, ion competition and ion suppression. Also misidentification of peptides and inaccurate detection of m/z-peaks cause missing values. In addition, missingness occurs in an abundance-dependent manner, where the mass spectrometer cannot detect abundances below certain thresholds. These missing values is often referred to as being below the level of quantitation [5, 10]. In labeled experiments there is also so-called batch-level abundance-dependent missingness, where the abundance of a protein is often either completely observed or missing for all samples in a batch [9]. Overall, missing values in MS data occur both in somewhat random patterns

in the overall data, but also in systematic patterns such as left-censoring of values below the limit of quantitation or batch-level abundance-dependent missingness.

In label-free LC-MS/MS experiments, the frequency of missing values generally ranges from 10% to 50% of the overall data, whereas the proportion of proteins that exhibit at least one missing value can be as high as 70-90% [10]. In labeled experiments, the proportion of missing data is usually low for a single batch of samples (~2%), but as data generated from multiple batches are combined the amount inflates. This is a consequence of the batch-level missingness and the fact that different proteins may be missing in different batches. For multi-batched experiments the amount of missing values is usually around 20% [11, 12].

The high amount of missing values and the many different events causing them severely complicates any downstream analysis of MS data. In general, substantial amounts of missing values in any dataset must be handled by statistical methods before valid and accurate analysis can be carried out [13]. The tackling of missing data in MS-based proteomics data is however not yet standardized and remains a challenge [5, 9, 10].

2.2 Concepts of missing data

Most statistical methods for data analysis rely on assumptions about the data and its representativeness of the overall population. When data are incomplete, as it often is for MS-based datasets, analysis results and interpretations may not be valid. In order to avoid this problem, it is crucial to understand the mechanisms that have generated missing data. Rubin (1976) formalized a framework for classifying mechanisms that lead to missingness, where the effects of different mechanisms on downstream inference are considered. The framework was refined by Little & Rubin (1987), establishing the foundations of analysis with missing data [13, 14]. In this section, concepts of this framework that are relevant for missing values in MS-based proteomics data are presented.

To establish some concepts of missing values, let $\mathbf{Y} = (y_{ij})$ denote a complete $n \times p$ data matrix without any missing values, where n denotes the number of samples and p denotes the number of proteins. Let $y_i = (y_{i1}, \ldots, y_{ip})$ denote the abundance of proteins $Y_1, \ldots, Y_p$ in the ith sample. Introducing missing data, let $\mathbf{M} = (m_{ij})$ be a matrix with the same dimension as $\mathbf{Y}$, such that $m_{ij} = 1$ if the abundance of the jth protein in the ith sample is missing, and $m_{ij} = 0$ otherwise. Thereby $\mathbf{M}$ is acting as a missingness indicator matrix for the abundance data $\mathbf{Y}$. Further, let $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$ denote the observed and missing elements of the dataset respectively. The *missingness mechanism* is defined as the conditional distribution of $\mathbf{M}$ given $\mathbf{Y}$, $f(\mathbf{M}|\mathbf{Y}, \phi)$, for some unknown parameters ϕ associated with the underlying mechanism of missingness. The distribution describes the relationships between the data and the probability of missingness; information that is important when analyzing any data set with missing values [13].

2.2.1 Types of missingness mechanisms

According to Rubin (1976), missingness mechanisms can be categorized into three different types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [14]. When the mechanism is MCAR, the probability of a value being missing does not depend on any values of the data matrix $\mathbf{Y}$, neither observed ($\mathbf{Y}_{\text{obs}}$) nor missing ($\mathbf{Y}_{\text{mis}}$). Any observation of any variable is as likely to be missing as any other, and the conditional distribution of $\mathbf{M}$ does not depend on $\mathbf{Y}$. That is,

$$f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\phi), \tag{2.1}$$

holds for all possible datasets $\mathbf{Y}$ and missingness parameters ϕ. In the case of MAR, the missingness mechanism does depend on values of the data, though only on the observed parts and not on the components that are missing. Formally,

$$f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\mathbf{Y}_{\text{obs}}, \phi), \tag{2.2}$$

holds for all missing values $\mathbf{Y}_{\text{mis}}$ and missingness parameters ϕ. When the missing-data mechanism is MNAR, the conditional distribution

$$f(\mathbf{M}|\mathbf{Y}, \phi), \tag{2.3}$$

does not simplify, as it depends on both observed and missing values of the data [13, 14, 15]. A summary of the three types of missingness is shown in Figure 2.3.

Understanding the type of missingness mechanism underlying the incompleteness of the data is important for handling the remaining data correctly. If the missing values are MCAR, the observed values are likely still representative of the population, seeing that the observed values are then essentially a random subsample of all values. In contrast, if the missingness is systematic, i.e. MAR or MNAR, inference and analysis may be biased [13]. MS data exhibit both random and systematic missingness patterns.

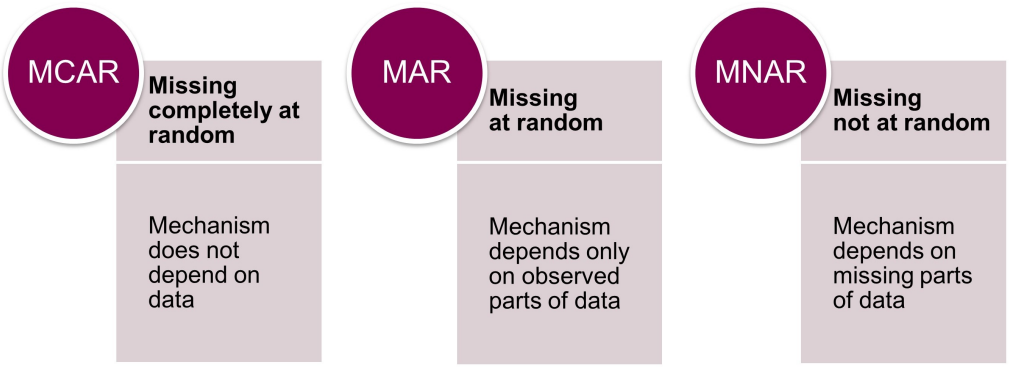


Figure 2.3: A summary of three types of missingness.

2.2.2 Ignorable missingness

One important distinction between MCAR, MAR and MNAR is the conditions under which the parameters of interest, θ, have valid and unbiased estimates. For data with

missing values, valid estimation of θ may or may not require knowing the parameters of the missingness mechanism ϕ. However, since the reason for missing values is rarely fully known, it is often impossible to estimate ϕ with any certainty. A context where ϕ is not needed for valid estimation of θ is therefore desirable [13, 16]. In this case, the missingness mechanism is *ignorable*.

The sufficient conditions of ignorable missingness mechanisms for likelihood inference (e.g. maximum likelihood estimation) are:

(i) the missing data are MAR; and
(ii) the parameters θ and ϕ are distinct, in the sense that the joint parameter space of (ϕ, θ) is the product of the parameter space of θ and ϕ.

The first condition implies that the missing data must be at least MAR, and thus the condition is fulfilled also when data is MCAR, i.e. given (2.1) holds, then by implication (2.2) holds as well. The second condition implies that knowing θ provides little information about ϕ and vice versa. Generally, the first condition is considered the most important, since if the data is MAR but distinctness does not hold, inference based on ignorable likelihood is still valid from the frequency perspective, but it is not fully efficient. For data that are fulfilling these conditions, valid maximum likelihood estimates for θ can be obtained based on a likelihood where the missingness mechanism is ignored $l(\theta|\mathbf{Y}_{\text{obs}})$, rather than on the full likelihood $l(\theta, \phi|\mathbf{Y}_{\text{obs}}, \mathbf{M})$. This considerably simplifies the inference (for derivation of these conditions, see Appendix A) [13]. Notably, ignorable missingness only implies that inference on θ without knowing ϕ is valid; it does not imply that that the missing data itself can be disregarded without introducing bias [13, 14].

2.2.3 Simulating missingness in MS data

Motivated by the fact that missing values in mass spectrometry-based data may be generated by different underlying mechanisms, we simulate data both by random and abundance-dependent mechanisms. Similar to previous sections, let $\mathbf{Y}$ be a $n \times p$ matrix of protein abundances and let the matrix $\mathbf{M}$ indicate which elements of $\mathbf{Y}$ are missing. It is reasonable to assume all samples of the mass spectrometry experiment are independent, implying independent rows $(y_i, m_i), i = 1, \ldots, n$. We may also assume that given the data $\mathbf{Y}$ and corresponding covariate data $\mathbf{C}$, the missingness of a protein is independent of the missingness, abundance and covariates of other proteins. The missingness mechanism can then be written as

$$P(\mathbf{M}|\mathbf{Y}, \mathbf{C}) = \prod_{ij} P(m_{ij} = 1|\mathbf{Y}, \mathbf{C}) = \prod_{ij} P(m_{ij} = 1|y_{ij}, \mathbf{c}_{ij}),$$

where $\mathbf{c}_{ij}$ is a vector of covariates associated with y_{ij}, if there are any. It has been proposed to model the missingness mechanisms $P(m_{ij} = 1|y_{ij}, \mathbf{c}_{ij})$ seen in label-free MS data by a bounded exponential function [17]:

$$g(y_{ij}; \mathbf{c}_{ij}, \phi_1, \phi_2, \phi_3) = K \min\{\exp(-\phi_1 - \phi_2 y_{ij} - \phi_3^T \mathbf{c}_{ij}), 1\}, \tag{2.4}$$

where the parameters ϕ_1, ϕ_2, ϕ_3 control the missingness mechanism and K is a constant to ensure reasonable probabilities. Specifically, when $\phi_2 = 0$, the probability function is

independent of the abundance of the missing element and missing values are MAR [13]. For MS-based data, it is generally assumed all MAR values are also MCAR, such that $\phi_3 = 0$ and there are no covariates $\mathbf{c}_{ij}$ considered that affect the missingness [8]. For $\phi_2 > 0$ the probability function monotonically decreases with the abundance value of y_{ij}, which is consistent with the abundance-dependent MNAR mechanism seen in mass spectrometry data [17]. In this case, the missingness is non-ignorable.

To model the batch-level abundance-dependent missingness present in labeled MS experiments (see Section 2.1), the mechanism should take into account that proteins tend to be completely observed or missing in all samples of a batch. In this case, the probability of missing values is dependent on the proteins' average abundance in the batch, and it may be appropriate to simulate missingness in a different manner [18]. In this thesis, the scope is limited to missingness mechanisms associated with label-free experiments modeled by (2.4).

2.3 Complete-case analysis

To demonstrate the effect that missing values may have on statistical analysis, consider a normally distributed variable Y with n observations. The variable could for example represent the abundance measurements of a single protein in n samples. Now assume some proportion of Y is missing and we wish to make inference about the mean. Let $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ denote observed and missing cases and let $n_{\text{obs}} < n$ denote the number of samples where Y is observed. In order proceed with inference, we must first decide on how to tackle the missing data.

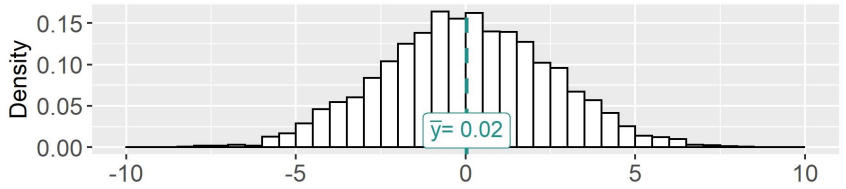
One method to proceed is to conduct the same analysis for the incomplete data as intended for the complete data, that is, to perform *complete-case analysis*. This would imply estimating the mean of Y by the sample mean of observed cases Y_{obs}, using only the reduced sample size n_{obs}. The standard error is then $s/\sqrt{n_{\text{obs}}}$, where s is the sample standard deviation of Y_{obs}.

If the missing data are MCAR, the set of observed values is a random subsample of all values. Consequently, complete-case analysis would indeed be valid and unbiased for data with MCAR missingness, though with reduced power as a consequence of the reduced sample size [13]. In Figure 2.4, the sample distribution of the the complete data (Figure 2.4a) is very similar to the distribution of the dataset with MCAR values (Figure 2.4b), even though 50% of data are actually missing in the second case. With complete-case analysis, the mean estimates are approximately identical.

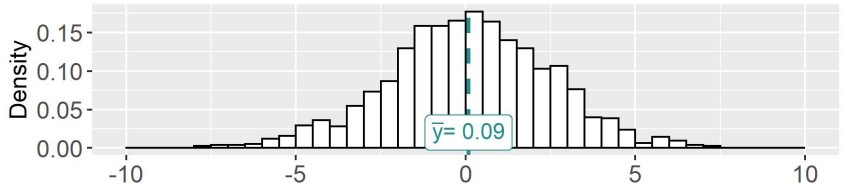
If the missing data are MNAR, the set of observed values is not generated from a random process and analysis based on the observed data will generally be biased [13]. Consider for example a left-censored dataset, where values are missing from the low-valued part (left tail) of the data. Complete-case analysis in this context results in upward biased estimates of the mean. However, the amount of biased largely depends on the censoring point (Figure 2.4). With a censoring point that is removing only a minor proportion of the data, the missing values have relatively small effects on the mean estimate bias (Figure

2.4c). In contrast, a strict censoring point means that the difference between observed and missing cases is large, leading to greater estimation bias (Figure 2.4d).

