METHODS FOR ADAPTING GLOBAL MASS SPECTROMETRY BASED METABOLOMICS TO THE CLINICAL ENVIRONMENT

Jacob Edward Wulff

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2018

Approved by:

Lloyd J. Edwards

Jianwen Cai

Michael G. Hudgens

Chirayath M. Suchindran

Jonathan McDunn

ABSTRACT

Jacob Edward Wulff: Methods for Adapting Global Mass Spectrometry based Metabolomics to the Clinical Environment
(Under the direction of Lloyd J Edwards)

Metabolomics is a maturing field with successful application to research areas such as biomarker discovery and mechanisms of disease. With the ability to profile hundreds or even thousands of biochemicals simultaneously, many of which are also used in various laboratory diagnostics, the technology has the potential to replace a battery of clinical tests with a single test. However, the current state of global analysis presents several challenges for the clinical environment. This dissertation addresses two of these challenges. First is handling of missing values with respect to comparing an individual sample against a reference population. Second is the semi-quantitative nature of the liquid chromatography mass spectrometry.

The first paper explores basic properties of metabolites, specifically the statistical distribution of metabolite concentrations and correlation between them. In human sample sets covering three different sample material appropriate for clinical testing, raw ion counts are shown to be vastly non-normal and consistently having a heavy right skew. Natural log-transformation is effective at removing this skewness and inducing Gaussian behavior, though departures from normality may persist in the tails of the distributions. Correlation between library-matched metabolites after removing artifact related features is also shown to be of only moderate degree in most cases.

In the second paper, application of the log transformation is used to account for missing values in estimating population parameters of a reference cohort. Missing values are largely

attributed to the true level falling below the detection limit of the instrument. Combining this assumption with the Gaussian model leads to two parametric approaches being introduced for the estimation of population parameters. These methods are shown to outperform standard imputation approaches in the field using a combination of simulations and real metabolomic datasets.

The third paper addresses merging multiple global LC-MS metabolomic sets of the same biological sample type together. Typical normalization methods meant to account for sample to sample variation are presented and compared to alternative approaches using technical replicates and within batch scaling. Concentrations from targeted analysis of eight clinical biomarkers are used to show the superiority of these alternative approaches.

To my wife Teresa, who had the strength not just to put up with me over these eight long years but help carry me through as well. And to my children Brianna and Junius, without whom this dissertation would have been done so many years earlier but with so much less personal meaning.

ACKNOWLEDGEMENTS

First and foremost, I wish to thank my committee chair, Dr. Lloyd Edwards, for his patience and guidance. He understood the challenges of completing a dissertation while working full time and raising a young family, without which this dissertation would not have been possible. His positive attitude and encouragement provided a much-needed light on a long, dark road.

I am grateful to my other committee members, Dr. Jianwen Cai, Dr. Michael Hudgens, Dr. Chirayath Suchindran and Jon McDunn, as well for their participation and comments which strengthen the results of this work.

My deepest appreciation to Dr. Matthew Mitchell for his many contributions over the years, for tolerating my frequent questions, for mentoring me on what it means to be a professional and for being a friend.

I am grateful to Dr. Adam Kennedy for his knowledge into biochemistry, his skill at navigating across multiple organizational entities and for being a good friend.

Thank you also to Metabolon and the Department of Molecular and Human Genetics at Baylor College of Medicine for permission and access to the data used in this research.

Finally, thank you to Lorraine Riordan and Marcus Maxon for their technical skills and always knowing the right thing to say.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

4MOP	4-methyl-2-oxopentanoate / 4-methyl-2-oxopentanoic acid
AD	Anderson-Darling
AHB	α-hydroxybutyrate / α-hydroxybutyric acid
ALS	Amyotrophic Lateral Sclerosis
ANCH	Anchor Normalization
ANOVA	Analysis of Variance
AVG	Single Imputation with Average
BAT	Batch Normalization
BHB	β-hydroxybutyrate / β-hydroxybutyric acid
BPCA	Bayesian Principal Component Analysis
CAP	College of American Pathologists
CE	Capillary Electrophoresis
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
CLIA	Clinical Laboratory Improvement Amendments
CLOW	Cyclic Lowess
CN	Contrast Normalization
CSF	Cerebral Spinal Fluid
CV	Coefficient of Variation
Da	Dalton; unit of molecular mass
DMSO	Dimethyl-sulfoxide
DW	de Wet-Venter
ECF	Empirical Characteristic Function

EDF	Empirical Distribution Function
eGFR	Estimated Glomerular Filtration Rate
EI	Electron Impact
ESI	Electrospray Ionization
FDA	US Food and Drug Administration
FDR	False Discovery Rate
GC	Gas Chromatography
GCC	Glycocholate / Glycocholic acid
GCCDC	Glycochenodeoxycholate / Glycochenodeoxycholic acid
GFR	Glomerular Filtration Rate
GLOG	Generalized Log
HESI	Heated Electrospray Ionization
HGD	Homogentisate Oxidase
HPLC	High-Performance Liquid Chromatography
IEM	Inborn Errors of Metabolism
IGT	Impaired Glucose Tolerance
IRAS	Insulin Resistance Atherosclerosis Study
kNN	k Nearest Neighbors
KS	Kolmogorov-Smirnoff
LB	Linear Baseline Scaling
LC	Liquid Chromatography
LDT	Laboratory Developed Test
LGP	1-linoleoylglycerophosphocholine

LOD	Limit of Detection
m/z	Mass to charge ratio
MA	Minus-Average
MAD	Median Absolute Deviation
MAR	Missing at Random
MCAD	Medium-chain Acyl-CoA Dehydrogenase
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MDRD	Modification of Diet and Renal Disease
mGFR	Measured Glomerular Filtration Rate
MI	Multiple Imputation
MICE	Multiple Imputation by Chained Equations
MIN	Single Imputation with Minimum
ML	Maximum Likelihood
MNAR	Missing Not at Random
MS	Mass Spectrometry
MSTUS	MS-Total Useful Signal
MVN	Multivariate Normality
NAFLD	Non-Alcoholic Fatty Liver Disease
NASH	Non-Alcoholic Steatosis
NLPCA	Non-linear Principal Component Analysis
NMR	Nuclear Magnetic Resonance Spectroscopy
NONE	No Imputation

OGTT	Oral Glucose Tolerance Test
OOB	Out of Bag
PBD	Peroxisomal Biogenesis Disorder
PCA	Principal Component Analysis
PPCA	Probabilistic Principal Component Analysis
PQN	Probabilistic Quotient Normalization
QQ	Quantile-Quantile
RF	Random Forest
RR	Rankit Regression
RSD	Relative Standard Deviation
SB	Smith-Bain
sCR	Serum Creatinine
SF	Shapiro-Francia
SI	Single Imputation
SLR	Simple Linear Regression
TC	Taurocholate / Taurocholic acid
TCDC	Taurochenodeoxycholate / Taurochenodeoxycholic acid
TIC	Total Ion Count
UPLC	Ultra High-Performance Liquid Chromatography
WB	Weisberg-Bingham

CHAPTER 1: LITERATURE REVIEW

1.1. Introduction to Metabolomics

Metabolomics is the field concerned with the study of small molecule biochemicals, collectively known as metabolites, which are the intermediates and end products of metabolism. The metabolome refers to the entire set of such small molecules present in a biological organism, or specific part thereof. Metabolomic analysis is thus tasked with identifying, measuring and understanding the metabolome. This requires a combination of analytical chemistry, biochemistry and biology. Metabolites are generally defined as having molecular weight < 1k Daltons (Da; grams/mol) which helps to differentiate metabolomics from areas that focus on certain classes or species such as proteomics or lipomics [1-3]. Metabolites include a diverse background of chemical compounds including amino acids, carbohydrates, nucleotides, peptides and lipids among others. Such a diverse set of compounds present an analytical challenge to measure but offer a wealth of information. Indeed, because metabolites are directly involved with cellular function, information drawn from a metabolomic study will tend to be more phenotypic rather than genotypic [4, 5]. Beebe and Kennedy [6] as well as Koen *et al*. [7] illustrate the importance of incorporating metabolic knowledge along with genetic and phenotypic information. In this way metabolomics may be seen as compliments to other omics fields such as proteomics, genomics or transcriptomics. Metabolomics offers a wide range of applications from cancer research to bioprocessing. For the purposes of the proceeding document our interest is focused on clinical application and disease screening. Inborn Errors of Metabolism, a large class of congenital disorders resulting in impaired metabolic function, will

have a large presence throughout the document due to the available data; however, the majority of the principles discussed are generalizable to other applications of metabolomics.

The work contained in this dissertation focuses on High Performance Liquid Chromatography Mass Spectrometry (HPLC-MS) which is the most popular method for obtaining high-throughput metabolomic data [8, 9]. A brief summary of this technology is provided here to help familiarize the reader with some the background of the paper.

1.2. Chromatography and Mass Spectrometry

Chromatography is used to separate dissolved compounds in complex mixture by exploiting the differential partitioning between two phases: the mobile phase and the stationary phase. The mixture to be separated is injected onto a fluidized bed containing the stationary phase, typically beads that are functionalized with lipophilic organic molecule. The mobile phase, a liquid state typically composed of an inert solvent, dissolves the mixture and is pumped through the stationary phase. Dissolved components partition between the liquid and surface of the stationary phase. Interaction with the surface varies by compound, causing the individual compounds to pass through the stationary phase at different. This is usually in the form of lighter compounds moving through more quickly while heavier compounds move more slowly and leads to a separation of the compounds over time. Chromatographic systems are identified by the type of mobile phase used. Gas chromatography (GC) refers to a gaseous mobile phase, usually a noble gas such as helium, while liquid chromatography (LC) refers to a liquid mobile phase, which is most commonly an aqueous solution containing an organic solvent (acetonitrile, methanol, etc.), acid (formic, phosphoric, etc.), and/or salt.

As the different compounds elute from the chromatogram they are injected into the mass spectrometer (MS). The MS system consists of three main components: an ion source, a mass

analyzer and a detector. The ion source ionizes the individual molecules of the compound. These ions are then passed to the mass analyzer, which separates them according to their mass to charge (m/z) ratio. Finally, the detector tracks the individual m/z ratios and the number of ions associated with each. MS can be used to identify a compound based on the resulting profile of m/z ratios. When an individual molecule is ionized, the chemical bonds holding the molecule together may break, leading to a fragment ion. The number of ways in which a compound may fragment and the proportion of molecules that form those fragments relate directly to the structure of the parent molecule in addition to the ionization state of the parent molecule and type of fragmentation chemistry that is employed. This leads to one of the fundamental powers of MS: Different molecules produce different fragmentation spectrums. By cross-referencing the fragmentation pattern of an unknown compound against a library of known spectrums the compound may be identified. It should be noted that sources of ionizations can be divided into two broad categories which also correspond to instrumentation used. Hard ionization methods, such as Electron Impact (EI), lead directly to the fragmentation of parent compounds and are well suited for GC systems. Soft ionization, which includes the popular Electrospray Ionization (ESI) and other atmospheric pressure chemical ionization methods, ionizes the compound in such a way that a charge is supported by the molecule without fragmentation. While it is possible to measure the m/z ratio of parent molecules in this state it is not possible to identify the compound without the daughter ion spectrum. Therefore, soft ionization methods usually involve another step to force the fragmentation. Tandem Mass Spectrometry, or MS-MS, involves multiple steps of the mass spectrometry process with fragmentation occurring in between [10]. Soft ionization techniques are well suited for LC systems.

In addition to the ion source, MS can be used to produce either positively or negatively

charged ions. Compounds, depending on their physical and chemical properties, may ionize and fragment much better in one ion mode over the other. Due to the wide range of physical and chemical diversity spanned by metabolites, it is advisable to incorporate multiple instrument conditions into the analysis in order to maximize coverage of the metabolome. The same mass analyzer paired to the same chromatography system can be used in both modes; however, the mobile phase will in most cases need to be different in order to facilitate positive ion formation or negative ion form [11].

GC-MS has been around for decades and by nature it is best suited for volatile compounds, or those that can be easily derivatized to a volatile state [12]. However, non-volatile and low vapor pressure, non-derivatizable compounds may fail to ionize under GC conditions. Despite these shortcomings, GC remained popular for many years until the development of high pressure pumps and other hardware for LC systems in the 1980s and 1990s. These advancements led to a superior level of sensitivity for the LC-MS system [13] making HPLC-MS the standard in metabolomic analysis.

Alternatives to the pairing of mass spec and chromatography do exist but tend to suffer from a much lower coverage of the metabolome in comparison, particularly with HPLC-MS. The majority of alternatives, in practices, are focused on specific compound classes or other niche roles in which LC-MS does not perform as well. For example, capillary electrophoresis (CE) is a separation technique that offers tremendous potential in terms of sensitivity and has been successfully coupled to MS for the purpose of metabolomics analysis [14]. But combination with ESI is not trivial and hence the full potential of CE-MS has yet to be realized. In its current state CE-MS is best suited for the measurement of highly polar compounds [15, 16]. Issaq *et al*. provides a good overview of the separation techniques available in metabolomics [17].

The main competition to LC-MS for identification of metabolites would be Nuclear Magnetic Resonance Spectroscopy (NMR). NMR exploits the electromagnetic spin of specific nuclei containing an odd number of electrons to produce a spectrum of the number of differently bonded target nuclei and the relative proportions of each. In a similar manner to the MS spectrum, the NMR spectrum can be used to identify compounds. NMR has a number of advantages over LC-MS, namely that it does not require chromatography to separate compounds and, as it does require the breaking of chemical bonds, it is non-destructive to the sample. NMR has successfully been used in many metabolomics studies [18, 19] but the sensitivity is far inferior to that of HPLC-MS [20, 21]. For more on NMR see Larive *et al*. [22] and for a general overview of the analytical methods in metabolomics see Verpoorte *et al*. [23].

1.3. The Mass Spectrum

As mentioned above, the MS process results in the breaking of chemical bonds, producing multiple ion features per compound. Consider a simple example with pentane, which is a five-carbon chain with chemical formula C5H12. When ionized in a mass spectrometer any of the C-C bonds may be broken and so the possible m/z ratios correspond to an ion with either 1, 2, 3 or 4 carbons. Note that in metabolomics the majority of analytes involved are too small to stably support more than a single charge, so the m/z ratio is generally equivalent to the fragment's mass. The relative heights of the peaks associated with each m/z relate to the relative stability of the ion (or corresponding neutral fragment) in question and serve to differentiate compounds with similar mass from one another. 2-methylbutane is a four-carbon chain in which the second carbon in the chain has a single carbon branch in place of a hydrogen. The chemical formula is identical to that of pentane, C5H12, but the branched structure of 2-methylbutane causes the four-chain carbon ion (peak centered around 57 m/z) to form more readily than it does with

pentane. *Figure 1.1*, curtesy of Chemguide [24], displays the fragmentation mass spectra of these two compounds. Global profiling will capture thousands to tens of thousands of features related to hundreds to thousands of metabolites.

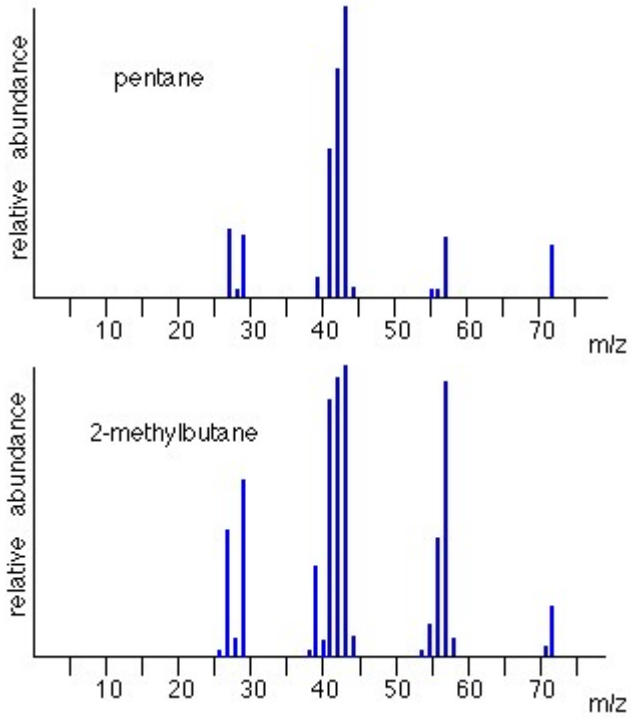


Figure 1.1: Mass Spectrums of pentane and 2-methylbutane. Both produce same mass fragments but at different relative proportions.

Separation of the ion features is performed by the mass analyzer with an emphasis on both resolution and sensitivity. Resolution refers to the ability to distinguish between two peaks, while sensitivity refers to the ability to detect an ion source. Among the most popular in metabolomics are quadrapole, linear ion trap (LIT), orbitrap, time-of-flight (TOF), and Fourier-transform (FT) [25]. MS-MS in inherent to some analyzers, like the ion traps and FT, allowing for multi-stage ionization within the same instrument. Some can be joined together, such as quadrapole-TOF configuration, to produce a similar multi-stage process and origination of the phrase “tandem MS”. Other analyzers can only operate in the traditional MS setup. For more specifics on the use of mass analyzers in metabolomics see Vékey [26].

1.4. Ion-centric vs. Chemo-centric

Being able to associate fragment features back to their parent compound requires more advanced instrumentation, such as MS-MS and/or pattern recognition software. When these resources are available, analysis may performed at the compound level which is referred to here as the chemo-centric approach. Measurement will be carried on the quantitation, or quant, ion.

This quant ion may be the parent ion itself, typical for MS-MS, or it may be a pre-determined daughter peak used to signify presence of the compound. In the latter case, the quant ion is usually the most dominant peak in the compound's spectrum. Minor peaks may also be used to confirm identification of the compound; however, these peaks are not used to quantify the compound. Absent the resources for the chemo-centric approach, data analysis must be carried out on all ion features without regard to parental origin. There are clear implications of the approach used on the resulting data. If an ion-centric approach is taken, the number of features can be up to 10,000 [27] whereas a chemo-centric approach will yield a much lower number of distinct biochemicals [28, 29]. Ion fragments associated with same parent molecule will extremely correlated, and thus resolution of these fragments into a single feature preferentially removes the most correlated variables. Finally, structural identification features allows for the removal of known contaminants which constitute the majority of features produced in MS metabolomics [30]. Removing these features not only reduces the size of the data but also likely alters characteristics of the data since these features are entirely the result of sample processing and lack any biological qualities as is the case with metabolites. It is worth pointing out that a chemo-centric approach does not by itself identify compounds by name. Naming requires referencing observed spectrums against an internal database, which can be very costly and time consuming to maintain in-house. Much of published literature has involved ion-centric data likely due to the challenges inherent to compound identification [31]. A few public and commercial databases are available [32-34]; however, precise matching can be difficult if the conditions (instrumentation, sample preparation, processing, etc.) in the study do not match those used to develop the library. Therefore, chemo-centric data may contain a mix of named and un-named biochemicals.

1.5. Data Structure and Routine Analysis

A metabolomics dataset is an array of n experimental sampling units, or more plainly just samples, and m metabolites (chemo-centric) or features (ion-centric). Orientation is arbitrary, but in these papers the convention is that the rows are the samples and the columns are the features. As such, denoting such a set as $\boldsymbol{Y}_{(n\times m)}$ gives

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix}$$

in which the sets of individual $y_{ji}'s$ are the observed value of the i^{th} feature in the j^{th} subject, $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$. In the case of MS, which is the focus of this dissertation, these observations are ion counts derived from the area of a peak associated with the metabolite or feature in question. Ion counts vary dramatically between metabolites with average values ranging from the tens of thousands up to hundreds of millions or even billions, and hence are treated as continuous variables for statistical analysis purposes. On occasion it may be useful to consider the matrix as a collection of features or samples. This amounts to segmenting the data into a series of columns (features):

$$\boldsymbol{Y} = [\, \boldsymbol{y}_{.1} \quad \cdots \quad \boldsymbol{y}_{.m}]$$

or a stack of rows (samples):

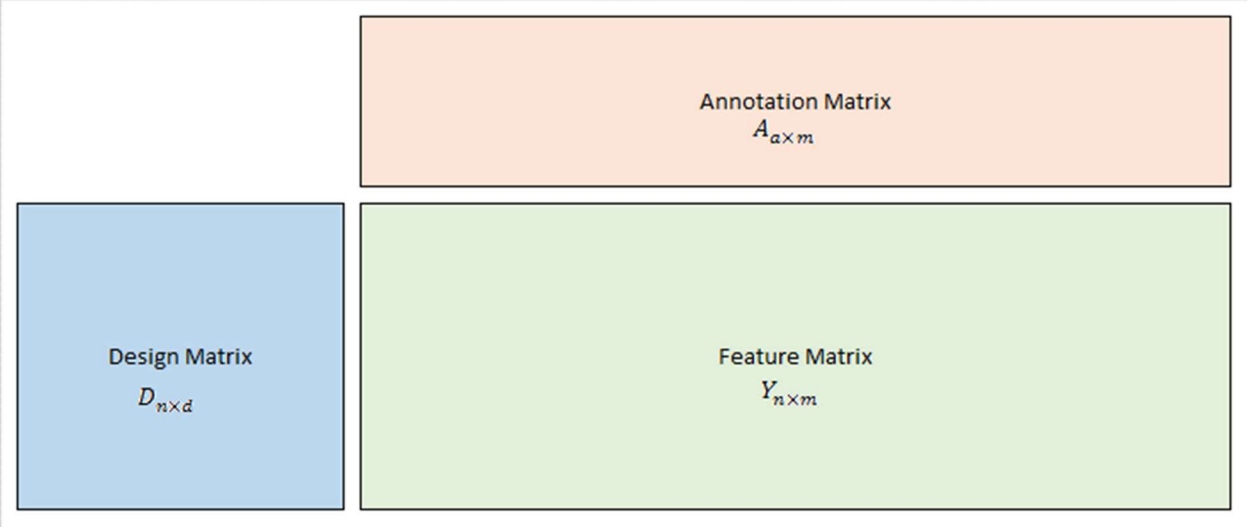
$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_{1.} \\ \vdots \\ \boldsymbol{y}_{n.} \end{bmatrix}$$

From here on the notation $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are used to generically indicate all the observed values for either the i^{th} feature in the data or the j^{th} subject. That is $\boldsymbol{y}_i = \boldsymbol{y}_{.i} = \{y_{1i}, \ldots, y_{ni}\}'$ and $\boldsymbol{y}_j = \boldsymbol{y}_{j.} = \{y_{j1}, \ldots, y_{jm}\}$.