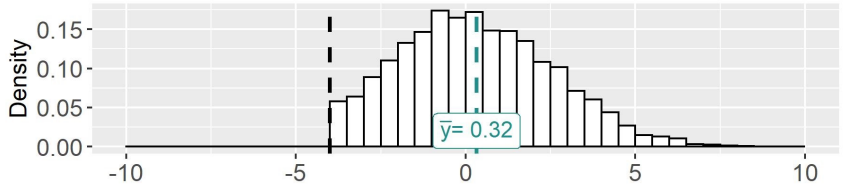
Complete-case analysis is fully valid only for MCAR data, whereas for MNAR or MAR data it results in biased estimates. The approach may however be justified provided that the bias is minimal. Factors that affect the degree of bias are the proportion of missing data, the missingness pattern, the extent to which complete and missing cases differ in value, and the parameters of interest. On that account, it is difficult to determine when complete-case analysis with systematic missingness is justified [13]. In MS-based proteomics data, all three types of missingness are usually present. Consequently, complete-case analysis of MS data is under most circumstances not recommended [5].



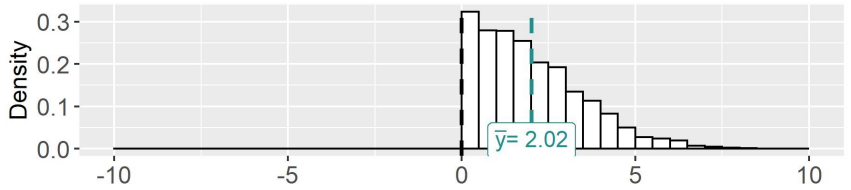
(a) No missing values.



(b) 50% MCAR values.



(c) Values below −4 censored (5% MNAR values).



(d) Values below 0 censored (50% MNAR values).

Figure 2.4: Demonstration of parameter estimation bias with different types of missing data. Dataset simulated from $\mathcal{N}(0, 2.5)$, $n = 3000$.

2.4 Imputation of missing values

A way to avoid the bias and reduction in statistical power associated with complete-case analysis of incomplete data is to impute the missing values [6]. *Imputation* is the

act of substituting missing values with values inferred from available information in the observed data. A synthetic complete dataset is created which can be analyzed using standard methods (Figure 2.5) [15].

There are a variety of imputation methods available that can be grouped into four categories: (i) single-value approaches; (ii) local-similarity approaches; (iii) ensemble approaches; and (iv) global-structure approaches. The choice of which category and method of imputation to use largely depends on the structure of the data but also the type of subsequent analysis.

Several of the available imputation methods have been suggested for dealing with the problem of missing values in MS-based proteomics data, but no formal framework has been established. A number of evaluations have been done in terms of imputation accuracy [5, 6, 7], but to our knowledge there has been no comprehensive comparison that focuses on the performance of imputation methods in terms of the ability to detect differentially abundant proteins (DAPs). We evaluated this aspect for five methods that generally have high imputation accuracy: k-Nearest Neighbors (kNN) imputation, Random Forest (RF) imputation, Multiple Imputation By Chained Equations (MICE), Penalized Expectation Maximum (PEM) imputation and Principal Component Analysis (PCA) imputation.

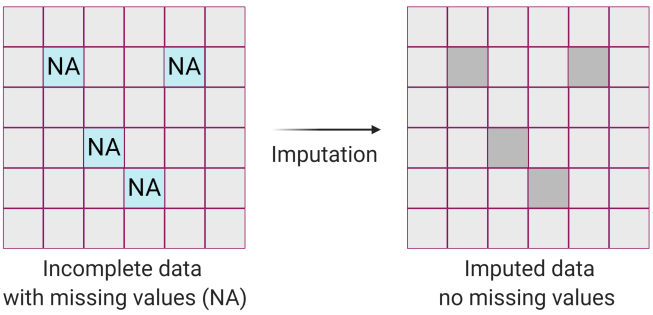


Figure 2.5: Illustration of imputation. Figure created with Biorender.com.

2.4.1 Single-value approaches

Single-value approaches for imputation replace missing values of a variable with a fixed value, such as the mean or median. Although simple in their methodology, these approaches will almost always underestimate the sample variance, yielding overestimated confidence levels in parameter estimation [15]. No single-value approach has been evaluated in this thesis, as it has been shown they exhibit poor imputation accuracy in MS contexts [5, 6].

2.4.2 Local-similarity approaches

Local-similarity approaches leverage information that is locally observed in clusters of the data in order to build the imputation model. In omics studies, these imputation methods are often applied to transposed data, i.e., a $p \times n$ matrix where proteins are in rows and samples in columns [6, 19]. By transposing the data, the local-similarity approaches

identify small clusters of proteins with similar abundance profiles; rather than clusters of similar samples. Under the assumption that such protein-clusters exist and are evenly distributed, the use of local structures may increase imputation accuracy [20, 21]. Due to the nature of MS data, we usually have datasets where $p \gg n$ holds; the number of proteins p are often many thousand, whereas the number of samples n are rarely more than a hundred. By transposing the data we also avoid problems associated with high-dimensionality for local-similarity approaches [22]. An example of a local-similarity approach is k-Nearest Neighbors imputation.

2.4.2.1 k-Nearest Neighbors (kNN) imputation

k-Nearest Neighbors (kNN) imputation is a method originally developed for microarray data. It leverages the correlation structure of the data by selecting a set of proteins with abundance profiles that are similar to that of the protein of interest, in order to impute its missing values. For each protein Y_j, $j \in (1, \ldots, p)$ with a missing value, the algorithm defines the neighborhood of Y_j as the k nearest proteins in Euclidean distance, confined to the samples for which the protein Y_j is not missing. Missing values of Y_j are then imputed by the average abundance of proteins within its neighborhood [19, 23]. For optimal performance k should be tuned based on a training set of data, however it has been suggested that the method is insensitive to the value of k within the range of 10-20 [19].

In a microarray setting, kNN imputation is very commonly used and the imputation accuracy has been well established. The major advantage of this method is the ability to successfully estimate missing values for genes (proteins) that are expressed in small clusters. The clusters do not themselves contribute significantly to the overall data structure, so other methods that rely on global parameters might not be as accurate in such cases [19]. However, this is very much dependent on the structure of the data and it is unclear if the same advantages apply to MS-based datasets.

2.4.3 Global-structure approaches

Global-structure approaches utilize the overall structure of the entire data to build the imputation model. These approaches are less flexible than the local-similarity ones, but can have the advantage of making more efficient use of the data and avoid over-fitting to observed cases [22]. For example, there are methods that perform dimension reduction in order to obtain a set of variables that capture a large amount of the variability in the overall data, to use for reconstructing missing values [6]. Penalized Expectation Maximum imputation and Principal Component Analysis imputation are examples of global-structure techniques.

2.4.3.1 Penalized Expectation Maximum (PEM) imputation

Penalized Expectation Maximum (PEM) imputation is an maximum penalized likelihood estimate (MPLE) approach to jointly estimate the mean abundances of proteins and the protein covariance matrix. It works by maximizing the joint likelihood of the observed data and the missingness pattern, assuming the data follows a multivariate normal distribution. It employs an Inverse-Wishart penalty on likelihood of the covariance matrix estimate

to ensure non-singularity. This form of penalty also allows for computational efficiency when calculating the MPLEs [17].

Expectation maximum algorithms have the advantage of producing unbiased estimates of parameters for both MCAR and MAR missing data, whereas for many other methods this may only be the case when missing values are MCAR [13]. The assumption of multivariate normally distributed data has the benefit of estimating the correlation matrix for the proteins. However, if those relationships are not linear, this could be a disadvantage. Thus, setting the penalty parameters in PEM can be challenging, but the authors suggest that the algorithm is robust to the choice of penalty parameters [24].

2.4.3.2 Principal Component Analysis (PCA) imputation

Principal Component Analysis (PCA) imputation works by the use of dimensionality reduction with PCA. The goal of PCA is to find an orthogonal coordinate system of principal components such that the variance of the data is maximized in each direction. By projection onto the first $q < p$ principal components, data can be reconstructed in a lower-dimensional space while still maintaining a certain amount of variance explained.

When performing PCA on incomplete datasets, there is no explicit solution to the maximization problem and an iterative approach must be adopted. The steps of the algorithm are: (1) initializing all missing values by e.g. mean values; (2) performing PCA on the completed data set to estimate the reconstruction formula with q principal components; (3) imputing missing values with a regularized version of the reconstruction formula. Steps (2) and (3) are repeated until the model convergences. The regularization step is used in order to avoid overfitting the model. The number of principal components, q, is selected by cross-validation [25].

2.4.4 Ensemble approaches

Similar to local-similarity approaches, the ensemble approaches utilize local structures in the data; but instead of basing imputation on one model, these methods produce and combine multiple models [6]. The combining of models can be done by bootstrap aggregation, as in Random Forest imputation, or by multiple imputation, as often done with Multivariate Imputation by Chained Equations.

2.4.4.1 Random Forest (RF) imputation

Random Forest (RF) imputation is based upon the ensemble machine learning algorithm with the same name. For each protein Y_j with a missing value, the algorithm builds a random forest with Y_j as the output variable and all other proteins as input. Samples where Y_j is observed are used as the training set, and missing values are imputed by predictions from the model. The random forest is built by creating an ensemble of decision trees using bootstrap sampling. The decision trees are de-correlated by a random sampling procedure and the final prediction model is the aggregation of all bootstrap trees. The RF algorithm predicts values with both low variance and low bias [22, 26]. An additional

advantage is that the imputation algorithm is non-parametric and makes no assumptions on the distributional aspects of the data [26].

2.4.4.2 Multivariate Imputation by Chained Equations (MICE)

Multivariate Imputation by Chained Equations (MICE) imputes missing values through an iterative series of regression models. The algorithm is initialized with a simple imputation method, e.g. mean value imputation. Then for one protein Y_j with missing elements at a time, Y_j is regressed upon all or some of the other proteins. Missing values of protein Y_j are imputed by predictions from the regression. The same procedure is repeated for all proteins with missing values. When a protein is acting as independent variable in any regression model, both observed and imputed values are used. In each iteration, the algorithm subsequently updates the imputed values, until a stopping criteria is reached [27]. In many applications, the stopping criteria is simply a fixed number of iterations.

The regression model used in MICE can be selected based on assumptions in the data, with examples being predictive mean matching, random forest or linear regression. In this thesis, we have used an algorithm that has been internally developed by Steven Novick, received through personal communication. The algorithm fits a Tobit linear regression model with left- and right-censoring in order to account for the values below the limit of quantitation seen in MS-data.

2.5 Datasets

2.5.1 Real quantitative MS dataset

By using a real dataset in the performance evaluation of the imputation methods, we ensure that the true complexity of MS-based data is considered. MS data is generally noisy and contains high levels of both biological and technical variations, as well as complex correlations from protein pathways and co-expression networks. For an imputation method to be used for MS data, is essential that it adequately considers these aspects.

For this reason, we conducted performance evaluations using the publicly available Cancer Cell Line Encyclopedia (CCLE) for quantitative protein expression profiling [28]. The CCLE dataset consists of protein abundances of samples corresponding to 375 cancer cell lines extracted from 22 lineages, with over 12,000 proteins identified over all samples. The experiments were batch-processed with 10-plex tandem mass tag (TMT), such that each batch included one reference channel and nine biological samples. The data has been normalized and batch-effects have been corrected for by the creators of the data [28]. For the analysis and evaluation of imputation methods, two cell lineages were selected to represent treatment groups: cell lines from skin- and lung tissue. These groups consisted of 33 and 77 samples respectively. In this original set, 12,743 proteins were identified with approximately 28% of values missing overall. To have a complete dataset for evaluation, all proteins with at least one missing value were removed, leaving 5270 proteins observed in all 110 samples.

2.5.2 Simulating MS data from a multivariate mixed-effects model

In addition to the evaluations with the real CCLE data, we conducted evaluations with simulated data. In order to simulate mass spectrometry-based data, we adopt a multivariate linear mixed-effects model [9]. In this case, the set of true DAPs is known from the model that generated the data. Further, the use of simulated data benefits from more control and knowledge about the data.

2.5.2.1 Background on multivariate linear mixed-effects models

The multivariate linear mixed-effects model (LMM) is a regression model that is particularly useful when the data has more than one source of random variability. The model structure is commonly used for clustered data, hierarchical data, repeated measurements or longitudinal data. The outcome variables in the LMM are described as a sum of unknown fixed effects, unknown random effects and noise. That is,

$$\mathbf{y}_k = \underbrace{\mathbf{X}_k\boldsymbol{\alpha}}_{\text{fixed effects}} + \underbrace{\mathbf{Z}_k\mathbf{b}}_{\text{random effects}} + \underbrace{\mathbf{e}_k}_{\text{noise}}$$

where $\mathbf{y}_k$ is a n_k-dimensional vector of outcomes for $k = 1, \ldots, N$ and N is the number of subjects. The matrices $\mathbf{X}_k$ and $\mathbf{Z}_k$ are a design matrices relating elements of $\mathbf{y}_k$ to the fixed effect $\boldsymbol{\alpha}$ and the random effects $\mathbf{b}_k$, respectively. The random effects, $\mathbf{b}_k$, are assumed to be independent and follow a multivariate normal distribution with mean vector zero and general covariance matrix $\mathbf{D}$. The noise (or error) term captures the variations not explained by the model. Usually it is assumed that all $\mathbf{e}_k$ are independent and follow a multivariate normal distribution with mean vector zero and covariance matrix $\boldsymbol{\Sigma}_k$ [29].