Generally, a complete metabolomic set, illustrated in *Figure 1.2*, will also include details on the samples and features, respectively. The meta-data, labeled here as $\boldsymbol{D}_{n\times d}$, contains sample details most notably the study design information such as group assignments as well as potential factors of interest and possible covariates. Similarly, in certain cases the features may be grouped or associated in some way such as by biochemical pathway function or molecular mass. This sort of feature annotation, designated as $\boldsymbol{A}_{a\times m}$, is more likely to occur with chemo-centric data.

The material, or matrix, that makes up a metabolomic sample is generally biofluid (plasma, serum, urine, saliva, sweat, etc) or tissue (muscle, heart, liver, etc.) from either plant or animal specimens, although technically any organic substance is permissible with the MS. Regardless of whether the dataset contains metabolites or the ion features, statistical analysis can be performed in roughly the same manner. However, the chemo-centric allows for the identification of artifacts and contaminants, which are features related to sample collection and processing and instrument performance [28]. For example, plasticizers, which include phthalates, can easily leach into the sample through contact with the test tube or other similar storage device. As another example, sample material derived from biopsy specimens can be contaminated by gel used during sonic

Figure 1.2: Metabolomic Data Components. Set of metabolomic data comprises three matrices: sample details (D), observed feature values (Y) and feature details (A).



disruption of preserved tissue samples [35]. By identifying and removing these artifacts, a greater level of reproducibility is achieved [36].

Common univariate statistical tests used in metabolomic studies include the t-test and ANOVA. Certain studies have even used the z-score to assess individual samples in a population [37]. Such applications are an important theme for this dissertation and the z-score itself will be discussed further in Chapter 4. Due to the large number of variables, some form of adjustment should be made whenever univariate p-values are being produced, with some form of false discovery rate (FDR) method being most practical [27]. Popular multivariate methods are principle components analysis, linear discriminant analysis, and partial least squares regression among others [38].

1.6. Metabolomics Workflow

The process of obtaining metabolomic data from a sample is intensive. There are six distinct phases in the metabolomics workflow which is shown in *Figure 1.3*. The first phase, sample collection, includes all activities related to obtaining physical specimens such as hypothesis formulation, experiment design, etc. Once the samples are collected, the samples are sent to the facility where metabolite data is obtained. All actions that take place in between sample collection and data analysis can be collectively referred to as pre-analytical. Here, the pre-analytical aspects have been divided i nto four phases to impart some understanding of the complexity and influence these steps have on data analysis. The descriptions provided here are intended to give a general understanding of the purpose served by each phase. For further information on these topics several texts and articles available [39-43].

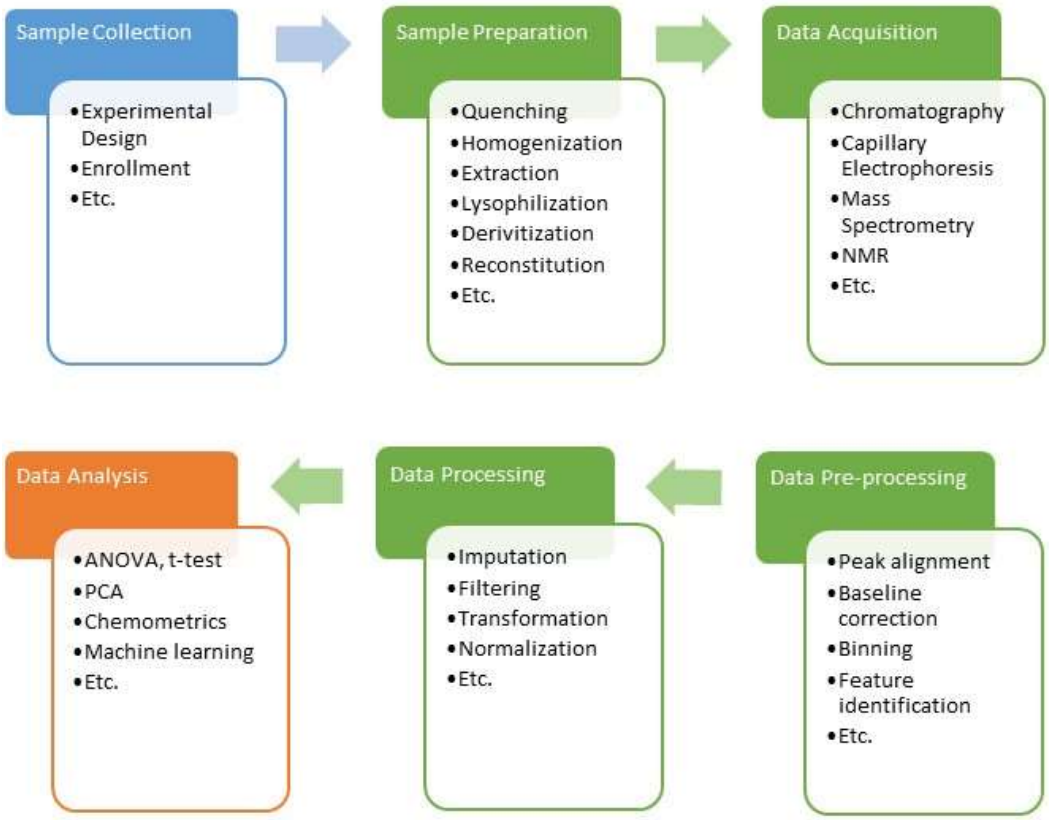


Figure 1.3: The metabolomics workflow. A metabolomics experiment begins with sample collection. Phases in green are collectively referred to as pre-analytical.

Sample preparation addresses two main goals. The first goal is to freeze or reduce metabolic activity. As intermediates of cellular function, metabolites often have rapid intracellular conversion and steps must be taken to preserve the status of the metabolome at the time the sample is obtained. This is known as “quenching”. The second goal is to lyse the cells contained in the sample to maximize recovery. As metabolites are contained within the interior of the cell, disrupting the cellular structure is helpful (or even necessary) in order to obtain measurable levels of the metabolites. This is known as “extraction”. Quotations are used because the description given here is an oversimplification. Specific workflows will vary depending on instrumentation, sample material and metabolite groups of interest. For example, derivatization is done in GC-MS but not in LC-MS or NMR. Instead these methodologies involve reconstitution.

Other actions that may take place in between sample collection and data collection include, but are not limited to, addition of chemical solvents, homogenization, centrifugation, lyophilisation, storage, etc. This makes it challenging to completely describe the sample preparation process concisely. Instead the focus here is on the primary purpose of sample preparation. For more complete details on the sample preparation process see Metabolome Analysis by Villas-Boas, Roessner, Hansen, Smedsgaard, and Nielsen [39] or Collection and Preparation of Clinical Samples by Chetwynd, Dunn and Rodriguez-Blanco [43].

After the samples have been prepared, the samples are processed through the instrumentation of choice. This produces data records associated with each sample. But before analysis can begin, this data must be compiled and organized in way that allows the spectral information to be combined across samples. This is the pre-processing phase and include peak alignment, baseline correction (NMR data), binning of features (ion-centric), or metabolite identification (chemo-centric) [38, 44-46]. The conclusion of these steps produces a data set consistent with that shown in *Figure 1.2*.

The final phase before data analysis is processing which involve manipulating certain characteristics of the data in order to maximize analysis results. Handling of missing values, which plays an important role in this dissertation, is one such step and is discussed in more detail in section 1.7.3 and is the subject of Chapter 3. Two other steps common to metabolomics are transformations and normalizations. Based on the orientation defined in *Figure 1.2*, transformations refer to operations applied to the columns (metabolites/features) while normalizations are applied to the rows (samples). A larger discussion on transformations can be found in section 2.3 and normalization is the subject of Chapter 4. Briefly, the intent of transformations is to improve some feature of the biochemicals whereas normalizations serve to

reduce variation between samples due to endogenous effects [31]. This purpose helps marks the distinction between pre-processing and processing. Pre-processing steps are necessary to make these feature values comparable between samples. In contrast, data processing can be regarded as optional, but generally useful to enhance the signal or other characteristic of the dataset for analysis. This leads to some ambiguity to the term "raw" data. The rawest forms of the data is that prior to any pre-processing steps. This includes the individual scans from the mass spectrometer, of which multiple scans are required to cover the entire chromatographic peak for a given metabolite. However, because these scans cannot be meaningfully resolved without pre-processing steps, the pre-processed data is sometimes referred to as "raw" data since this is data from which optional transformations or normalizations can be made. Plainly, this amounts to the difference between "raw scan" data versus "raw peak" data. In this dissertation the latter definition of raw is used. While pre-processing can influence the data, it is not the subject of the present document. Therefore, in this dissertation the term "raw data" implies the resulting data given whatever pre-processing steps have been selected.

Normalization seeks to remove sources of systematic variation by dividing feature values of samples according to some metric associated with this variation. In biological organisms many processes are subject to tight physiological control. Homeostasis therefore affects many matrices, including plasma and cerebral spinal fluid [47]. But some matrices are subject to a high degree of sample-to-sample variation due to endogenous characteristics. Such is the case with urine, where volume levels can vary greatly depending on water consumption, diet, exercise and other physiological condition [47] Creatinine levels have long been associated with kidney filtration and are a common normalizer for urine, although osmolality has also been used [47, 48]. In cell culture studies metabolite levels can similarly be influenced by the amount of cellular material

present [21, 49]. Cell specific normalizers are cell count, protein concentration and DNA concentration [21]. Universally available normalizers, many of which are adopted from other -omic fields, include total ion count, median ion count or housekeeping metabolites [49]. Normalization of this type is commonly used due the fact that it can be used in any matrix and is easy to calculate. However, it is susceptible to individual outliers and in cases where the experiment affects a large proportion of the variables, such normalization can actually remove the signal from the data. Many modifications have been made to total ion count for this reason, and improved normalization methods remain an active area of research [50]. Regardless of the normalizer, the underlying assumption is always the same: certain factors, pre-processing steps or biological, related to a sample cause an elevation or suppression of variable levels across the whole spectrum. The procedure is popular in proteomics where all variables are of the same class (proteins) but in metabolomics, where the variables span a number of classes and functions, the effectiveness is less clear.

1.7. Clinical Utility and Challenges

As stated previously, Inborn Errors of Metabolism (IEM) form a class of genetic diseases involving disorders of metabolism. In most cases a genetic defect leads to the coding of defective enzymes, which in turn inhibit the conversion of various substrates required for proper metabolic function. Individual symptoms and disease pathology can vary greatly depending upon the defective gene and resulting enzyme involved. Alkaptonuria, for example, is caused by a mutation in the gene responsible for creating the enzyme homogentisate oxidase (HGD) [51]. This results in a defective version of the enzyme homogentisate 1,2-dioxygenase, preventing the body from processing the amino acids phenylalanine and tyrosine [52, 53]. Buildup of the intermediate homogentisic acid occurs, the oxidated form of which is called alkpton, damaging

cartilage and heart valves which can negatively impact quality of life. Major symptoms may not present until the 3rd or 4th decade of life as it can take over 30 years for homogentisic acid to accumulate to dangerous levels in the body. Medium-chain acyl-CoA dehydrogenase (MCAD) deficiency presents in early childhood causing hypoglycemia and liver dysfunction [54]. Fatty acid beta-oxidation provides energy to the body after glucose and glycogen are depleted. A deficiency of the MCAD enzyme can quickly lead to fatty acid catabolism during periods of increased energy needs. Infants are particularly susceptible, and it is believed that prior to expanded newborn screening MCAD deficiency was responsible for a number of SIDs case [55, 56]. Comparing the symptoms and pathology of alkoptonuria and MCAD deficiency helps illustrate how different one IEM case can be from another. As a whole, IEMs can affect any organ and can occur at various stages of life. There is no single diagnostic test to identify someone as having an IEM. Instead, diagnosis of an IEM begins physicians who suspect a case ordering a targeted test covering biochemicals that best fit the observed symptoms or to rule out a possible candidate. Based on the results of this first round testing, a second round of tests will likely be ordered. And so on. This amounts to essentially a guess-and-check approach to diagnosing IEMs which may be time consuming.

Recently, global metabolomics have been used to identify a wide range of IEMs [37, 57]. In these studies, samples of human plasma and urine from subjects with known IEMs were profiled. In the majority of IEM cases, metabolites that were informative of the subjects underlying disease were significantly expressed when compared to control group, thus demonstrating a proof-of-concept for using global metabolomics as a screen for IEMs. However, several items related to the instrumentation and data processing of LC-MS metabolomics complicate the clinical utility of this technology. First is the need for a normal reference population through

which outlier metabolites can be identified. MS data is inherently semi-quantitative and are subject to run-to-run instrument variation. Yet it is infeasible to run a large set of control samples with every run of the instrument. Thus, practical application requires the ability to bridge different instrument runs together. Effectively dealing with these missing values and day-to-day instrument variation as part of the data analysis, particularly in a clinical environment, serves as the motivation for this dissertation. Second is that global MS metabolomics suffers from a high rate of missing values, which complicate patient assessment in a clinical setting. The purpose of this dissertation is to identify the optimal statistical methodology and data processing steps related to these issues when applying global MS based metabolomics in the clinical setting.

1.7.1. Relative Quantitation

Although MS is a very common analytical tool, it is not inherently quantitative. This means the ion count values returned by the mass spectrometer are not directly indicative of concentration. Instead the quantitation is relative, which is to say that the relative ion counts between samples run together can be inferred. For example, suppose the observed ion counts for creatinine in two samples analyzed in the same batch are 100,000 and 130,000. From this it can be inferred that the second sample contains approximately 30% more creatinine than the first sample, although the exact amount of either is unknown. Determining the exact amount involves elaborate methods. For instance, the gold standard method utilizes a stable isotope dilution assay which incorporates spiking known concentrations of the purified isotope associated with the compound of interest along with process controls to monitor extraction efficiency and matrix effects [58-60]. This is the process behind targeted analysis, so named because it focuses on the analytes of specific compounds (i.e. targets) and the instrumentation is typically optimized for these compounds. The challenges of targeting high throughput data is obvious. When utilizing

the global approach, all the metabolites detected may not be known up front. Even if they were, identifying associated isotopes, obtaining purified reserves and fitting the full dilution assay structure into workflow. As a result, global methods currently exist separately from targeted methods. A consequence of this is that separate runs of high-throughput MS, or "untargeted", datasets cannot be directly compared. Overcoming this limitation so that samples can be compared across multiple batches is one goal of this dissertation.

To better understand the semi-quantitative nature, it is helpful to explain how the ion counts returned by MS relate to the true concentration. The mass spectrometer is attempting to ionize the individual molecules of any given metabolite. The ion count is dependent on two things. One is the number of molecules present which depends not just on concentration but also molar mass, the number of molecules for a given mass of the compound. A smaller, lighter compound will thus have more molecules than larger, heavier compound when both are present at the same concentration. The second item influencing ion counts levels is the propensity of the molecule to stably carry a charge. This is referred to as ionization efficiency and is dependent on many things including chemical structure, meaning it will vary by biochemical, but sample handling and instrumentation. Here, instrumentation refers to the specific components, such as mobile and stationary phases, ionization method and even sample material. Some of these factors are static, i.e. molar mass, or can be controlled with choice of instrument, i.e. ionization method. But others will vary from run to run of the instrument. Certain components of the mass spectrometer, like the column/stationary phase, degrade over time and must be replaced, the exact solution of the mobile phase may vary from lot to lot, various sample preparation steps are performed manually, etc. While the instrument and sample processing can be monitored for performance in various

ways to indicate when components are needing to be replaced, these items induce variability causing the ionization efficiency to vary from run to run.

Rocke and Lorenzato presented an analytical chemistry model with two sources of error to more accurately estimate instrument error [61]. The argument for two components is based on empirical experience in which low level concentrations tend to have a near constant error rate but at higher levels the error tends to increase as the concentration does. This phenomenon was answered by two error components: a linear component and a component that is multiplicative of the concentration. This premise forms the basis of model 1.1. Letting y_{ji} be the observed ion count and x_{ji} be the true concentration of metabolite j in subject i, the model used here is as follows:

$$y_{ji} = \alpha + c_i x_{ji} e^{\eta_{ji}} + \varepsilon_{ji} \qquad \text{model 1.1}$$

with $\eta_{ji} \sim N(0, \sigma_\eta^2)$, $\varepsilon_{ji} \sim N(0, \sigma_\varepsilon^2)$. Both errors are independent and identically distributed random variables with ε_{ji} being the linear error component and η_{ji} representing the proportional error term. The intercept α represents the ion count returned when no amount of the metabolite is present and is analogous to the background level of the instrument. The coefficient c_i represents, generally, the overall ionization efficiency of the system for the biochemical in question. This model accounts for the observed behavior in analytical methods since at low levels ε_{ji} tends to dominate the observed error as smaller values of x_{ji} results in less contribution from η_{ji}. However, as x_{ji} increases the greater the impact of η_{ji} causing this term to dominate at high concentrations. This is also compatible with the relative assumption between samples as the expected value of between any two samples i and i' for a given feature is

$$E[y_{ji}] = \alpha + c_i x_{ji}$$

$$E[y_{ji'}] = \alpha + c_i x_{i'j}$$

α is mostly a nuisance parameter and many pre-processing steps involve removing counts that are not at least 3x to 5x above the baseline level. Therefore, this term may often be assumed as small enough to be ignored. Ignoring the intercept term gives the expected ratio of sample i to sample i' as $x_{ji}/x_{ji'}$ which is consistent with relative quantitation.

While c_i is certain to vary from feature to feature, it is unknown if α, σ_η or σ_ε do. Background and error can theoretically vary from feature to feature, though careful monitoring of instrument performance should indicate when this occurs and trigger remedial steps. After all, it is not uncommon for background levels to increase over time in MS instruments. This would imply, since features are processed over time within a sample and samples are then processed sequentially through the instrument, the background could vary not only by feature but by sample as well. This quickly leads to a model with more unknowns than observations, requiring additional assumptions to solve. Since instrument drift often tends to be linear, assuming linearity and examining the levels of housekeeping features, spiked in during sample preparation, in quality control samples interspersed through the instrument run provides an easy way to monitor for such drift. Significant inflation of either error component would lead to large increases in the variation of housekeeping markers, which would be grounds for re-analysis. Hence it is reasonable to treat the background and error components as fixed.

Lastly, to extend the model to multiple instrument runs, all model components related to instrument performance should be adjusted accordingly. As a result, considering, hypothetically, the same set of samples run over multiple days gives

$$y_{jik} = \alpha_k + c_{jk} x_{ji} e^{\eta_{jik}} + \varepsilon_{jik} \qquad \text{model 1.2}$$

where $k \in \{1, \ldots.\}$ indexes the instrument run. $\eta_{jik} \sim N(0, \sigma^2_{\eta k})$, and $\varepsilon_{jik} \sim N\left(0, \sigma^2_{\varepsilon k}\right)$. Obviously, from run to run the concentration of the sample does not change. The slope and ionization factor can change though, as can the distribution of the errors and background. Examining the expected value of the same sample from one run to another gives:

$$E\left[y_{jik}\right] = \alpha_k + c_{jk} x_{ji} * e^{\sigma^2_{\eta k}/2}$$

$$E\left[y_{jik'}\right] = \alpha_{k'} + c_{jk'} x_{ji} * e^{\sigma^2_{\eta k'}/2}.$$

Again, ignoring the intercept terms leads to, on average, a proportional relationship between the two runs with the second run differing by a factor of $(c_{jk}\, e^{\sigma^2_{\eta k}/2})/(c_{jk'} e^{\sigma^2_{\eta k'}/2})$ compared to the first run. Combination of multiple instrument runs can be accomplished by estimation of this ratio. This is the subject of Chapter 4. The effect of the varying background from run to run may result in some features, particularly those of low level, being lost or having more missing values (see section 2.1.2.) in certain runs as their detectability fails to rise sufficiently above the background level. Different error variances may result in certain runs being more precise than others; however, if all runs are deemed to be acceptable then combined error should also be acceptable provided the ratio $(c_{jk}\, e^{\sigma^2_{\eta k}/2})/(c_{jk'} e^{\sigma^2_{\eta k'}/2})$ is adequately accounted for.

1.7.2. Missing Values

Not all of the y_{ji}'s will have an observed ion count. Compared to other omics fields the rate of missing values in global MS metabolomics data is quite high [33, 62-64]. Theoretically, missing values can be caused by a number of technical issues, such as ion suppression [65], but the technology used is very similar to that used in proteomics where the missing rate is much lower [66]. The major difference is that metabolite species are of a much lower molecular mass and

abundance (average ion count) than proteins, and metabolomics workflow is optimized to these molecules. In a well-functioning instrument in which technical issues are minimal, missing values can largely be attributed to the true value falling below the background level of the instrument [62, 64]. In extreme cases the metabolite may be completely missing from the sample, a situation most likely to hold for drug metabolites or other xenobiotics. This detection limit depends on certain physical and chromatographic properties of each chemical compound meaning that it will vary from compound to compound.

Missing values can technically occur for other reason [67, 68]. If the analytes of two compounds elute close together with one peak being relatively large peak and the other being relatively small, the dominate peak may prevent the lesser peak from being recorded. In this case the missing value is not due to the abundance of a neighboring compound and un-related to the level of the lost compound. Another possibility is that a peak is missed because it occurs in-between scan intervals or because a certain scan records poorly, in which case missingness is also unrelated to the true intensity level (this can be avoided by increasing the scan rate, but doing so may negatively impact other areas of instrument performance). A literature survey returned no studies on the proportion each of these contribute to the overall missing rate. Given the wide range technological methods used in the field this is likely to depend the analytical approach and type of instrumentation. But it is increasingly of interest to treat missing values in MS as limit of detection [69], and in a well-functioning MS system this would theoretically be the dominant reason for such occurrences.

Due to the prevalence of missing values, addressing them forms an important step in the analysis workflow. From the experience of the authors, in large chemocentric MS analysis almost every sample is guaranteed to experience a missing value in at least one feature and

roughly half of the features will experience some amount of missing data. The spirit of high throughput metabolomics is the ability to identify a large quantity of metabolite features at once. Imagine in a study comparing healthy subjects to some disease in which a feature is found only to be present in the disease subjects. Such a feature would hold enormous research value, not just for disease identification but also potential treatment. So, in one sense low filled compounds can be the most alluring to researchers. Any step which reduces the number of identified features chips away at the potential and attraction of global metabolomics.