In many aspects, the main advantage of the LMM is the possibility to combine both fixed and random effects. Fixed effects often correspond to the effect of interest in a comparative study. The assumption for fixed effects is that the seen effect sizes are all estimates of a single true effect. Any variance between effect sizes is assumed attributable to noise only. For random effects, it is assumed that the seen effect sizes are estimates of their own true effect, which are distributed around an average. The variance is assumed attributable to both noise and actual across-subject variance. In other words, random effects are realizations of a random variable of which we model the distribution [29]. In MS data, the treatment group-associated differences can be represented by fixed effects, since the true difference is expected to be the same for all samples. For bach-processed experiments, we should include also batch-effects to our model. Taking into account that these effects are not generally expected to be the same for all batches, we model them as random effects.

Another advantage of the multivariate LMM is the use of multivariate outcome variables, corresponding to the p proteins under analysis. In contrast to a univariate model, a multivariate one allows for jointly modeling multiple proteins with some correlation structure taken into account. Hence, with the multivariate LMM co-expressed proteins (i.e. proteins within the same expression network or biological pathway) can be simulated [9].

2.5.2.2 Simulation model for MS data

Assume we are to simulate data from labeled (batch-processed) MS experiments, for some number of experimental batches each with some number of biological samples. Let $k = 1, 2, \ldots, N$ denote the experimental batches and let n_k denote the number of biological samples in the kth batch. For each batch k, we simulate a dataset of $\log_2$-transformed abundances with n_k samples in rows and p proteins in columns. Denote this $n_k \times p$ matrix by $\mathbf{Y}_k$ and denote its vectorization by $\mathbf{y}_k$, i.e.,

$$\mathbf{y}_k = \text{vec}(\mathbf{Y}_k) = [Y_{1k}^T, \ldots, Y_{pk}^T]^T,$$

where Y_{jk}, $j \in (1, \ldots, p)$ is the vector of abundances measured for the jth protein in the n_k samples of the kth batch. With this notation, the complete dataset can be simulated using a multivariate mixed-effects model for p-variate outcomes.

$$\mathbf{y}_k = \mathbf{X}_k\boldsymbol{\alpha} + \mathbf{Z}_k\mathbf{b}_k + \mathbf{e}_k, \tag{2.5}$$

where $\mathbf{X}_k$ is a known $n_k p \times q$ design matrix, $\boldsymbol{\alpha}$ is a vector of fixed effects, $\mathbf{Z}_k$ is a known $n_k p \times h$ covariate matrix and $\mathbf{b}_k$ is a vector of random effects for which $\mathbf{b}_k \sim \mathcal{N}_h(\mathbf{0}, \mathbf{D}_{h\times h})$ is assumed. The random error vector $\mathbf{e}_k$ is assumed to follow a multivariate normal distribution, $\mathbf{e}_k \sim \mathcal{N}_{n_k p}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{S_k})$, where $\otimes$ denotes the Kronecker product. The matrix $\mathbf{S}_k = \text{diag}(\sigma^2, \ldots)$ is an $n_k \times n_k$ matrix where the diagonal elements σ^2 are the sample variances, which are assumed equal for all samples. The $p \times p$ matrix $\boldsymbol{\Sigma}$ captures the unexplained covariance among the p response variables $Y_{1k}, \ldots, Y_{pk}$, that is, the covariance among proteins [9].

The model in (2.5) can be used to simulate protein-specific effects for the p outcome variables, by letting the $q \times 1$ vector $\boldsymbol{\alpha}$ contain a fixed intercept and the treatment group-associated effect for each of the p proteins, such that $q = 2p$. Then $\mathbf{X}_k$ will be a $n_k p \times 2p$ design matrix containing the treatment group indicators and the indicators of proteins. That is, $\mathbf{X}_k = \mathbf{I}_p \otimes \mathbf{X}_k^*$, where $\mathbf{I}_p$ is the identity matrix and $\mathbf{X}_k^*$ is the $n_k \times 2$ sample treatment group indicator matrix shared among the p proteins. Denote the fixed intercept and the treatment group associated fixed effect related to the jth protein by $\alpha_{0(j)}$ and $\alpha_{1(j)}$ respectively. With this notation, $\alpha_{0(j)}$ corresponds to the expected baseline abundance of the jth protein and $\alpha_{1(j)}$ is the expected log-fold change. Proteins simulated with non-zero $\alpha_{1(j)}$ are then differentially abundant.

The multivariate nature of the model in (2.5) moreover allows for simulation of co-expressed proteins by the use of a non-diagonal error covariance matrix $\boldsymbol{\Sigma}$. In this setup, a non-zero pairwise error covariance $\Sigma_{jj'}$ indicates correlation between the two proteins $j \neq j'$. Assuming all protein correlations are accounted for by the covariance of error terms, the random effects in $\mathbf{b}_k$ are independent among the p response variables. Thus $\mathbf{b}_k$ is one dimensional, containing only the batch effect of the kth batch, such that $b_k \sim \mathcal{N}(0, D)$ and $\mathbf{Z}_k$ is a vector of 1's.

Finally, simulated $\log_2$-abundances for each of N experiments (batches) can be combined into one overall $n \times p$ data matrix, where $n = \sum_{k=1}^{N} n_k$ is the total number of samples in all experiments. Denote this overall complete data matrix as $\mathbf{Y}_{n\times p} = (y_{ij})$, where

y_{ij} represents the $\log_2$-abundance of the jth protein in the ith sample of all experiments combined.

For simplicity, in the conducted simulations we have further assumed that samples are not batch-processed but rather generated from a label-free experimental setup, thereby the random effects term can be removed from the model ($b_k = 0$). This simulation setup could also represent a scenario where all batch effects are properly corrected for in the data pre-processing. Simulating samples that are not batch-processed implies $n_k = 1$ for all $k = 1, \ldots, N$, so that $n = N$. Taking into account the above assumptions and simplifications, we can rewrite the resulting simulation model for label-free MS experiments with p proteins in n samples as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{e}, \tag{2.6}$$

where $\mathbf{y}$ is the vector of abundances and $\mathbf{X}_k$ is a known design matrix for the fixed treatment effects $\boldsymbol{\alpha}$. The noise is assumed to follow a normal distribution $\mathbf{e} \sim \mathcal{N}_{np}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{S})$, with the diagonal $n \times n$ matrix $\mathbf{S} = \text{diag}(\sigma^2, \ldots)$ of sample variances and the non-diagonal $p \times p$ matrix $\boldsymbol{\Sigma}$ of protein covariances.

2.6 Statistical analysis

2.6.1 Welch's t-test for differential abundance

Welch's t-test is a two-sample location test for equal sample means. It assumes that both populations are normally distributed, while allowing unequal sample variances [30]. In the context of MS data, the test is applied for each protein individually to identify statistically significant group-differences in abundance means, that is, to identify DAPs. The null and alternative hypotheses, denoted H_0 and H_A respectively, for each protein $j = 1, \ldots, p$ are:

$$H_0 : \mu_{(0)j} = \mu_{(1)j},$$
$$H_A : \mu_{(0)j} \neq \mu_{(1)j},$$

where $\mu_{(0)j}$ and $\mu_{(1)j}$ denotes the mean $\log_2$ abundance of protein j in the control and treatment population respectively. That is, the null hypothesis of the jth test is that the jth protein is a DAP, with the alternative hypothesis that it is not. The test statistic for the jth protein is defined as

$$t = \frac{\bar{Y}_{(0)j} - \bar{Y}_{(1)j}}{\sqrt{s^2_{(0)j}/n_{(0)j} + s^2_{(1)j}/n_{(1)j}}},$$

where $\bar{Y}_{(0)j}, \bar{Y}_{(1)j}$ are the sample mean $\log_2$ abundances, $s_{(0)j}, s_{(0)j}$ are the sample standard deviations and $n_{(0)j}, n_{(0)j}$ are the sample size, for control and treatment group, respectively. The test statistic approximately follows a t-distribution under the null hypothesis, under the given assumptions. The degrees of freedom associated with the variance estimate is approximated by

$$\nu = \frac{\left(\frac{s^2_{(0)j}}{n_{(0)j}} + \frac{s^2_{(1)j}}{n_{(1)j}}\right)^2}{\frac{\left(s^2_{(0)j}/n_{(0)j}\right)^2}{n_{(0)j}-1} + \frac{\left(s^2_{(1)j}/n_{(1)j}\right)^2}{n_{(1)j}-1}}$$

[30]. Associated with each performed statistical test is a P-value, which indicates the probability of obtaining a test statistic at least as extreme as the one observed. In other words, the P-value indicates the amount of evidence there is against the null hypothesis. If the P-value is below or equal to the significance level α, there is sufficient evidence to reject the null hypothesis and the test is considered significant. The significance level corresponds to the probability of rejecting the null hypothesis even though it is true. The probability that the null hypothesis is rejected when it is false is called the statistical power of the test [31].

The rejection rules of a statistical can test can also be determined from the confidence interval (CI) associated with the test. The two sided (1-α)100% CI for Welch's t-test is derived from

$$\bar{Y}_{(0)j} - \bar{Y}_{(1)j} \pm t_{\alpha/2,\nu}\sqrt{\frac{s^2_{(0)j}}{n_{(0)j}} + \frac{s^2_{(1)j}}{n_{(1)j}}},$$

where $t_{\alpha/2,\nu}$ is the critical t-value for significance level α and ν degrees of freedom. The CI gives a range of values within which we are reasonably confident that the population parameter lies. It is a random interval that in (1-α)100% of cases will contain the true estimate of the effect size. CIs together with point estimates give information about the uncertainty of estimates. Rejecting the null hypothesis based on a P-value $\leq \alpha$ is equivalent to that the CI does not contain the null hypothesis value [31].

The motivation for using a simple Welch's t-test, rather than more sophisticated methods such as the R package Limma [32], is the direct interpretability of the result. Limma is an Empirical Bayes procedure, with the effective functionality of shrinking protein variances used to compute the test statistic. However, missing or imputed values present in the data is also a possible source for alterations to the estimated standard error. Therefore, in order to distinguish the actual consequences of having missing or imputed data, we employ Welch's t-test instead.

2.6.2 Correcting for multiple testing

For a single hypothesis test, the significance level α controls the probability of the test to generate a false positive (FP). However, as one test per protein will be performed simultaneously, there will on average be $\alpha \times p$ FPs among all tests, where p is the number proteins. Generally, the number of proteins is large, hence the problem of multiple testing must be corrected for in order to avoid inflation in the number of FPs. There are a number of methods available for correcting the effect of multiple testing, where we have chosen the Benjamini-Hochberg procedure for false discovery rate (FDR) correction. With this method, the P-values for each test are adjusted (P_{adj}) in order to control the FDR. This method is generally considered less strict compared to competing methods (e.g. the Bonferroni procedure) [22, 31]. Details on the Benjamini-Hochberg procedure are found in Appendix A.

2.7 Measures of performance

Each hypothesis test performed for a protein can be regarded as a binary classification problem, by classifying the decision made based on the test as either positive or negative. The decision is considered positive if H_0 is rejected and negative if we fail to reject H_0. Further, a ground truth can be established: cases where H_0 is in fact false are considered real positives; and cases where H_0 is in fact true are considered real negatives. Hence, there are four possible combinations of test-based decisions and ground truths (Figure 2.6), namely: true positive (TP); false positive (FP); true negative (TN); and false negative (FN) [22].

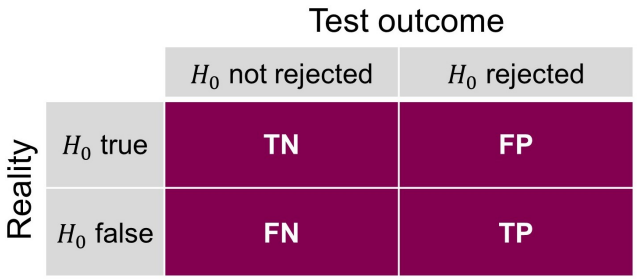


Figure 2.6: Possible outcomes of a statistical test; true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

When Welch's t-test has been performed for each protein of a dataset, the outcome of tests are summarized in a confusion matrix, containing the number of TPs, FPs, TNs and FNs. Given the ground truth is known, the performance in terms of the ability to detect significantly abundant proteins can be derived. Different metrics can be used to this end, where we have considered the true positive rate (TPR) and the false positive rate (FPR). The TPR is the proportion of real positives (P) that were correctly identified as positives. The TPR is the proportion of real negatives (N) that were falsely identified as positives. Formally,

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}},$$

[22]. Framing the performance metrics in the MS context, the TPR is the proportion of true DAPs that were indeed identified, and the FPR is the proportion of non-DAPs that were falsely identified. This way, the TPR and FPR indicate the ability to detect DAPs, given the data at hand. Finally, the TPR and FPR derived from datasets where missing values have been imputed by different methods can be compared in order to evaluate the performance of the imputation methods. The ground truth indicates whether or not the protein under testing is *truly* a DAP; a fact that cannot actually be known about a real MS dataset. For this reason, we defined the ground truth as the DAPs detected in the complete data. When using simulated data, the ground truth set of DAPs is given by the simulation model.

2.7.1 Clinical relevance

Statistical significance does not necessarily imply practical importance. In drug development, a successful study demonstrates not only a significant result but also a clinically relevant effect size [33]. For this reason, we apply a second condition in the binary classification of positive or negative hypothesis tests, in addition to the condition on statistical significance. The second condition is; for a test-outcome to be considered positive, the estimated $\log_2$ fold-change must be clinically relevant. That is, $|\bar{Y}_{(0)j} - \bar{Y}_{(1)j}| > R$ for some number R determined by target effect size. Statistical significance and clinical relevance can be interpreted in terms of confidence intervals (Figure 2.7).