However, data quality is also important to obtaining confident, reproducible results. Filtering, which removes features when the proportion of missing values is above a certain percentage, is a compromise that mitigates the impacts of missing values and is gaining some popularity. Removing the most sparsely populated compounds improves reproducibility, but still runs the risk of removing potential markers from being discovered. And because global metabolomics operates on the front line of biomarker discovery, markers that are missed at this stage may never be recovered later on. Whether filtering is applied liberally or conservatively, missing values persist to some degree and must be dealt with somehow during statistical analysis. With global metabolomics functioning mostly as a biomarker screening tool, researchers often desire a complete dataset from which to work and try various approaches. Imputation, which involves inserting a value in place of the missing observations, is therefore common. For example, average imputation replaces the missing values in a given feature with the mean of the observed samples in that feature. Similarly, minimum imputation, which is analogous to LOD, replaces missing values with the observed minimum. Numerous other types of imputations methods have been used in metabolomics, with none being found to be universally “best”. In fact, the optimal approach is likely specific to goals and study design [62-64]. Here it is important to note that the

clinical setting differs in several critical ways from the usual biomarker discovery setting. First, in contrast to the group-based designs of biomarker discovery, clinical analysis involves comparing one patient against a control population or set of reference values derived from such a control population. Second, interest is only in the patient sample which means that imputation of control population to achieve a complete data set is not strictly necessary. All that is needed are accurate measures of relevant parameters that are derived from the control population. This is a novel issue for high through put metabolomics and is the subject of Chapter 5.

1.7.3. Metabolite Distribution

As a maturing field of study, there exists several fundamental unknowns in metabolomic data. Among these include the distribution of metabolites themselves. Under the proposed ion intensity model 1.1, y_{ji} has error components that are both normal and log-normal. In the previous section it was discussed how these components influence the distribution for a fixed level abundance level. But, the overall distribution of y_{ji} will also depend on the distribution of x_{ji}, the true concentration of the population. Since metabolomics studies often involve factorial designs leading to t-tests and ANOVA while simultaneously employing a low number of samples, normality is a convenient assumption. There is biological reason to believe metabolites are normally distributed. Many biological matrices, including blood/plasma, are subject to homeostasis. Tight process regulation could easily lead to a normal behavior. Clearly though, the ion intensity model and log-normal behavior of η makes the validity of such an assumption questionable. If x_{ji} is normal than the product $x_{ji}e^{\eta_{ji}}$ is not strictly normal or log-normal, though may still be close to normal if σ_η^2 is small relative to the population variance σ_i^2. Some examination of normality has been done [27], though it has primarily used ion-centric data which contains a high proportion of artifacts [30]. Since these features are closely tied to the

performance of the instrument, it is likely that most influential error component is σ_ε^2 which indicates a normally distributed variable. It is the experience of this author that in chemo-centric data metabolites are most accurately described as log-normal. There is some existing evidence to support the use of log transformation with metabolites [70, 71]. This could be explained partially by the property of the log-normal distribution that holds the product of two log-normal distributions, which would the case for $x_{ji}e^{\eta_{ji}}$ if x_{ji} is log-normal, is another log-normal with variance equal to the sum of the individual variances. This combined variance would help to diminish the relative contribution of σ_ε^2. By establishing the distribution of metabolites, the door is opened for parametric approaches to handling missing values and is hence a natural place for the papers in this dissertation to begin.

REFERENCES

[1] L. M. Samuelsson and D. G. Larsson, "Contributions from metabolomics to fish research," *Mol Biosyst,* vol. 4, no. 10, pp. 974-9, Oct 2008.

[2] U. Roessner *et al.*, "Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems," *Plant Cell,* vol. 13, no. 1, pp. 11-29, Jan 2001.

[3] P. Puri *et al.*, "A lipidomic analysis of nonalcoholic fatty liver disease," *Hepatology,* vol. 46, no. 4, pp. 1081-90, Oct 2007.

[4] O. Fiehn, "Metabolomics--the link between genotypes and phenotypes," *Plant Mol Biol,* vol. 48, no. 1-2, pp. 155-71, Jan 2002.

[5] W. Weckwerth, "Metabolomics in systems biology," *Annu Rev Plant Biol,* vol. 54, pp. 669-89, 2003.

[6] K. Beebe and A. D. Kennedy, "Sharpening Precision Medicine by a Thorough Interrogation of Metabolic Individuality," *Comput Struct Biotechnol J,* vol. 14, pp. 97-105, 2016.

[7] N. Koen, I. Du Preez, and d. T. Loots, "Metabolomics and Personalized Medicine," *Adv Protein Chem Struct Biol,* vol. 102, pp. 53-78, 2016.

[8] H. G. Gika, G. A. Theodoridis, R. S. Plumb, and I. D. Wilson, "Current practice of liquid chromatography-mass spectrometry in metabolomics and metabonomics," *J Pharm Biomed Anal,* vol. 87, pp. 12-25, Jan 2014.

[9] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrom Rev,* vol. 26, no. 1, pp. 51-78, 2007 Jan-Feb 2007.

[10] A. D. McNaught and A. Wilkinson, "tandem mass spectrometer," IUPAC. Compendium of Chemical Terminology (the "Gold Book")2nd ed.: Blackwell Scientific Publications, 1997. [Online]. Available.

[11] M. Y. Fong, J. McDunn, and S. S. Kakar, "Identification of metabolites in the normal ovary and their transformation in primary and metastatic ovarian cancer," *PLoS One,* vol. 6, no. 5, p. e19963, 2011.

[12] A. T. JAMES and A. J. MARTIN, "Gas-liquid partition chromatography; the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid," *Biochem J,* vol. 50, no. 5, pp. 679-90, Mar 1952.

[13] J. J. Pitt, "Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry," *Clin Biochem Rev,* vol. 30, no. 1, pp. 19-34, Feb 2009.

[14] T. Soga, "Capillary electrophoresis-mass spectrometry for metabolomics," *Methods Mol Biol,* vol. 358, pp. 129-37, 2007.

[15] R. Ramautar, G. W. Somsen, and G. J. de Jong, "CE-MS in metabolomics," *Electrophoresis,* vol. 30, no. 1, pp. 276-91, Jan 2009.

[16] R. Ramautar, G. W. Somsen, and G. J. de Jong, "CE-MS for metabolomics: developments and applications in the period 2012-2014," *Electrophoresis,* vol. 36, no. 1, pp. 212-24, Jan 2015.

[17] H. J. Issaq, E. Abbott, and T. D. Veenstra, "Utility of separation science in metabolomic studies," *J Sep Sci,* vol. 31, no. 11, pp. 1936-47, Jun 2008.

[18] S. K. Davies, J. G. Bundy, and A. M. Leroi, "Metabolic youth in middle age - predicting ageing in Caenorhabditis elegans using metabolomics," *J Proteome Res,* Sep 2015.

[19] M. Austdal *et al.*, "First Trimester Urine and Serum Metabolomics for Prediction of Preeclampsia and Gestational Hypertension: A Prospective Screening Study," *Int J Mol Sci,* vol. 16, no. 9, pp. 21520-38, 2015.

[20] Z. Pan and D. Raftery, "Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics," *Anal Bioanal Chem,* vol. 387, no. 2, pp. 525-7, Jan 2007.

[21] L. P. Silva, P. L. Lorenzi, P. Purwaha, V. Yong, D. H. Hawke, and J. N. Weinstein, "Measurement of DNA concentration as a normalization strategy for metabolomic data from adherent cell lines," *Anal Chem,* vol. 85, no. 20, pp. 9536-42, Oct 2013.

[22] C. K. Larive, G. A. Barding, and M. M. Dinges, "NMR spectroscopy for metabolomics and metabolic profiling," *Anal Chem,* vol. 87, no. 1, pp. 133-46, Jan 2015.

[23] R. Verpoorte, Y. H. Choi, N. R. Mustafa, and H. K. Kim, "Metabolomics: back to basics," *Phytochemistry Reviews,* vol. 7, no. 3, pp. 525-537, 2008/10/01 2008.

[24] J. Clark. (2009). *Chemguide: helping you to understand chemistry*. Available: http://www.chemguide.co.uk/

[25] G. A. Gowda and D. Djukovic, "Overview of mass spectrometry-based metabolomics: opportunities and challenges," *Methods Mol Biol,* vol. 1198, pp. 3-12, 2014.

[26] K. Vékey, "Mass spectrometry and mass-selective detection in chromatography," *J Chromatogr A,* vol. 921, no. 2, pp. 227-36, Jul 2001.

[27] M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart, and O. Yanes, "A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data," *Metabolites,* vol. 2, no. 4, pp. 775-95, 2012.

[28] A. M. Evans, M. W. Mitchell, H. Dai, and C. D. DeHaven, "Categorizing Ion-Features in Liquid Chromatography/Mass Spectrometry Metabolomics Data," *Journal of Metabolomics,* vol. 2, no. 3, 2012.

[29] N. G. Mahieu and G. J. Patti, "Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites," *Anal Chem,* vol. 89, no. 19, pp. 10397-10406, Oct 2017.

[30] B. P. Bowen and T. R. Northen, "Dealing with the unknown: metabolomics and metabolite atlases," *J Am Soc Mass Spectrom,* vol. 21, no. 9, pp. 1471-6, Sep 2010.

[31] M. M. W. B. Hendriks, F. A. v. Eeuwijk, R. H. Jellema, and J. A. Westerhuis, "Data-processing strategies for metabolomics studies," *TrAC, Trends in analytical chemistry (Regular ed.),* vol. 30, no. 10, pp. 1685-1698, 2011.

[32] R. Tautenhahn, K. Cho, W. Uritboonthai, Z. Zhu, G. J. Patti, and G. Siuzdak, "An accelerated workflow for untargeted metabolomics using the METLIN database," *Nat Biotechnol,* vol. 30, no. 9, pp. 826-8, Sep 2012.

[33] J. Xia, N. Psychogios, N. Young, and D. S. Wishart, "MetaboAnalyst: a web server for metabolomic data analysis and interpretation," *Nucleic Acids Res,* vol. 37, no. Web Server issue, pp. W652-60, Jul 2009.

[34] J. M. Halket, D. Waterman, A. M. Przyborowska, R. K. Patel, P. D. Fraser, and P. M. Bramley, "Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS," *J Exp Bot,* vol. 56, no. 410, pp. 219-43, Jan 2005.

[35] B. J. Trock, "Application of metabolomics to prostate cancer," *Urol Oncol,* vol. 29, no. 5, pp. 572-81, 2011 Sep-Oct 2011.

[36] J. M. Perkel, "Metabolomics: Sifting through complex samples," ed: Biocompare, 2013.

[37] M. J. Miller *et al.*, "Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism," *J Inherit Metab Dis,* Apr 2015.

[38] K. H. Liland, "Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis," *TrAC, Trends in analytical chemistry (Regular ed.),* vol. 30, no. 6, pp. 827-841, 2011.

[39] S. G. Villas-Boas, U. Roessner, M. A. E. Hansen, J. Smedsgaard, and J. Nielsen, *Metabolome Analysis An Introduction* (Wiley Interscience Series In Mass Spectrometry). Hoboken, NJ: John Wiley & Sons, Inc, 2007.