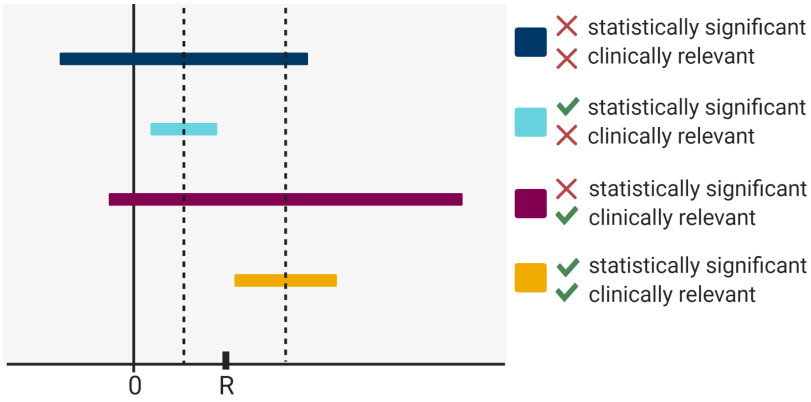


Figure 2.7: Representation of statistical significance and clinical relevance in terms of CIs for a null hypothesis of zero fold-change given a level of confidence. The CI of a statistically significant test does not cover 0. The CI of a clinically relevant test has its center above the determined clinical relevance threshold R. Figure created with BioRender.com.

2.7.2 Level of uncertainty

When evaluating the performance of different strategies for dealing with missing data in MS experiments, there is interest in determining whether statistically significant and clinically relevant DAPs can be detected, as well as in determining with what amount of certainty the detection can be made. This information can be obtained from the corresponding CIs [31].

The CI of a test indicates the range within which we are $(1-\alpha)100\%$ confident that the true estimate lies. The width of the CI conveys the associated uncertainty, and is defined by the critical value of the test statistic $t_{\alpha/2,\nu}$, the sample standard deviations $s_{(0)j}$, $s_{(1)j}$ and the sample sizes $n_{(0)j}$, $n_{(1)j}$. With complete-case analysis of incomplete data, the sample sizes will decrease with the amount of missing values, generally leaving wider confidence intervals. For imputed datasets, the sample sizes will correspond to those of the complete data, but depending on the predictions made by the imputation method the sample standard deviations may be skewed. This could in turn affect the width of confidence intervals. In some cases, the increase in width will leave non-significant test. It is however possible that two tests on different samples from the same population differ greatly in certainty, even though they are both statistically significant (Figure 2.8).

When testing for DAPs in imputed or incomplete data we expect an increase in the

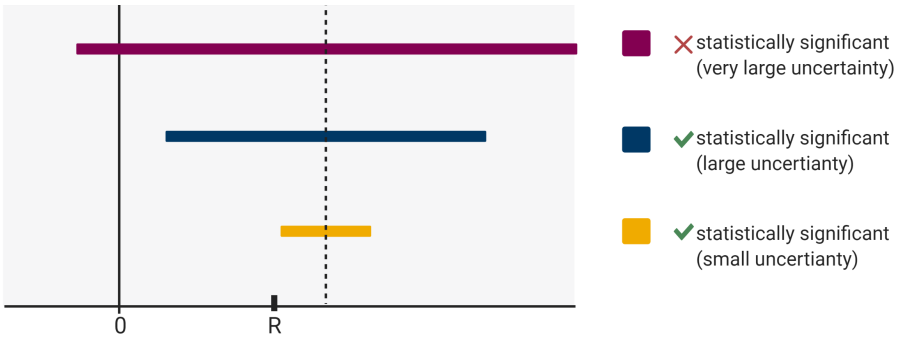


Figure 2.8: Illustration of CIs, given the same $(1-\alpha)100\%$ confidence level and est. fold-changes. The widths of the CIs convey the uncertainty associated with the tests. For the first CI (burgundy), the large amount of uncertainty results in a non-significant test. Both the next two CIs (blue and yellow) are significant, but the amount of uncertainty differs substantially. Here, performance evaluations that also account for uncertainty may be of value. Figure created with BioRender.com.

width of CIs to some extent, compared to when using the complete data. However, in order to establish a best practice for imputation, we want to evaluate how this increase in uncertainty differs; both between different imputation methods and in relation to complete-case analysis of incomplete data. Considering that a wider CI implies a larger P-value (given a hypothesis test) [31], this comparison may be made on the P-values rather than on the CIs.

We implement a third and a final condition for positive test-outcomes. A hypothesis test that is both significant and relevant is considered positive only if the obtained P-value is at most Q times larger than the P-value obtained with complete data. That is, given a test performed using the complete data and the corresponding test for an alternative (incomplete or imputed) data, it must hold that:

$$\frac{P\text{-value} \mid \text{alternative data}}{P\text{-value} \mid \text{complete data}} < Q,$$

where Q defines some threshold. Note that it is assumed that the observed elements in the alternative data are identical to corresponding elements in the complete data. The value of the threshold Q can be set differently depending on the comparisons that are to be made. Given a complete and an alternative dataset, a small Q implies that only those tests that generated a small increase in uncertainty will be considered positives for the alternative data. Using a larger Q implies that larger increases in uncertainty are accepted. What ranges of Q that are regarded as small or large highly depends on the data, and on the P-values obtained for tests in complete data.

3

Methods

Simulations and analyses were in R 3.6.0 on a high-performance computing cluster with Slurm Workload Manager (v.19.05.7). The packages used for imputation are given in Table 3.1, as a complement to the employed Conda (v.4.5.12) environment.

3.1 Generating real complete data

For analysis with real MS data, the normalized proteomics data from the Cancer Cell Line Encyclopedia (CCLE) was used. As described in Section 2.5.1, cell lines from skin- and lung tissue were selected to represent treatment groups. All proteins with a missing value were removed, leaving 5270 proteins observed in all 110 samples. In order to ensure reasonable computation time, a subset of 1000 proteins was randomly selected in each simulation to form the complete dataset.

3.2 Simulating complete data

In each simulation, a complete dataset was simulated from the model in (2.6) with $p = 1000$ proteins. The number of samples n was set either to 10, 50 or 100, always with equal number of samples in treatment and control groups. The mean abundance of each protein was set to $\alpha_0 = 20$ for datasets where MCAR values were to be introduced. For datasets where abundance-dependent MNAR missingness were to be introduced, the mean abundance of each protein was generated from $\text{Unif}(17, 23)$, in order to increase the variation of abundances in the data.

The sample variance, σ^2, of every simulation was fixed to 4. The protein covariance matrix, $\mathbf{\Sigma}$, was fixed and constructed by making one random selection of 1000 proteins from the CCLE dataset and estimating the covariance matrix for those. The random selection was made from complete cases of the lung and skin tissue samples, in order to mimic the analysis performed with real data.

In each simulation, 10% of all proteins were selected at random to be simulated as differentially abundant, where the expected fold change was set large enough to allow for statistical power to detect the effect in complete data, given the number of samples. The values were set as $\alpha_1 = 3.5$ for $n = 10$, $\alpha_1 = 2.5$ for $n = 50$, and $\alpha_1 = 2$ for $n = 100$. The remaining 90% of proteins were simulated with zero expected fold-change.

3.3 Analysis

Next, missing values were introduced with either MCAR or MNAR mechanisms. To simulate MCAR values, elements from the complete overall data matrix were selected at random to be removed until a given proportion of the entire data was missing. To introduce MNAR values the mechanism (2.4) described in Section 2.2.1 was used with $K = 150$, $\phi_1 = \phi_3 = 0$ and $\phi_2 = 0.3$. The parameter values were selected based on the produced probability distribution of missing values, ensuring a contrast between lower and higher abundances. Using the probabilities of missingness derived for each element in the data, a random sample of elements to set as missing was drawn so that a given proportion of the data were removed. For each of the mechanisms, missing values were introduced at 10%, 30% and 50% of the overall data. Finally, the missing values of each generated incomplete dataset were imputed, using each of the evaluated imputation methods (kNN, RF, MICE, PEMM and PCA) using the packages specified in Table 3.1. Most parameters of the imputation algorithms were implemented with default values, with some exceptions given in the table. The evaluated imputation methods are all available as open-source packages in R, but for imputation with MICE an internal and not yet published version developed by Steven Novick, was used.

Table 3.1: Reviewed imputation methods, package used for implementation in R and non-default parameters. For implementation of MICE, an internal (closed-source) version developed by Steven Novick was used.

Method	R package	Non-default R parameters
kNN [19]	*impute* v.1.60.0 [34]	`colmax=1`, `rowmax=1`
RF [26]	*missForest* v.1.4 [35]	`parallelize="forests"` used w. 8 cores
MICE [27]	internal	
PEM [17]	*PEMM* v1.0 [24]	`phi=0`
PCA [25]	*missMDA* v.1.17 [36]	`ncp` tuned by k-fold CV [37]

3.4 Evaluation of performance

For each generated complete, incomplete and imputed dataset, DAPs were tested for using Welch's t-test, where the estimated $\log_2$ fold-change and corresponding FDR-adjusted P-value were derived for each protein in the data. A positive outcome was counted when three conditions were met: (1) adjusted P-value (P_{adj}) below 0.05; (2) the estimated $\log_2$ fold-change was greater than R; and (3) the obtained P_{adj} was at most Q times larger than the P_{adj} obtained for complete data. The threshold for clinical relevance was set to $R = 1$ since in MS-based analyses we are rarely interested in smaller target effect sizes. The value of Q was set differently depending on the data: using real data $Q = 10^3$; and using simulated data $Q = 10^{18}$. The difference in value is motivated by the differences in complexity and noise between real and simulated data.

Comparing the test-outcomes to the ground truth, the TPR and FPR associated with each dataset were derived. With real data, the true set of expected log fold-changes is not known. Therefore, the DAPs detected for the complete datasets were used as ground

truth in order to derive performance metrics for incomplete and imputed datasets. For simulated data, the real DAPs is determined by the simulation model; proteins simulated with a non-zero fold-change correspond to the ground truth. In this case, performance metrics for the complete data can be derived as well.

For every dataset and every combination of parameters, 1000 iterations were performed. Using the real data, this resulted in 1,000 complete, 6,000 incomplete datasets (for three missingness proportions per complete dataset) and 30,000 imputed datasets (for five imputation methods per incomplete dataset) (Figure 3.1a). Using simulated data, this generated 3,000 complete datasets (for three different samples sizes), 18,000 incomplete datasets and 90,000 imputed datasets (Figure 3.1b).

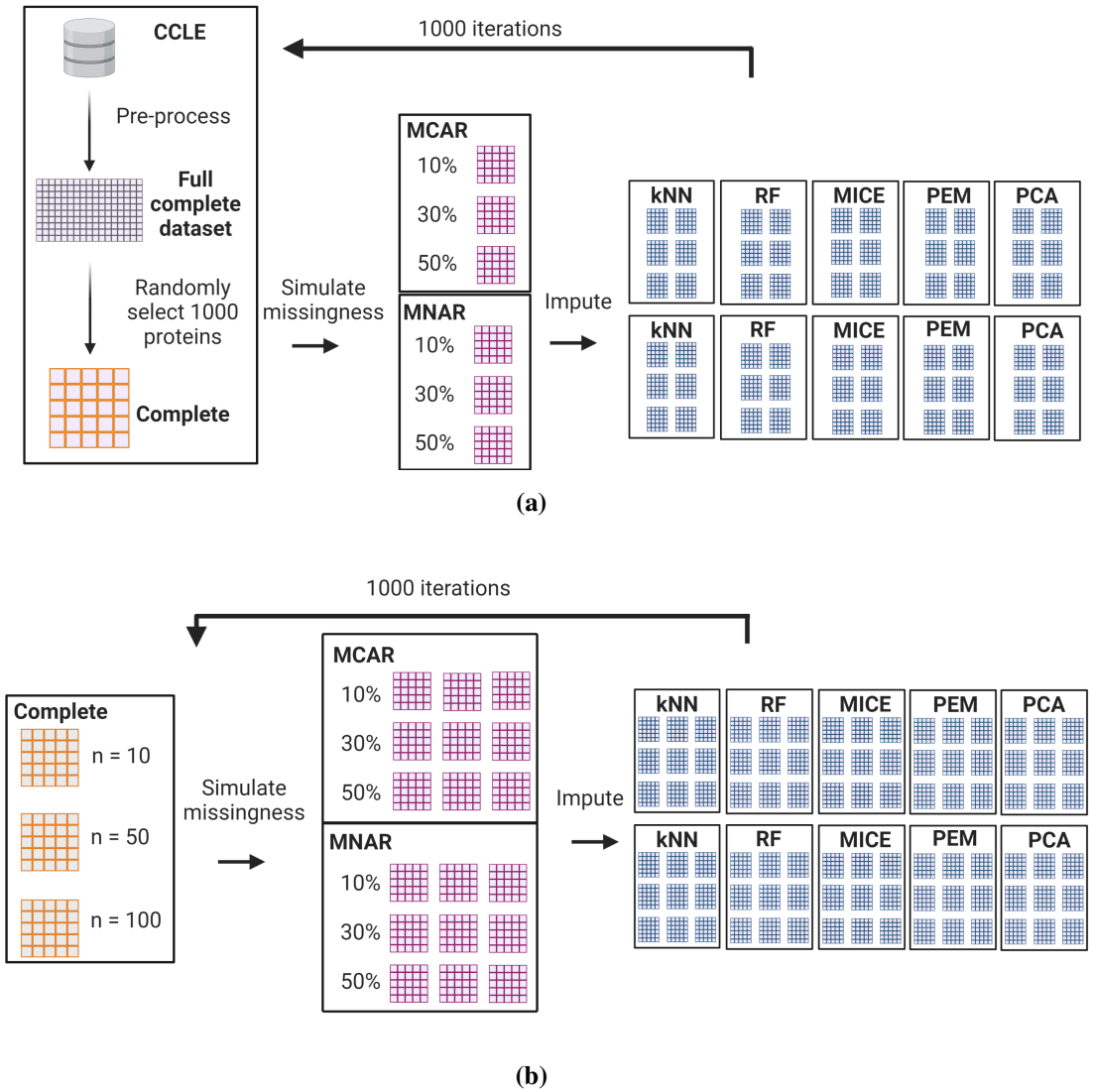


Figure 3.1: Scheme of data generation and associated parameters, starting with (a) real data (b) simulated data. Figures created with BioRender.com.