[40] S. P. Putri and E. Fukusaki, *Mass spectrometry-based metabolomics : a practical guide*. Boca Raton: CRC Press, Taylor & Francis Group, 2015, pp. xiv, 280 pages.

[41] M. Lämmerhofer and W. Weckwerth, *Metabolomics in practice : successful strategies to generate and analyze metabolic data*. Weinheim: Wiley-VCH, 2013, pp. xxv, 415 pages.

[42] M. Cuperlović-Culf, D. A. Barnett, A. S. Culf, and I. Chute, "Cell culture metabolomics: applications and future directions," *Drug Discov Today,* vol. 15, no. 15-16, pp. 610-21, Aug 2010.

[43] A. J. Chetwynd, W. B. Dunn, and G. Rodriguez-Blanco, "Collection and Preparation of Clinical Samples for Metabolomics," *Adv Exp Med Biol,* vol. 965, pp. 19-44, 2017.

[44] Y. Xi and D. M. Rocke, "Baseline correction for NMR spectroscopic metabolomics data analysis," *BMC Bioinformatics,* vol. 9, p. 324, Jul 2008.

[45] M. Katajamaa and M. Oresic, "Data processing for mass spectrometry-based metabolomics," *J Chromatogr A,* vol. 1158, no. 1-2, pp. 318-28, Jul 2007.

[46] A. C. Sauve and T. P. Speed, "Normalization, baseline correction and alignment of high-throughput mass spectrometry data," in *Genomic Signal Processing and Statistics*, 2004.

[47] B. M. Warrack *et al.*, "Normalization strategies for metabonomic analysis of urine samples," *J Chromatogr B Analyt Technol Biomed Life Sci,* vol. 877, no. 5-6, pp. 547-52, Feb 2009.

[48] S. Ganti and R. H. Weiss, "Urine metabolomics for kidney cancer detection and biomarker discovery," *Urol Oncol,* vol. 29, no. 5, pp. 551-7, 2011 Sep-Oct 2011.

[49] B. Cao *et al.*, "GC-TOFMS analysis of metabolites in adherent MDCK cells and a novel strategy for identifying intracellular metabolic markers for use as cell amount indicators in data normalization," *Anal Bioanal Chem,* vol. 400, no. 9, pp. 2983-93, Jul 2011.

[50] D. Ryan, K. Robards, P. D. Prenzler, and M. Kendall, "Recent and potential developments in the analysis of urine: a review," *Anal Chim Acta,* vol. 684, no. 1-2, pp. 8-20, Jan 2011.

[51] G. H. Reference. (2018). *HGD gene*. Available: https://ghr.nlm.nih.gov/gene/HGD

[52] N. C. f. A. T. Sciences, "Alkaptonuria," ed, 2016.

[53] J. Barwell and E. Boskey. (2016). *Alkaptonuria*. Available: https://www.healthline.com/health/alkaptonuria

[54] N. O. f. R. Disorders. (2005). *Medium Chain Acyl CoA Dehydrogenase Deficiency*. Available: https://rarediseases.org/rare-diseases/medium-chain-acyl-coa-dehydrogenase-deficiency/

[55] T. Hegyi, B. Ostfeld, and K. Gardner, "Medium chain acyl-coenzyme A dehydrogenase deficiency and SIDS," *N J Med,* vol. 89, no. 5, pp. 385-92, May 1992.

[56] L. D. Keppen and B. Randall, "Inborn defects of fatty acid oxidation: a preventable cause of SIDS," *S D J Med,* vol. 52, no. 6, pp. 187-8; disscussion 188-9, Jun 1999.

[57] A. Kennedy *et al.*, "Utilizing Metabolomics of Human Urine to Screen for Multiple Inborn Errors of Metabolism," *Genetic Testing and Molecular Biomarkers,* vol. Under Review, 2016.

[58] E. Varga *et al.*, "Stable isotope dilution assay for the accurate determination of mycotoxins in maize by UHPLC-MS/MS," *Anal Bioanal Chem,* vol. 402, no. 9, pp. 2675-86, Mar 2012.

[59] M. Granvogl, P. Koehler, L. Latzer, and P. Schieberle, "Development of a stable isotope dilution assay for the quantitation of glycidamide and its application to foods and model systems," *J Agric Food Chem,* vol. 56, no. 15, pp. 6087-92, Aug 2008.

[60] M. Milton and R. Wielgosz, "Uncertainty in SI-traceable measurements of amount of substance by isotope dilution mass spectrometry," *Metrologia,* vol. 37, no. 3, p. 199, 2000.

[61] D. M. Rocke and S. Lorenzato, "A Two-Component Model for Measurement Error in Analytical Chemistry," *Technometrics,* vol. 37, no. 2, pp. 176-184, 1995.

[62] P. S. Gromski *et al.*, "Influence of missing values substitutes on multivariate analysis of metabolomics data," *Metabolites,* vol. 4, no. 2, pp. 433-52, 2014.

[63] E. G. Armitage, J. Godzien, V. Alonso-Herranz, Á. López-Gonzálvez, and C. Barbas, "Missing value imputation strategies for metabolomics data," *Electrophoresis,* vol. 36, no. 24, pp. 3050-60, Dec 2015.

[64] O. Hrydziuszko and M. R. Viant, "Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline," *Metabolomics,* vol. 8, no. 1, pp. 161-174, 2012.

[65] T. G. Payne, A. D. Southam, T. N. Arvanitis, and M. R. Viant, "A signal filtering method for improved quantification and noise discrimination in fourier transform ion cyclotron resonance mass spectrometry-based metabolomics data," *J Am Soc Mass Spectrom,* vol. 20, no. 6, pp. 1087-95, Jun 2009.

[66] D. Albrecht, O. Kniemeyer, A. A. Brakhage, and R. Guthke, "Missing values in gel-based proteomics," *Proteomics,* vol. 10, no. 6, pp. 1202-11, Mar 2010.

[67] E. G. Armitage and C. Barbas, "Metabolomics in cancer biomarker discovery: current trends and future perspectives," *J Pharm Biomed Anal,* vol. 87, pp. 1-11, Jan 2014.

[68] E. Mattarucchi and C. Guillou, "Comment on "Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery"," *Anal Chem,* vol. 83, no. 24, pp. 9719-20; discussion 9720-1, Dec 2011.

[69] 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 Suppl 16, p. S5, 2012.

[70] P. Manini, G. De Palma, R. Andreoli, M. Goldoni, and A. Mutti, "Determination of urinary styrene metabolites in the general Italian population by liquid chromatography-tandem mass spectrometry," *Int Arch Occup Environ Health,* vol. 77, no. 6, pp. 433-6, Aug 2004.

[71] T. H. Herdt, J. B. Stevens, W. G. Olson, and V. Larson, "Blood concentrations of beta hydroxybutyrate in clinically normal Holstein-Friesian herds and in those with a high prevalence of clinical ketosis," *Am J Vet Res,* vol. 42, no. 3, pp. 503-6, Mar 1981.

CHAPTER 2: DISTRIBUTION AND CORRELATION

2.1. Overview

Global metabolomics has been developed as a screening platform. This primary function can be attributed to cost and difficulty in obtaining sample material, which in turn leads to metabolomics having experimental designs with small sample sizes [1]. Another characteristic of metabolomic sets are the frequent use of factor designs such as water-stress conditions in plants [2] or disease staging [3]. This approach fits well with the semi-quantitative nature of MS where only relative difference between samples are meaningful. One cannot use global MS to answer how many days of drought it takes phenols or flavonoids to fall below a certain mg/dl benchmark, for example. This would be more appropriate for targeted analysis [4].

However, it is possible through global metabolomics to examine what biochemicals have significantly higher or lower levels in water stressed plants compared to non-stressed plants. The low sample size, semi-quantitative nature and group-based comparisons combine to position global metabolomics primarily as a tool for identifying the most promising candidates from hundreds, or even thousands of features. This group-based approach to biomarker identification has naturally led to wide spread use of two sample t-test and ANOVA, which have become the go-to statistical procedures for metabolomics. The hypothesis tests here involve the means of the groups, which for the simple two group setting is

$$H_{0i}: \mu_{1i} = \mu_{2i}$$

$$H_{Ai}: \mu_{1i} \neq \mu_{2i}$$

For $i = 1, \ldots, m$. The features can then be ranked by p-value. However, an underlying

assumption to these univariate methods is a normal distribution. Normality is also central to certain multivariate procedures including MANOVA and linear discriminant analysis. Yet, biochemicals themselves are not biologically constrained to a Gaussian behavior [5, 6]. Generally, this concern would be avoided by using non-parametric equivalents, such as Wilcoxon rank sum or Kruskal Wallis, but the lower power of these tests coupled with the low sample size make this option undesirable or even impractical. Focusing for a minute on the Wilcoxon test, the hypothesis tested is

$$H_0: \textit{Distribution of Group 1} = \textit{Distribution of Group 2}$$

$$H_A: \textit{Distribution of Group 1} \neq \textit{Distribution of Group 2}$$

In metabolomics though, it is possible to have as few as three observational units per group [7]. A comparison of two such groups using the Wilcoxon test gives the lowest possible obtainable p-value as 0.1, and this is before any adjustments for multiple comparisons are made. Hence parametric hypothesis testing is preferred, and normality is, therefore, a desirable trait. Yet little research has been devoted to the distributional behavior of metabolites. Vinaixa *et al*. [1] have previously assessed normality in ion-centric datasets, finding the majority of the features from four separate data sets "pass" for normal. Three different types of sample material were covered by these sets: two retina, one serum and one cell culture. Given the differences between ion-centric and chemo-centric data (Section 1.9) [8], it is natural to extend this discussion to the chemo-centric setting. This chapter does so using LC-MS metabolomic data with a variety of assessment tools and transformations. The results suggest that metabolites can be reasonably regarded as log-normal, which is to say that normality is achieved following a natural log transformation.

Another item of interest is the dependence or association between metabolites. While univariate approaches have historically dominated in metabolomics, certain multivariate procedures, most notably principal component analysis, are also used widely in metabolomics. Correlation plays an extremely important in multivariate statistics. For example, variable correlation can have significant impact on the power and Type I error rate of MANOVA [9], correlation plays a fundamental role when choosing between LASSO and Elastic Net [10], and many network analyses, including Gaussian graphical models, rely on correlation or partial correlation. Literature on correlation analysis has been small and inconsistent in metabolomics. Vinaixa *et al.* [1] commented that LC-MS metabolomics is "multi-correlated" to the point that certain features border on being collinear, whereas Camacho *et al.* [11] observed "the large majority of metabolite pairs showed little or no correlation". Clearly this is an issue for which greater understanding is beneficial.

This chapter is devoted to examining the distribution of metabolites, specifically focusing on normality, and correlation. Three separate cohorts comprising samples of human plasma, urine and cerebral spinal fluid (CSF) were profiled using LC-MS based chemocentric metabolomics. Each set is composed of between 30 and 40 healthy individuals creating reasonably large sets of single populations with which to explore metabolite properties. All three sets were indexed against a propriety library allowing for biochemical identification and arrangement by biochemical pathway [8, 12]. This level of annotation allows for further assessment within and between the different types of datasets.

2.2. Assessing Normality

Normality is one of the richest subjects in statistics with the literature spanning from the early 1900s to the present day [13, 14]. Summary measures can be used to assess normality.