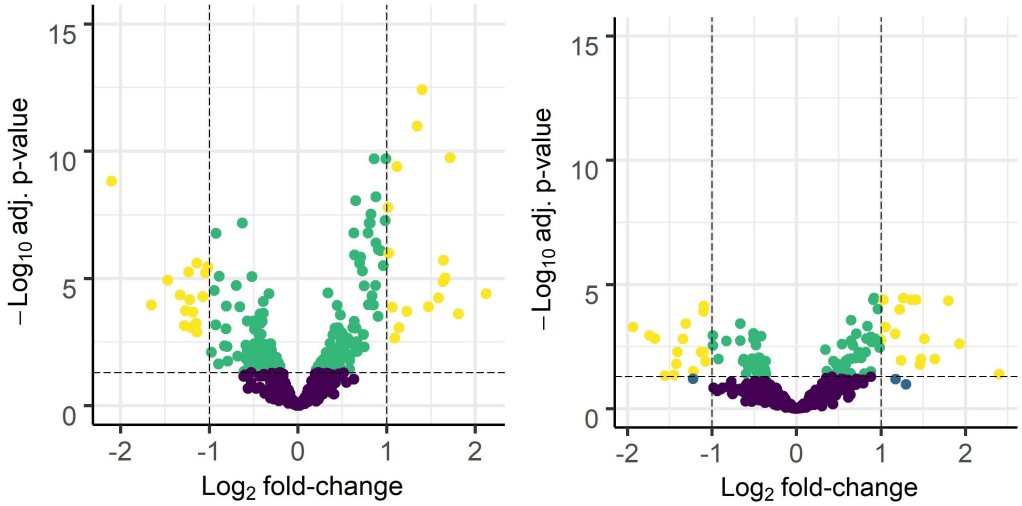
4

Results

The conducted simulations and analyses were aimed to answer the question on whether missing values affect the ability to detect differentially abundant proteins (DAPs) in MS-based data. A further aim was to evaluate the performance of commonly used imputation methods (kNN, RF, MICE, PEM and PCA) in terms of the ability to detect DAPs. Performance was measured in TPR and FPR from univariate analyses with Welch's t-test, where MCAR and MNAR values had been introduced into data from the Cancer Cell Line Encyclopedia, and data simulated from a multivariate-mixed effects model respectively. A test was classified as positive if it was both statistically significant and clinically relevant, while also yielding a uncertainty level similar to that obtained with complete data ($P_{\text{adj}} \leq 0.05$, est. $\log_2$ fold-change >1, and P-value ratio criteria Q). For detailed description of methods see Section 3. Based on the results, we can establish suggestions on how to deal with missing values, considering proportion of missing values, type of missingness and sample size.

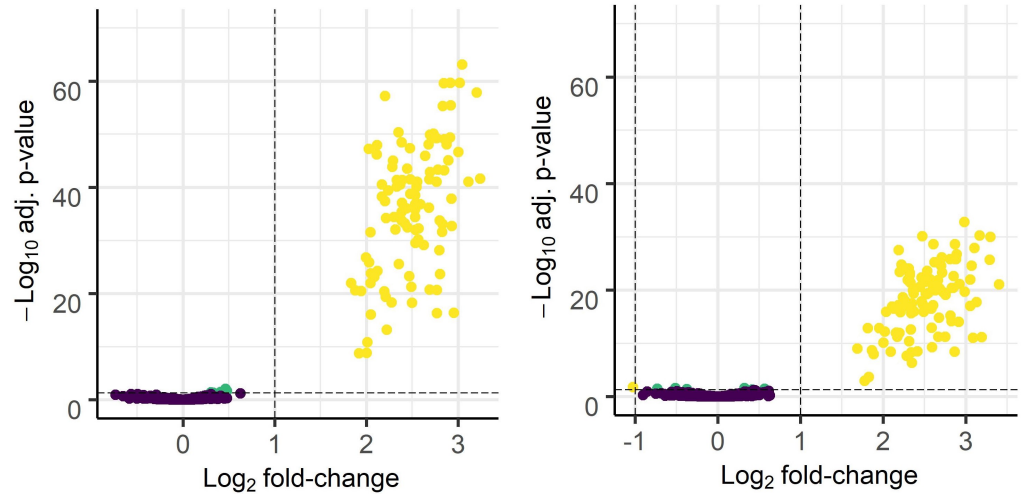
4.1 Establishment of performance criteria

The performance criterion Q was used to restrict the allowed increase in uncertainty, comparing an alternative data (incomplete or imputed) to complete data, in order to count tests as positives (see Section 2.7.2). By using this criterion, not only did the evaluations determine if DAPs could be detected, but also considered the level of certainty with which they were detected. The threshold was determined by evaluating the distributions of P_{adj}-values over estimated $\log_2$ fold-change. In Figure 4.1 this distribution is shown for one instance of the real dataset; either as complete data or incomplete data with 50% MCAR values. When no data were missing, the spread of P_{adj} values was down to approximately $10^{-12.5}$. For 50% data missing, the smallest P_{adj} was around 10^{-5}. Based on this difference in P_{adj} distributions, the value $Q = 10^3$ was selected as the threshold for how many times larger a P_{adj}-value for alternative data may be compared to the P_{adj}-value for the complete data, in order to count the test as positive. Similarly, in Figure 4.2 the distribution of P_{adj}-values over estimated $\log_2$ fold-change for one iteration of simulated data is shown. In this case, the smallest P_{adj} was around 10^{-60} for the complete dataset and 10^{-30} for the incomplete dataset with 50% MCAR. Based on the difference, we selected $Q = 10^{18}$ when using simulated data. Supplementary figures for 10% MCAR values and for PCA-imputed data can be found in Appendix B.



(a) Complete data **(b)** 50% MCAR

Figure 4.1: FDR-adjusted P-value on negative $\log_{10}$ scale against estimated $\log_2$ fold-change for one iteration of analysis with real data ($p = 1000$). Points colored according to: (yellow) significant and relevant; (green) significant but non-relevant; (purple) non-significant and non-relevant; and (blue) relevant but non-significant. Significance level $\alpha = 0.05$ and clinical relevance level $\log_2$ fold-change > 1.



(a) Complete data **(b)** 50% MCAR

Figure 4.2: FDR-adjusted P-value on negative $\log_{10}$ scale against estimated $\log_2$ fold-change for one iteration of analysis with real data ($p = 1000$, $n = 100$). Points colored according to: (yellow) significant and relevant; (green) significant but non-relevant; (purple) non-significant and non-relevant; and (blue) relevant but non-significant. Significance level $\alpha = 0.05$ and clinical relevance level $\log_2$ fold-change > 1.

4.2 False positive rates

For each performed analysis with complete, incomplete or imputed data, the FPR was calculated. Figures 4.3 and 4.4 show the FPR distributions with MCAR values, for real and simulated data respectively. Corresponding figures for MNAR values are found in Appendix C. As seen in these figures, the FPR was below 0.05 for all methods under all conditions tested. Since 0.05 is the expected rate of false positives given the FDR adjustment and the significance level of 5%, the results indicate that the FPR was controlled below the significance level. This means that no artificial DAPs was introduced by any imputation methods under any of the tested conditions.

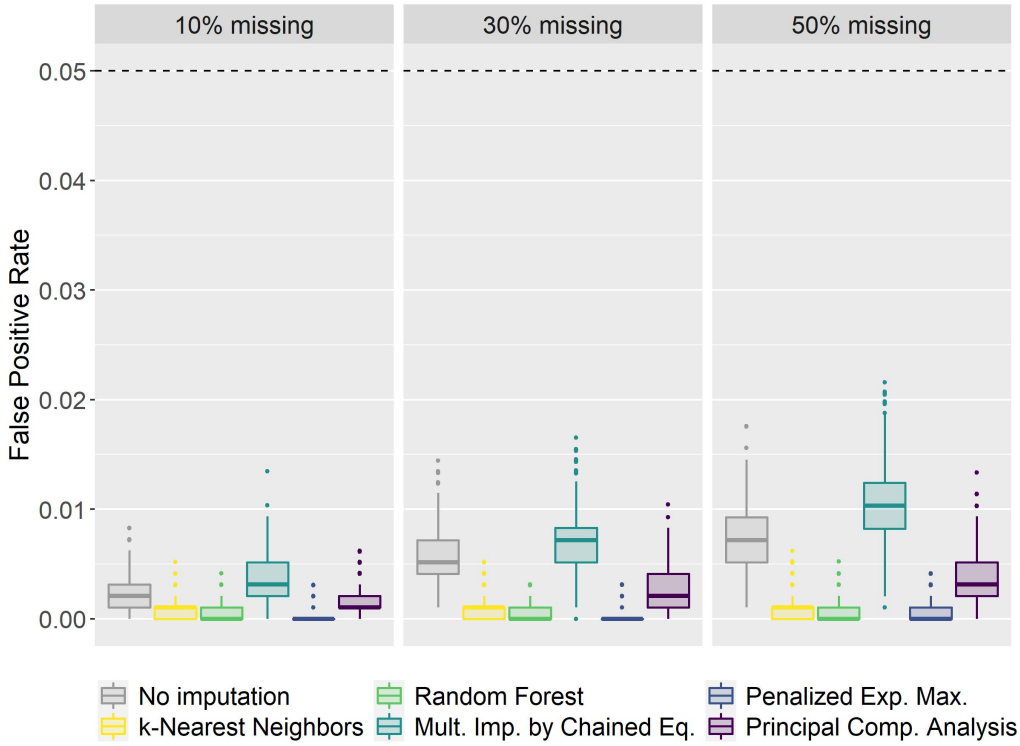


Figure 4.3: False positive rate for real data with three proportions of MCAR missingness and different strategies of tackling missing values. A test was considered positive if three conditions were met: (1) adj. *P*-value ≤0.05; (2) est. $\log_2$ fold-change greater than 1; and (3) adj. *P*-value not more than 10^3 times larger than adj. *P*-value for complete data. Test-outcomes were compared to the ground truth of DAPs detected when using the complete data.

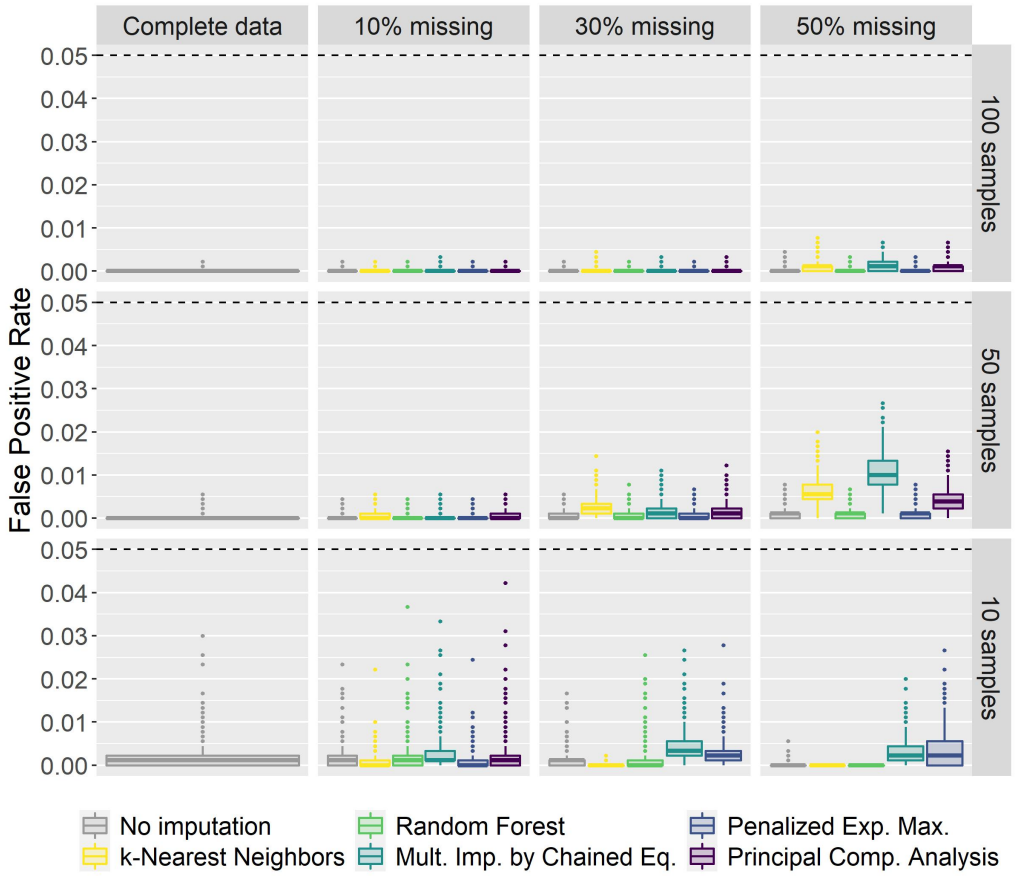


Figure 4.4: False positive rate for simulated data with three sample sizes, three proportions of MCAR missingness and different strategies of tackling missing values. A test was considered positive if three conditions were met: (1) adj. *P*-value ≤0.05; (2) est. $\log_2$ fold-change greater than 1; and (3) adj. *P*-value not more than 10^{18} times larger than adj. *P*-value for complete data. Test-outcomes were compared to simulated ground truth DAPs.