Most of these metrics revolve around the concepts of skewness and kurtosis. Skewness attempts to measure the amount of asymmetry while kurtosis relates to the amount of overall "tailedness", "shoulderdness", or "peakedness" of a distribution [15, 16]. Skewness is most commonly expressed numerically as the third standardized moment, which for Y_i a random metabolite feature, is defined as:

$$\gamma_i = E\left[\frac{\left(Y_i - \mu_i\right)^3}{\sigma_i^3}\right]$$

where μ_i and σ_i are respectively the mean and variance of $\boldsymbol{y_i}$. Similarly, kurtosis is the fourth standardized moment:

$$v_i = E\left[\frac{\left(Y_i - \mu_i\right)^4}{\sigma_i^4}\right]$$

Other definitions of these statistics exist, such as skewness based on the standardized difference between the mean and median [17, 18], but all are geared toward the same fundamental concepts of skewness and kurtosis. A recent summary of skewness and kurtosis measures is given in Cain *et al.* [19] though both are usually covered in most elementary statistics books. The definitions above are specifically shown because they are prominently used in formal hypothesis tests and considered the most recognizable.

The range of γ_i is the entire real line. Values of $\gamma_i < 0$ indicate a distribution in which the left tail is longer or heavier while $\gamma_i > 0$ indicates the right tail is so. Distributions, which include the normal, in which $\gamma_i = 0$ are said to be non-skewed. For kurtosis, v_i is strictly greater than 1. For any normal distribution $v_i = 3$, and as a result the term *excess kurtosis* subtracts 3 from the usual kurtosis value in order to index against the normal. A value of $v_i < 3$, also known as leptokurtic, are taken to imply longer tails and a thinner central peak than a normal. Values of $v_i > 3$, or

platykurtic, indicate shorter tails and wider shoulders than a normal.

For a given sample, inference about the normality of the population can be made from the corresponding sample statistics $\hat{\gamma}_i$ and $\hat{\nu}_i$, which are found by substituting the sample mean and standard deviation in place of their population level parameters and averaging over the sample. Interpretation is subjective, but a general rule of thumb is $-1 < \hat{\gamma}_i < 1$ and $2 < \hat{\nu}_i < 4$. However, statistics meant to address skewness or kurtosis are mathematical constructs and need not correspond exactly to any philosophical interpretation [20, 21]. For example, γ_i does not adhere to any strict rule or direction when one tail is very long and the other is very heavy. It has been shown that statistical tests based on these measures can be misleading [22]. As a result, skewness and kurtosis are not often used without other diagnostics.

Graphical approaches to assessing normality include simple stem and leaf plot, histogram and the quantile-quantile plot. This plot compares the observed quantiles of the sample against their expected value under a normal distribution. The main attraction to graphical approaches is the visual format and, like summary statistics, the ability to indicate where deviation from normality is occurring. However, interpretation can be subjective and is challenging to implement in omic datasets containing hundreds to thousands of features.

In contrast to summary and graphical methods, hypothesis testing provides a concrete framework for determining if a sample is normal that can be applied easily to large datasets, though the informational value behind a test statistic and p-value is limited beyond a simple decision tool. All normality tests surveyed in this review utilize the same null hypothesis, namely for $\boldsymbol{y_i} = \{y_{1i},\ldots,y_{ni}\}'$

$$H_{0i} : F(\boldsymbol{y_i}) = \varphi\left(\frac{\boldsymbol{y_i} - \mu_i}{\sigma_i}\right)$$

where $\varphi(t)$ is the cumulative distribution function of the standard normal distribution. Some

measure of discrepancy is calculated between the observed sample and what would be expected under H_{0i}. A consequence leads to a well-known dilemma with normality testing: major departures from normality may not trigger rejection if the sample size is small while minor departures may be deemed "significant" if the sample size is very large [23, 24]. But in terms of formal decision making, normality tests provide the only option and for that reason are quite widely used when determining normality.

Normality tests can be broadly classified depending on the mechanism used to detect departure from normality. Here we choose a classification scheme similar to Baringhaus *et al.* [25] which separates tests based on (1) the empirical distribution function (EDF), (2) correlation, (3) moment, (4) empirical characteristic function (ECF) and (5) miscellaneous approaches. In the descriptions that follow let $\boldsymbol{X_n}$ be a random vector of size n with x_j the j^{th} element of $\boldsymbol{X_n}$, $x_{(j)}$ the j^{th} ordered element of $\boldsymbol{X_n}$, $\bar{x} = n^{-1} \sum_{j=1}^{1} x_j$ the sample mean and $s = \sqrt{(n\text{-}1)^{-1} \sum_{j=1}^{n} (x_j - \bar{x})^2}$ the sample standard deviation.

2.2.1. EDF Tests

As the name suggests these tests make use of the empirical distribution function and includes those from Kolmogorov-Smirnoff, Cramer-von Misses and Anderson-Darling. Using the notation above while letting $V_j = \varphi\left(x_{(j)} - \bar{x}/s\right)$ and defining:

$$\widehat{F}_n(t)=\frac{1}{n}\sum_{j=1}^{n} I(V_j \le t)\,,\ 0 \le t \le 1$$

each of the three tests can be written using the function:

$$R_n(t) = \sqrt{n}\,(\widehat{F}_n(t) - t)$$

For the tests above, we have:

Cramer – von Mises: $$\int_0^1 R_n^2(t)\,dt$$

Kolmogorov – Smirnoff: $$\sup_t |R_n(t)|$$

Anderson – Darling: $$\int_0^1 R_n^2(t) \Big/ t\,(t\text{-}1)\,dt$$

Cramer-von Mises (CM) is the oldest of these test with Cramer introducing his version of the test statistic in 1928 [26]. Von Mises, working independently, produced a very similar test statistic in 1931 [27]. Being so similar in content and introduced so closely together in time, the test has come to share the names of both individuals. Shortly after introduction of the CM test, Kolmogorov introduced a different test statistic in 1933 with important contributions made in subsequent years by Smirnoff regarding the distribution of the variables involved [28-31]. The result came to be known as the Kolmogorov-Smirnoff test (KS) and has become a stalwart of normality testing. Use of this test continues into the present day and it remains a fixture in many introductory statistics books despite arguments that it should not be used as a normality test due its poor performance [32, 33]. In its defense, KS can be used to test that two samples are derived from the same distribution and it is able to handle ties, which many of the more statistically powerful tests do not tolerate well. Worth noting is that as a normality test KS specifically tests that the data is Standard Normal and thus use of this test requires centering and standardizing the sample. Lilliefors went on to extend KS to test for normality without specifying the mean or variance [34]. Anderson and Darling (AD) introduced their test in 1952, which gives more weight to the tails of the distribution than KS does [35]. More recently, Vasicek provides another EDF test using a test statistic based on the sample entropy of the density function [36]. Of the EDF tests mentioned here, AD is generally regarded as the most powerful and therefore the most advisable [37, 38].

2.2.2. Correlation Tests

These tests compare the observed ordered statistics from the sample against their expected value under a standard normal distribution. Letting $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}$ be the ordered statistics of n independent draws from a standard normal. Under H_0 the ordered sample statistics can be modeled as a linear function of the $z_{(j)}$'s such that

$$x_{(j)} = \mu + \sigma m_j + \varepsilon_j$$

where $m_j = E[\, z_{(j)} \,]$ and the ε_j's represent random error satisfying $E[\varepsilon_j]=0$ and $cov(\varepsilon') = \sigma^2 \dot{\Sigma}$ for all $j \in \{1, 2, \ldots, n\}$ with $\dot{\Sigma}$ being the covariance matrix of $\boldsymbol{z} = (z_{(1)}, \ldots, z_{(n)})'$. When the sample is derived from a normal distribution the ordered sample vector $\boldsymbol{x} = \{z_{(1)}, \ldots, z_{(n)}\}$ will be closely linear with the vector of ordered standard normal expectations, $\boldsymbol{m} = \{m_{(1)}, \ldots, m_{(n)}\}$. Therefore, the square of the correlation coefficient ρ^2 provides a measure of assessing normality. In practice calculating m_j is not trivial and the choice of what to use for $\boldsymbol{m}$ leads to the various tests in this category. For **w** the chosen representative of $\boldsymbol{m}$ we have ρ^2 as

$$\rho^2(\boldsymbol{w}, \boldsymbol{x}) = \left(\frac{\sum_{j=1}^{n} (w_j - \overline{w})(x_{(j)} - \overline{x})}{\sqrt{\sum_{j=1}^{n} (w_j - \overline{w})^2 \sum_{j=1}^{n} (x_{(j)} - \overline{x})^2}} \right)^2$$

In terms of **w** better known correlations tests can be written as:

Shapiro-Francia (SF) $\quad \boldsymbol{w} = \boldsymbol{m}$

Shapiro-Wilk (SW) $\quad \boldsymbol{w} = \dot{\Sigma}^{-1}\boldsymbol{m} \Big/ \left\| \dot{\Sigma}^{-1}\boldsymbol{m} \right\|$ $\quad \|a\|$=Euclidean Norm of a

de Wet-Venter (DW) $\quad w_j = \varphi^{-1}\left({}^{j}/_{n+1} \right)$

Weisberg-Bingham (WB) $$w_j = \varphi^{-1}\left(\left.j-\frac{3}{8}\right/ n-\frac{3}{4}+1\right)$$

Smith-Bain (SB) $$w_j = \varphi^{-1}\left(\left.j-\frac{1}{2}\right/ n\right)$$

Each of these tests are shown to be asymptotically equivalent but in large sample sizes values of w are easily calculated for DW, WB and SB [25]. Practical estimation of ***m*** was also extended by Royston for $n \leq 2{,}000$ and then by Rahman and Govidarajulu for $n \leq 5{,}000$ [39-41]. The expected values of normal ordered statistics will be discussed further in Chapter 3.

A major advantage for correlation tests is the ease by which they can be applied to a censored sample by essentially restricting the portion of ṁ that matches with the observed data [42]. Additionally, Shapiro-Wilk has consistently demonstrated strong power across a plethora of non-normal distributions and has become the gold standard for overall power [43].

2.2.3. Moment Tests

Normality tests utilizing the moments of the distribution in their test statistic are known as moment tests and have existed conceptually since at least the late 19th century [44]. Given their behavior in the normal distribution, the third and fourth moments are natural metrics for testing. These are commonly denoted, respectively, by $\sqrt{b_1}$ and b_2 and defined as:

$$\sqrt{b_1} = \frac{\frac{1}{n}\sum (x_i - \bar{x})^3}{s^3} \quad \text{and} \quad b_2 = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum (x_i - \bar{x})^2\right)^2} - 3$$

$\sqrt{b_1}$ is equivalent to the moment-based estimate of skewness (while b_2 is the moment based estimate of excess kurtosis. As such, tests employing them are sometimes more plainly referred to as tests of skewness or kurtosis. Possibly the most well-known test of skewness comes from

D'Agostino who provides a transformation for b_1 which follows a closely normal distribution when the sample is derived from a normal distribution [45]. Anscombe and Glynn detail a similar procedure for b_2 [46]. Hosking utilizes L-moments to provide an alternate test for skewness [47]. Lin and Mudhokar base their Z_2 (initially named Z) test on the first and second moments utilizing the property that $\bar{x}$ and s^2 are independent if and only if the sample is drawn for a normal distribution [48]. The Z_2 test is shown to be particularly strong against skewed alternatives. Later Mudholkar *et al*. developed the Z_3 statistic based on the first and third sample moments and demonstrates strong power against non-normal kurtosis [49].