4.3 True positive rates

For each performed analysis, the proportion of true DAPs that were detected (TPR) was calculated. The obtained TPR distributions for datasets with MCAR values are shown in Figures 4.5 and 4.6, for real and simulated data respectively. The corresponding results for MNAR values were close to identical to the MCAR case (Appendix D).

Figure 4.5 depicts a trend of decreased TPR with increased proportion of missing values. When 10% of the data were missing, all methods had a median TPR above 0.75, suggesting that imputation does not increase the TPR compared to removing missing values. For 30% missingness irrespective of data or missingness type, the median TPR was less than 0.8. In this case, there was a slight increase in median TPR from no imputation (0.73) to imputation by PCA or MICE (0.8 and 0.77 respectively). For 50% missingness the TPR was 0.25-0.75 depending on imputation method. Here, the median TPR obtained with PCA and MICE (0.64 and 0.69 respectively) is evidently higher than that obtained

without imputation (0.33). Accordingly, we would not recommend imputation if only a small amount of data is missing (≤30%). For large proportions of missingness (>30%) we recommend imputation with PCA or possibly MICE.

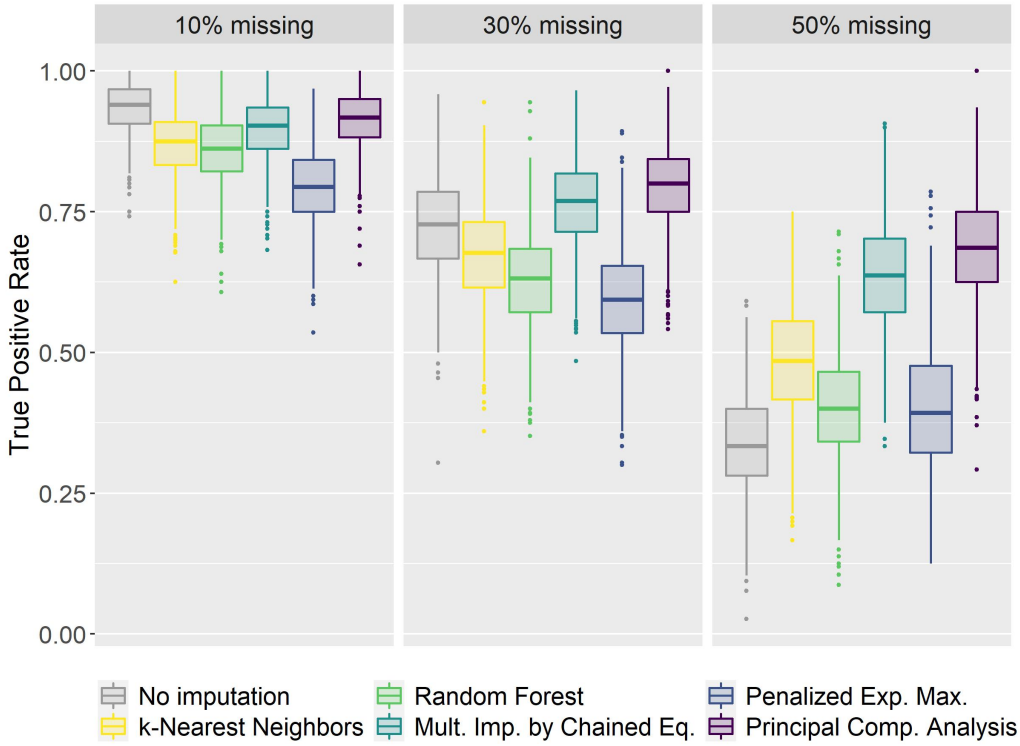


Figure 4.5: True positive rate for real data with three proportions of MCAR missingness and different strategies of tackling missing values. A test was considered positive if three conditions were met: (1) adj. P-value ≤0.05; (2) est. $\log_2$ fold-change greater than 1; and (3) adj. P-value not more than 10^3 times larger than adj. P-value for complete data. Test-outcomes were compared to the ground truth of DAPs detected when using the complete data.

For simulated data, we performed the evaluations using different sample sizes (n = 10, 50, 100) and varying proportions of missingness (MCAR in Figure 4.6). For complete data, the TPR was centered at 1 with close to no dispersion. This was also the case with 10-30% of data missing in large sample sizes (n=50, 100), irrespective of method.

With 50% missingness and n = 100, the TPR for non-imputed data was notably low (median at 0.48), but when imputation was performed the TPR remained close to 1. Therein, we would recommend imputation by any of the evaluated methods for large n and high percentage of missing data (considering only the results from simulated data).

The results with n = 10 indicated poor performance for all methods at 50% missing values. Guided by these results, statistical inference without imputing missing values is recommended for small n and large proportions of missingness. Notably, PCA did not give any results at all at 30% or 50% missing values, hence these cases are not seen in Figure 4.6. PCA-imputation is initialized with mean imputation, and since the PCA transform

cannot deal with zero column-variance, the algorithm fails unless at least two values are observed per column. For small sample sizes, it is probable that at at least one protein will have fewer observed cases than that; hence the algorithm is likely to fail in this case.

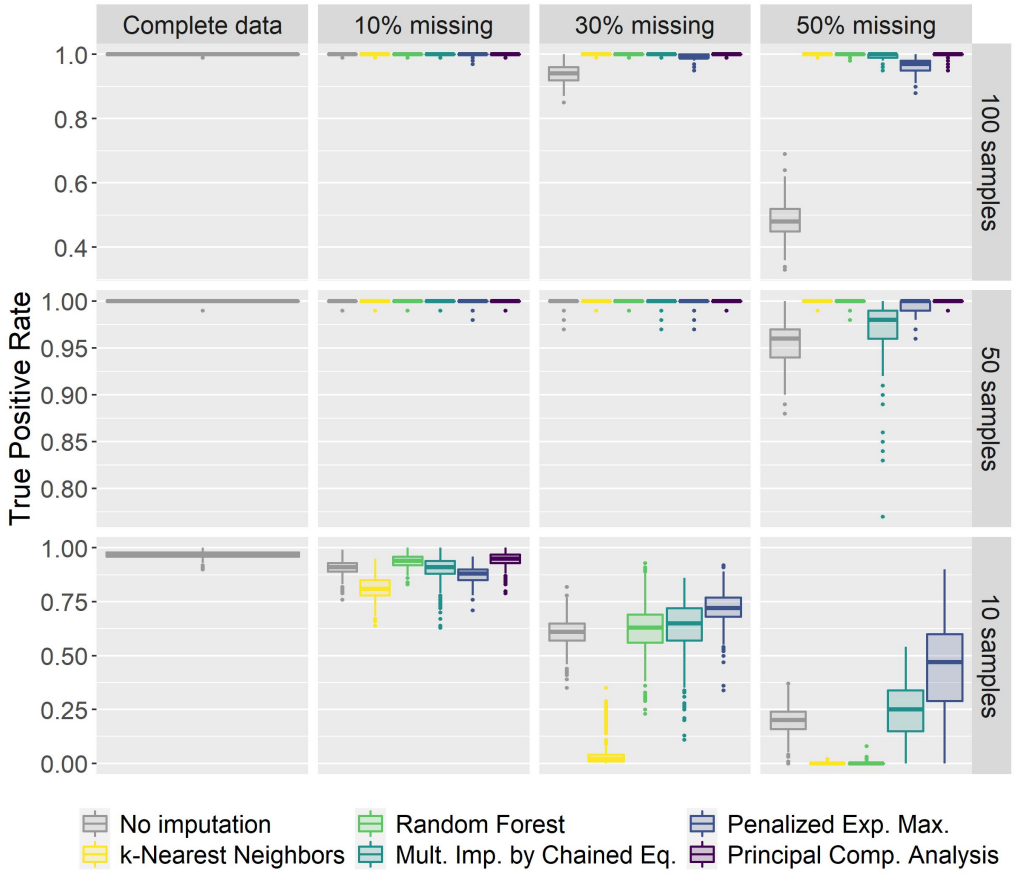


Figure 4.6: True positive rate for simulated data with three sample sizes, three proportions of MCAR missingness and different strategies of tackling missing values. A test was considered positive if three conditions were met: (1) adj. P-value ≤ 0.05; (2) est. $\log_2$ fold-change greater than 1; and (3) adj. P-value not more than 10^{18} times larger than adj. P-value for complete data. Test-outcomes were compared to simulated ground truth DAPs.

4.4 Computation time

In the evaluation of performance of different strategies for dealing with missing data, the required computation time of the algorithms is also of interest. We derived the average run time of the analysis procedure including imputation of missing values, for each evaluated method and different proportions of missingness in real data. Times were averaged over 100 iterations on a high-performance computing (HPC) cluster using a single core. For information on respective R packages see methods in Section 3.1. Note that in these comparisons, no method was run with parallelization, in order to allow for proper comparison on run time.

Comparing the required run time for different proportions of MCAR values (Table 4.1), the amount of missingness seemed to have little effect in general. However, there were differences between imputation methods. Imputation with MICE and RF had the longest average run time, around one hour for both. In contrast, analysis with imputation by kNN was on average as fast as without imputing (< 1 minute). Using PCA and PEM both had average run times of 3 to 7 minutes at different proportions of missing values. Corresponding times for MNAR missingness were very similar (Table 4.2). Considering the run times only, we would use other imputation algorithms instead of RF and MICE.

Table 4.1: Average run time in minutes for analysis using different missing data strategies in real data (with 110 samples and 1000 proteins) at different proportions of MCAR missing values. Computations run in R 3.6.0 on a HPC cluster with a single core. See Table 3.1 for information on used imputation packages.

	Average run time (min)		
	10% MCAR	30% MCAR	50% MCAR
No imputation	0.7	0.6	0.7
kNN	0.7	0.7	0.7
RF	53.7	58.6	61.6
MICE	76.8	66.5	62.0
PEM	9.7	7.3	3.6
PCA	3.2	4.7	6.5

Table 4.2: Average run time in minutes for analysis using different missing data strategies in real data (with 110 samples and 1000 proteins) at different proportions of MNAR missing values. Computations run in R 3.6.0 on a HPC cluster with a single core. See Table 3.1 for information on used imputation packages.

	Average run time (min)		
	10% MNAR	30% MNAR	50% MNAR
No imputation	0.7	0.7	0.8
kNN	0.7	0.7	0.8
RF	54.6	56.6	54.6
MICE	72.2	69.0	76.7
PEM	7.8	7.0	2.6
PCA	3.5	5.0	7.4

5

Discussion

Since the FPR was controlled in all analyses, no artefacts would be introduced by any imputation method under any of the tested circumstances. For large proportion proportions of missing data (50%) irrespective of missingness type: for small sample sizes ($n \leq 10$) imputation is advised against; and for large sample sizes ($n \geq 50$) imputation with PCA is recommended. For smaller proportions of missing data (10-30%) imputation did not increase the TPR compared to removing missing values and is therefore not explicitly recommended; but PCA-imputation can be performed without reduction in performance, if a complete dataset is necessary for subsequent analysis.

In a published review [6], the same conclusions have been made regarding FPR. However, the results differ in the evaluation of TPR, where they have suggested PEM closely followed by RF and PCA as the highest performing methods in this aspect when $n = 75$. Notably, the previous review [6] did not considered the level of uncertainty associated with tests, but rather only used the significance level for the binary classification. Although insightful, such evaluations merely indicate the rate of true DAPs detected, but do not account for differences in effect size uncertainty between the methods. Note also that the specific version of the MICE algorithm used in this thesis is internally developed and not yet open-source, hence it has not been included in any previous review.

Further, our results indicated that the analyses were not affected by missingness type. This conclusion is not in line with previous studies, which suggest that methods that assume MCAR, such as kNN and RF, are generally not as accurate in their imputation of left-censored MNAR values [5]. Indeed, the MNAR values are non-ignorable for likelihood-based inference. Hence the majority of the evaluated methods of dealing with missing data, particularly the complete-case analysis, could be expected to yield larger estimation bias when the missingness is MNAR [13, 15]. However, the extent of estimation bias associated with the non-ignorable (MNAR) missing values largely depends on the structure of both the missingness mechanism and the data. It could be the case that the function used to simulate MNAR data, in reality does not introduce considerable estimation bias, given the data in this evaluation.

The results obtained in real data (Figure 4.5) indicated that the TPR obtained without imputing decreased with the proportion of missing data. As expected, reduced effective sample sizes resulted in reduced power to detect DAPs. In contrast, analyses with simulated data of similar sample size as the real dataset ($n = 100$), indicated that performance was not affected by missingness until 50% of the data were missing (Figure 4.6). This could

indicate that the resulting signal-to-noise ratio in the simulated data were lower than in the real CCLE dataset. This is also supported by the distribution of adjusted P-values for the simulated data, where DAPs were associated with very small adjusted P-values (ranging from 10^{-60} to 10^{-10}) in the complete dataset. Hence it is possible that only large amounts of missing data will result in insufficient power to detect DAPs.

5.1 Limitations

The results for real- and simulated data differed, which could indicate that the model from which data were simulated did not fully capture the complex data structure of real data (e.g. CCLE data). It is however possible that it represented other MS-based datasets better, such as the Clinical Proteomic Tumor Analysis Consortium (CPTAC) as the original authors suggested [9]. This aspect could have been investigated by further varying the parameters of the simulation model, or by using additional real world MS dataset to compare with.

It should also be noted the generated probabilities in the MNAR-function are highly dependent on data and selected parameters, hence the estimation bias caused by the non-ignorability is as well. The comparison on how different methods perform with MCAR data in contrast with MNAR data should therefore be further evaluated. Given discrepancies between this study and the literature, there is a need for further studies to come to a consensus. Different parameter values could be tested, or, a different methodology for simulating left-censored values could be used [5].

Different thresholds for the ratio between adjusted P-values obtained for incomplete- or imputed data compared to that obtained for complete data (Q), should be evaluated since the value of Q may influence of the obtained TPR distributions. Alternatively, other methods could have been used to capture the level of uncertainty in the evaluation of performance. Possible alternatives could be to use Receiver Operation Characteristics curves [38], or to compare percentage increase in CI widths.

6

Conclusion

This study showed that large proportions of missing values (≥30%) affect the ability to identify DAPs in MS-based data. By evaluation of five commonly used imputation methods (kNN, RF, MICE, PEM and PCA), we found that the FPR was controlled below 5% in all analyses, indicating that no artefacts were introduced by any method under any of the conditions tested. The evaluations showed that for large sample sizes ($n \geq 50$) and large proportions of missing values (50%) imputation by PCA is likely to improve the true positive rate. For smaller proportions of missing data (10-30%), imputation did not increase the TPR; but PCA-imputation can be performed without reduction in performance. However, if the sample size is small ($n \leq 10$) imputation is advised against, due to poor performance of all evaluated imputation methods.

Bibliography

[1] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[2] D. I. Papac and Z. Shahrokh, "Mass spectrometry innovations in drug discovery and development," *Pharmaceutical research*, vol. 18, no. 2, pp. 131–145, 2001.

[3] K. T. Do, S. Wahl, J. Raffler, S. Molnos, M. Laimighofer, J. Adamski, K. Suhre, K. Strauch, A. Peters, C. Gieger, *et al.*, "Characterization of missing values in untargeted ms-based metabolomics data and evaluation of missing data handling strategies," *Metabolomics*, vol. 14, no. 10, p. 128, 2018.

[4] P. Wang, H. Tang, H. Zhang, J. Whiteaker, A. G. Paulovich, and M. Mcintosh, "Normalization regarding non-random missing values in high-throughput mass spectrometry data," in *Biocomputing 2006*, pp. 315–326, World Scientific, 2006.

[5] R. Wei, J. Wang, M. Su, E. Jia, S. Chen, T. Chen, and Y. Ni, "Missing value imputation approach for mass spectrometry-based metabolomics data," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[6] L. M. Bramer, J. Irvahn, P. D. Piehowski, K. D. Rodland, and B.-J. M. Webb-Robertson, "A review of imputation strategies for isobaric labeling-based shotgun proteomics," *Journal of Proteome Research*, 2020.

[7] M. Kokla, J. Virtanen, M. Kolehmainen, J. Paananen, and K. Hanhineva, "Random forest-based imputation outperforms other methods for imputing lc-ms metabolomics data: a comparative study," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–11, 2019.

[8] Y. V. Karpievitch, A. R. Dabney, and R. D. Smith, "Normalization and missing value imputation for label-free lc-ms analysis," *BMC bioinformatics*, vol. 13, no. S16, p. S5, 2012.

[9] J. Wang, P. Wang, D. Hedeker, and L. S. Chen, "Using multivariate mixed-effects selection models for analyzing batch-processed proteomics data with non-ignorable missingness," *Biostatistics*, vol. 20, no. 4, pp. 648–665, 2019.

[10] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies," *Journal of proteome research*, vol. 15, no. 4, pp. 1116–1125,

2016.

[11] J. D. O'Connell, J. A. Paulo, J. J. O'Brien, and S. P. Gygi, "Proteome-wide evaluation of two common protein quantification methods," *Journal of proteome research*, vol. 17, no. 5, pp. 1934–1942, 2018.

[12] A. Brenes, J. Hukelmann, D. Bensaddek, and A. I. Lamond, "Multibatch tmt reveals false positives, batch effects and missing values," *Molecular & Cellular Proteomics*, vol. 18, no. 10, pp. 1967–1980, 2019.

[13] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019.

[14] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[15] S. Van Buuren, *Flexible imputation of missing data*. CRC press, 2018.

[16] C. K. Enders, *Applied missing data analysis*. Guilford press, 2010.

[17] L. S. Chen, R. L. Prentice, and P. Wang, "A penalized em algorithm incorporating missing data mechanism for gaussian parameter estimation," *Biometrics*, vol. 70, no. 2, pp. 312–322, 2014.

[18] L. S. Chen, J. Wang, X. Wang, and P. Wang, "A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments," *The annals of applied statistics*, vol. 11, no. 1, p. 114, 2017.

[19] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[20] B.-J. M. Webb-Robertson, H. K. Wiberg, M. M. Matzke, J. N. Brown, J. Wang, J. E. McDermott, R. D. Smith, K. D. Rodland, T. O. Metz, J. G. Pounds, *et al.*, "Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics," *Journal of proteome research*, vol. 14, no. 5, pp. 1993–2001, 2015.

[21] P. Keerin, W. Kurutach, and T. Boongoen, "Cluster-based knn missing value imputation for dna microarray data," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 445–450, IEEE, 2012.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[23] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," 1999.

[24] L. Chen and P. Wang, *PEMM: A Penalized EM algorithm incorporating missing-data*

mechanism, 2014. R package version 1.0.

[25] J. Josse and F. Husson, "Handling missing values in exploratory multivariate data analysis methods," *Journal de la Société Française de Statistique*, vol. 153, no. 2, pp. 79–99, 2012.

[26] D. J. Stekhoven and P. Buehlmann, "Missforest - non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[27] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[28] D. P. Nusinow, J. Szpyt, M. Ghandi, C. M. Rose, E. R. McDonald III, M. Kalocsay, J. Jané-Valbuena, E. Gelfand, D. K. Schweppe, M. Jedrychowski, *et al.*, "Quantitative proteomics of the cancer cell line encyclopedia," *Cell*, vol. 180, no. 2, pp. 387–402, 2020.

[29] G. Verbeke, "Linear mixed models for longitudinal data," in *Linear mixed models in practice*, pp. 63–153, Springer, 1997.

[30] B. L. Welch, "The generalization of student's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

[31] J. A. Rice, *Mathematical statistics and data analysis*. Nelson Education, 2006.

[32] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, 2004.

[33] R. E. Kirk, "Practical significance: A concept whose time has come," *Educational and psychological measurement*, vol. 56, no. 5, pp. 746–759, 1996.

[34] T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu, *impute: impute: Imputation for microarray data*, 2019. R package version 1.60.0.

[35] D. J. Stekhoven, *missForest: Nonparametric Missing Value Imputation using Random Forest*, 2013. R package version 1.4.

[36] J. Josse and F. Husson, "missMDA: A package for handling missing values in multivariate data analysis," *Journal of Statistical Software*, vol. 70, no. 1, pp. 1–31, 2016.

[37] J. Josse and F. Husson, "Selecting the number of components in principal component analysis using cross-validation approximations," *Computational Statistics & Data Analysis*, vol. 56, no. 6, pp. 1869–1879, 2012.

[38] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *Journal of biomedical*

informatics, vol. 38, no. 5, pp. 404–415, 2005.

A

Additional background

A.1 Derivation of conditions for ignorable missingness in likelihood inference

The distribution of the actual observed data, i.e. of the variables $(\mathbf{Y}_{\text{obs}}, \mathbf{M})$, is obtained by integrating $\mathbf{Y}_{\text{mis}}$ out of the joint density of $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ and $\mathbf{M}$. That is,

$$f(\mathbf{Y}_{\text{obs}}, \mathbf{M}|\theta, \phi) = f(\mathbf{M}|\mathbf{Y}_{\text{obs}}, \phi) \times \int f(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}|\theta) d\mathbf{Y}_{\text{mis}}. \tag{A.1}$$

If the missingness mechanisms is at least missing at random (MAR), such that $f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\mathbf{Y}_{\text{obs}}, \phi)$ holds, the joint distribution in (A.1) can be simplified:

$$\begin{aligned} f(\mathbf{Y}_{\text{obs}}, \mathbf{M}|\theta, \phi) &= f(\mathbf{M}|\mathbf{Y}_{\text{obs}}, \phi) \times \int f(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}|\theta) d\mathbf{Y}_{\text{mis}} \\ &= f(\mathbf{M}|\mathbf{Y}_{\text{obs}}, \phi) \times f(\mathbf{Y}_{\text{obs}}|\theta) \\ &\propto f(\mathbf{Y}_{\text{obs}}|\theta). \end{aligned} \tag{A.2}$$

Inference about θ can be obtained from maximizing the joint likelihood of (θ, ϕ). The full likelihood of the unknown parameters θ and ϕ is any function of (θ, ϕ) proportional to $f(\mathbf{Y}_{\text{obs}}, \mathbf{M}|\theta, \phi)$:

$$\mathbf{L}_{\text{full}}(\theta, \phi|\mathbf{Y}_{\text{obs}}, \mathbf{M}) \propto f(\mathbf{Y}_{\text{obs}}, \mathbf{M}|\theta, \phi), \qquad (\theta, \phi) \in \Omega_{\theta,\phi}. \tag{A.3}$$

If we ignore the missingness mechanism, i.e. ignore the parameter ϕ, we may instead base inference on the likelihood of θ alone. The likelihood of θ based on the observed data, ignoring the missingness mechanism, is any function of θ proportional to $f(\mathbf{Y}_{\text{obs}}|\theta)$:

$$\mathbf{L}_{\text{ign}}(\theta|\mathbf{Y}_{\text{obs}}) \propto f(\mathbf{Y}_{\text{obs}}|\theta), \qquad \theta \in \Omega_{\theta}. \tag{A.4}$$

Consequently, if the missingness mechanisms is at least MAR and $\Omega_{\theta,\phi} = \Omega_{\theta} \times \Omega_{\phi}$, we have

$$\mathbf{L}_{\text{ign}} \propto \mathbf{L}_{\text{full}} \propto f(\mathbf{Y}_{\text{obs}}|\theta). \tag{A.5}$$

Under these conditions, the simpler likelihood $\mathbf{L}_{\text{ign}}$ that ignores the missingness mechanism can be used instead of $\mathbf{L}_{\text{full}}$ for likelihood based inference about θ [13].

If only the first condition of ignorable missingness is met, there is possibly still information about θ in the factor $f(\mathbf{M}|\mathbf{Y}_{\text{obs}}, \phi)$ in equation A.2 that is being ignored. However, it does not affect the sampling distribution of any of the statistics which is conditioned on the fixed but unknown parameters (θ, ϕ). Therefore, inference based on the simpler likelihood is valid, but not fully efficient, also if only the first condition is met [13].

A.2 Benjamini-Hochberg procedure for FDR control

The expected proportion of false positives (FPs) among all positive outcomes is called the false discovery rate (FDR):

$$\text{FDR} = \text{Exp}\left[\frac{\text{FPs}}{\text{FPs} + \text{TPs}}\right].$$

The Benjamini-Hochberg procedure is a method to control the FDR at the significance level α. The procedure begins with ordering the p hypothesis by ascending p-values, denoting $p_{(j)}$ the p-value at the jth position. Then by finding

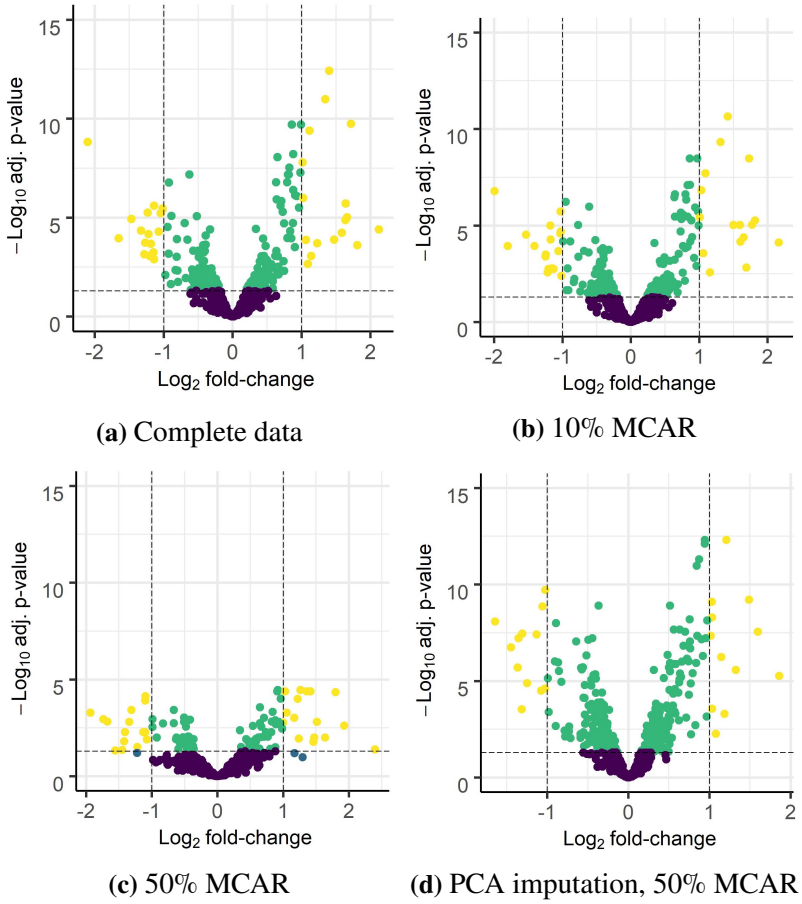
$$L = \underset{j \in (1,\ldots,p)}{\text{argmax}}\ p_{(j)} < \frac{\alpha j}{p},$$

we reject all hypotheses for which $p_{(j)} < p_{(L)}$. Assuming independent tests, the method ensures estimated FDR $\leq \alpha$ and the number of FPs is expected to be controlled at the significance level [22].

B

Supplementary results: establishment of performance criteria

Distributions of P_{adj}-values over estimated $\log_2$ fold-change for real data (Figure B.1) and simulated data (Figure B.2). With both real and simulated data: 10% missingness did not noticeably affect the P_{adj}-values; with 50% missingness P_{adj}-values were severely increases; and by imputation low P_{adj}-values were restored.



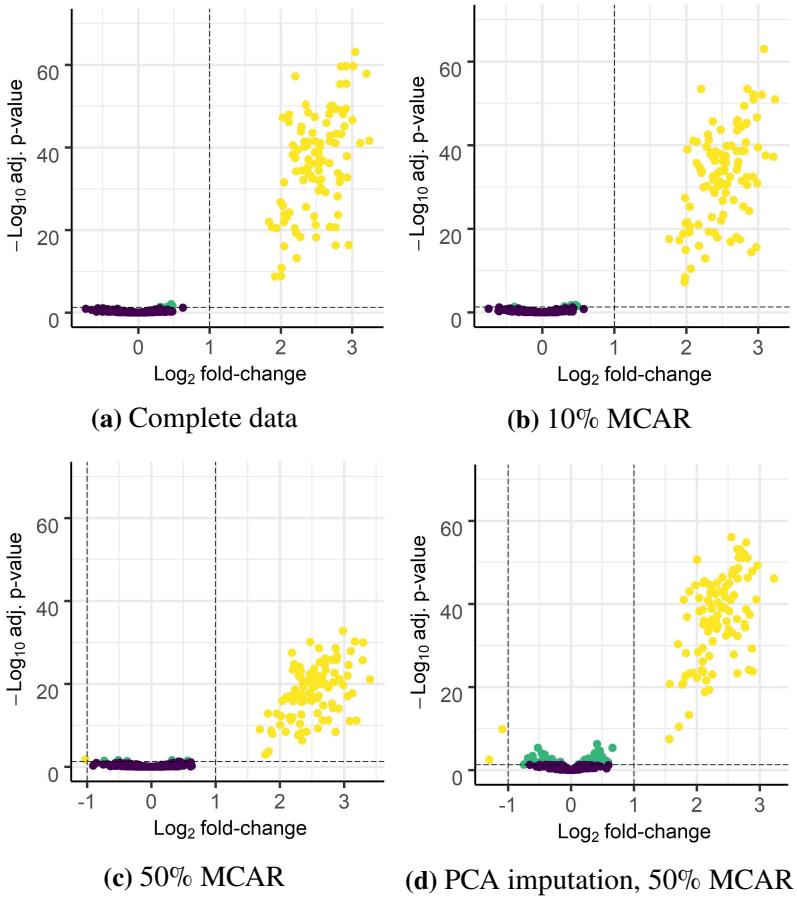
(a) Complete data

(b) 10% MCAR

(c) 50% MCAR

(d) PCA imputation, 50% MCAR

Figure B.1: FDR-adjusted P-value on negative $\log_{10}$ scale against estimated $\log_2$ fold-change for one iteration of analysis with real data ($p = 1000$). Points colored according to: (yellow) significant and relevant; (green) significant but non-relevant; (purple) non-significant and non-relevant; and (blue) relevant but non-significant.



(a) Complete data **(b)** 10% MCAR

(c) 50% MCAR **(d)** PCA imputation, 50% MCAR

Figure B.2: FDR-adjusted P-value on negative $\log_{10}$ scale against estimated $\log_2$ fold-change for one iteration of analysis with real data ($p = 1000$). Points colored according to: (yellow) significant and relevant; (green) significant but non-relevant; and (purple) non-significant and non-relevant.

C

Supplementary results: false positive rate

Figures C.1 and C.2 shows the distribution FPR when MNAR values were introduced, for real and simulated data respectively.

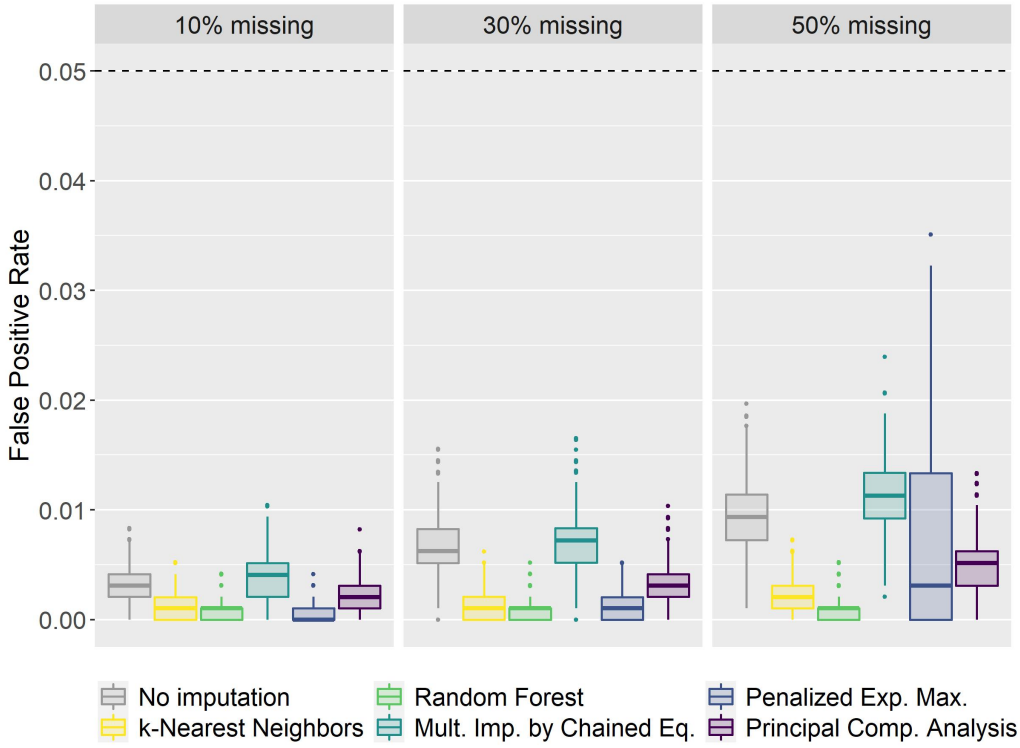


Figure C.1: False positive rate for real data with three proportions of MNAR missingness and different strategies of tackling missing values. A test was considered positive if three conditions were met: (1) adj. P-value $\leq$0.05; (2) estimated $\log_2$ fold-change greater than 1; and (3) P-value not more than 10^3 times larger than the P-value for complete data. Test-outcomes were compared to the ground truth of DAPs detected when using the complete data.

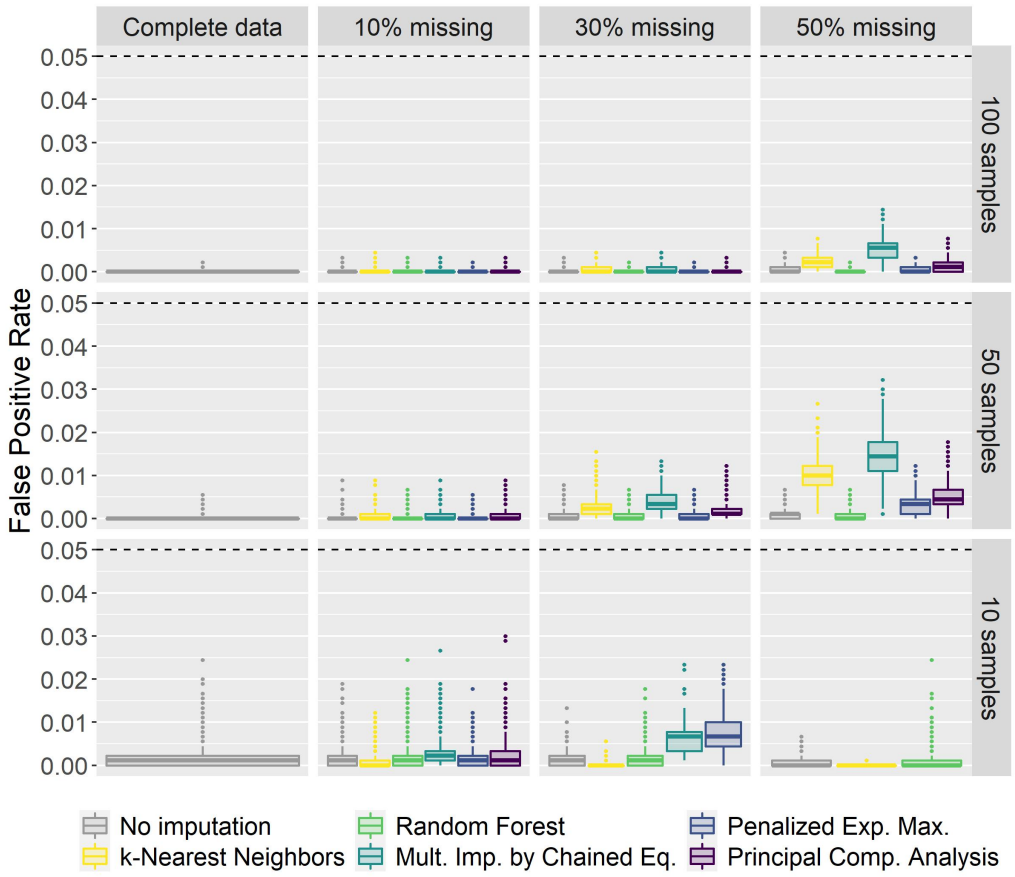


Figure C.2: False positive rate for simulated data with three sample sizes, three proportions of MNAR missingness and different strategies of tackling missing values. A test was considered positive if three conditions were met: (1) adj. P-value $\leq$0.05; (2) estimated $\log_2$ fold-change greater than 1; and (3) adj. P-value not more than 10^{18} times larger than the adj. P-value for complete data. Test-outcomes were compared to the ground truth of simulated DAPs.

D

Supplementary results: true positive rate

Figures D.1 and D.2 shows the distribution TPR when MNAR values were introduced, for real and simulated data respectively.

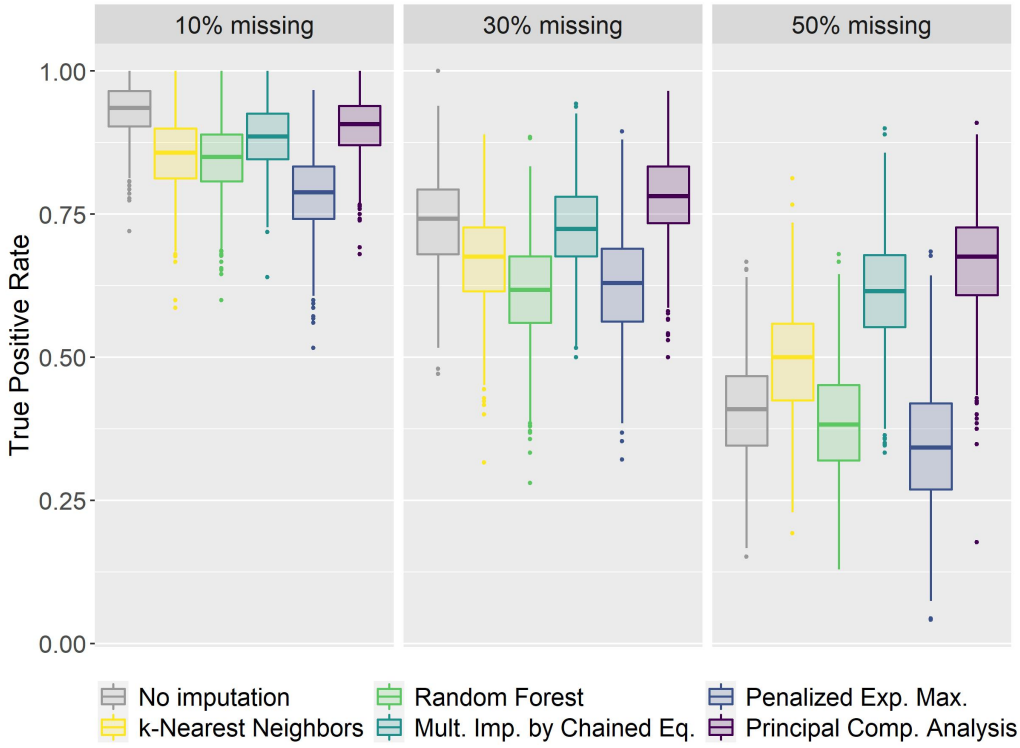


Figure D.1: False positive rate for real data with three proportions of MNAR missingness and different strategies of tackling missing values. A test was considered positive if three conditions were met: (1) adj. P-value ≤ 0.05; (2) estimated $\log_2$ fold-change greater than 1; and (3) P-value not more than 10^3 times larger than the P-value for complete data. Test-outcomes were compared to the ground truth of DAPs detected when using the complete data.

Figure D.2: False positive rate for simulated data with three sample sizes, three proportions of MNAR missingness and different strategies of tackling missing values. A test was considered positive if three conditions were met: (1) adj. P-value ≤ 0.05; (2) estimated $\log_2$ fold-change greater than 1; and (3) adj. P-value not more than 10^{18} times larger than the adj. P-value for complete data. Test-outcomes were compared to the ground truth of simulated DAPs.