The greatest advantages of moment tests are also tied to their greatest disadvantages. Since these test statistics typically focus on either skewness or kurtosis it is very easy to amend them toward one-sided alternatives. For example, a researcher may only be interested to know if the data is left skewed without care to the data being right skewed. But as a tradeoff the examination of only one aspect of non-normality leads to considerably lower power compared to other classes of tests that are sensitive to all aspects of non-normality. In order to improve the overall power and create an omnibus test, skewness and kurtosis statistics may be combined as with the K^2 test of D'Agostino-Pearson [50], the combination of Z_2 and Z_3 [49], the tests by Bowman-Shenton [51] and that by Jarque-Bera [52]. More recently Bai and Ng present an omnibus test for time series data that is serially correlated [53]. These omnibus tests result in tests that are well powered against a wide range of distributions and the use of the individual statistics can be used to identify sources of non-normality [54]. However, this second point is refuted based on the correlation between $\sqrt{b_1}$ and b_2 being quite high even for large sample sizes [20, 22]. As a result, the diagnostic value of moment tests is questionable.

2.2.4. ECF Tests

ECF tests are based on upon the characteristic function, which for the normal distribution is

$$\phi(t) = exp(it\mu - \frac{1}{2}t^2\sigma^2)$$

and the empirical characteristic function for a sample of size n is

$$\phi_n(t, X_n) = \frac{1}{n}\sum_{j=1}^{n} exp(itx_j)$$

The earliest ECF test surveyed by this review was by Koutrouvelis whose test statistic divided the characteristic function into real and imaginary parts [55]. Around the same time Murota and Takechu proposed a test based on $a_n(t) = \left|\phi(t, {}^{X_n}/_{S})\right|$ which demonstrated good power against symmetric distributions but not against skewed distributions [56]. Epps and Pulley provide a test that is robust against both skewed and kurtic alternatives and is probably the most well-known of ECF tests [44]. Their test statistic is

$$T = \int_{-\infty}^{\infty} \left|\phi_n\left(t, \frac{X_n - \bar{x}}{s}\right) - \hat{\phi}(t, X_n)\right|^2 dG(t)$$

in which $\hat{\phi}(t, X_n) = exp\left(it\bar{x} - {}^{1}/_{2}\, t^2 s^2\right)$ and dG(t) is a weight function initially suggested as $\left(\sqrt{2\pi}\right)^{-1} exp({}^{-t^2}/_{2})$.

The Epps-Pulley test has been shown to have comparable power to other well-known normality tests with recommendation for its relatively easy computation [25]. ECF tests have received much praise in the multivariate realm. Whereas EDF and moment tests are criticized for lacking consistency [57], Csorgo's application of the empirical distribution function to the multivariate setting [58] was, at the time, praised as being the only multivariate normality (MVN) test "genuinely" multivariate in nature [59]. Application of the ECF to testing MVN is still ongoing [60].

2.2.5. Other Tests

Tests that do not clearly fit into any of the previous categories are grouped together as a separate category. These include tests based on Hermite polynomials [61-63] and the moment generating function [64]. Of particular interest to this dissertation is the test by Sigut *et al*. [65], which combines Bayesian elements with large deviation theory [66].

Unlike most normality tests Sigut *et al.*, through the use of so-called experts, offers the ability not just to discriminate between normal and non-normal but also provide diagnostic value. Let C_0 be the class of any sample of size n drawn from a $\mathrm{N}(0,1)$ population and let C_1be the class of all samples of a non-normal population with the same size and parameters. The problem is then to determine if X_n belongs to C_0 or C_1. The authors introduce the notion of an expert, based upon the optimal decision rule under a Bayesian framework given by:

$$\frac{P(C_1)p(X_n|C_1)}{P(C_0)p(X_n|C_0)}$$

in which $P(C_i)$ and $p(X_n|C_i)$ are the a priori probabilities and CDFs for i equal to 0 or 1. The error associated with such a decision rule is given by:

$$E=P(C_0)E_0+P(C_1)E_1$$

where E_0 and E_1 are the errors associated with misclassifying samples from C_0 and C_1 respectively. E_0 can be rewritten as:

$$P_0\left(\frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\log\frac{p(X_n|C_1)}{p(X_n|C_0)} \geq \frac{1}{n}\log\frac{P(C_1)}{P(C_0)}\right)$$

with P_0 being shorthand for the probability of the event in the case that all observations are from the normal class. Using Chernoff's theorem, it can be shown that as n tends to infinity this probability relates to the Chernoff distance between $p(X_n|C_0)$ and $p(X_n|C_1)$, which is independent of the prior probabilities and decays exponentially with the sample size. These arguments are

applicable to both E_1 and E_0 as well establishing the expert as an asymptotically optimal decision rule. The authors then extend the concept to a sum of experts

$$\frac{P(C_{11})p(X_n|C_{11})}{P(C_0)p(X_n|C_0)}+\frac{P(C_{12})p(X_n|C_{12})}{P(C_0)p(X_n|C_0)}+\ldots+\frac{P(C_{1r})p(X_n|C_{1r})}{P(C_0)p(X_n|C_0)}$$

showing that the previous results hold for this combination. In practice care must be taken so that the chosen experts provide a comprehensive basis for "non-normal" as samples derived from non-normal populations that are not reasonably "close" to any of the experts become problematic. In response the Johnson family of distributions is recommended with four parameterizations intended to cover the plane of non-normality defined by the population skewness and kurtosis given, respectively, as:

$$\sqrt{\beta_1}=\frac{E\left[(X-\mu)^3\right]}{(E[(X-\mu)^2])^{3/2}}$$

and

$$\beta_2=\frac{E\left[(X-\mu)^4\right]}{(E[(X-\mu)^2])^2}$$

Specifically, the quadrants of interest are:

i. $\sqrt{\beta_1} < 0,\ 1 < \beta_2 < 3$

ii. $\sqrt{\beta_1} < 0,\ 3 < \beta_2 < 10$

iii. $\sqrt{\beta_1} > 0,\ 1 < \beta_2 < 3$

iv. $\sqrt{\beta_1} > 0,\ 3 < \beta_2 < 10$

Distributions that differ from normal in both skewness and kurtosis will be identified by the expert associated with a specific type of deviation. For distributions that deviate only by skewness, then a combination of either (i) and (ii) or (iii) and (iv) will be large, depending on

whether the distribution is left skewed or right skewed. Similarly, the combination of (i) and (iii) or (ii) and (iv) will identify distributions that are leptokurtic or platykurtic with zero skew.

2.3. Transformation

Data that is found to be non-normal may be altered through (non-linear) transformation to achieve normality. Feature transformation is actually quite common across the omic sciences including metabolomics. Some of the more popular transformations are shown in Table 2.1. Many involve a linear mechanism via centering (mean/median), scaling (variance / standard deviation) or both, and offer no change in distribution. This is sufficient in the typical metabolomics study, focusing on biomarker discovery in a sea of features with wildly varying abundance levels, where transformations are primarily used to remove variation within and between features. For example, when conducting univariate t-tests or analysis of variance (ANOVA) it is helpful to remove heteroscedasticity. In the case of multivariate analysis via principle component analysis then it may be desirable for the compounds to have equal weight [67].

Although normality is not often the primary focus, some transformations (g-i) are non-linear and do offer the possibility to alter the distribution. Both the power and Box-Cox transformations are directly intended for such use and can be directly applied to metabolomic data as ion counts are strictly greater than 0. They do, as with the generalized log, require estimation of an additional parameter δ from the data. With the power and Box-Cox transformations this parameter can be estimated using profile likelihood and is dependent on a fully observed sample. The natural log transformation is frequently used to help induce normality, being in fact a special case of the Power and Box-Cox when $\lambda_j = 0$. It is a popular choice in general and evidence supports its value in metabolomics [68]. Unfortunately, it has the unattractive consequence of

Table 2.1: Common Transformations in Metabolomics

	Name	Calculation
(a)	Centering	$y_{ij}^{*} = y_{ij} - \bar{y}_i$
(b)	Autoscaling	$y_{ij}^{*} = \frac{y_{ij} - \bar{y}_i}{s_i}$
(c)	Range Scaling	$y_{ij}^{*} = \frac{y_{ij} - \bar{y}_i}{max(y_i) - min(y_i)}$
(d)	Pareto Scaling	$y_{ij}^{*} = \frac{y_{ij} - \bar{y}_i}{\sqrt{s_i}}$
(e)	Vast Scaling	$y_{ij}^{*} = \frac{\bar{y}_i (y_{ij} - \bar{y}_i)}{s_i^2}$
(f)	Level Scaling	$y_{ij}^{*} = \frac{y_{ij} - \bar{y}_i}{\bar{y}_i}$
(g)	Log_a	$y_{ij}^{*} = \log_a(y_{ij})$
(h)	Power	$y_{ij}^{*}(\lambda_j) = \begin{cases} y_{ij}^{\lambda_j} & \text{for } \lambda_j \neq 0 \\ \ln(y_{ij}) & \text{for } \lambda_j = 0 \end{cases}$
(i)	Box-Cox	$y_{ij}^{*}(\lambda_j) = \begin{cases} y_{ij}^{\delta} - 1 \big/ \lambda_j & \text{for } \lambda_j \neq 0 \\ \ln(y_{ij}) & \text{for } \lambda_j = 0 \end{cases}$
(j)	Generalized Log	$y_{ij}^{*}(\delta, \alpha) = \ln\left((y_{ij} - \alpha) + \sqrt{(y_{ij} - \alpha)^2 + \delta}\right)$

greatly inflating the transformed variance in features with lower abundance. This phenomenon is consistent with the two-component model introduced in Section 1.7. Returning to this model, the ion count y_{ji} for some metabolite j and sample i is

$$y_{ji} = \alpha + c_j x_{ji} e^{\eta_{ji}} + \varepsilon_{ji}$$

with $\eta \sim N(0,\sigma_\eta^2)$, $\varepsilon \sim N(0,\sigma_\varepsilon^2)$, x_{ji} is the true concentration and c_j represents the overall ionization efficiency of the instrument. For this exercise it convenient to combine the true concentration

and instrument ionization as $\mu_{ji} = c_j x_{ji}$ representing the true average ion count. When μ_{ji} is large, the middle term dominates the expression and the distribution of y_{ji} is reasonably close to a log-normal and the log transformation performs well. But, when μ_{ji} is small the last term dominates and y_{ji} behaves more as a normal random variable. The mean and variance of y_{ji} follow as

$$E\left[y_{ji}\right] = \alpha + \mu_{ji}$$

$$Var\left(y_{ji}\right) = \mu_{ji}^2 \Lambda_\eta^2 + \sigma_\varepsilon^2$$

with $\Lambda_\eta^2 = \mathrm{e}^{\sigma_\eta}(\mathrm{e}^{\sigma_\eta} - 1)$, which is the variance of a log-normal variable with mean 0 and standard deviation σ_η (on the log scale). From the observation above $\mu_{ji} \to 0$ implies that $y_{ji} \xrightarrow{D} N$. Using the delta method, the asymptotic variance when applying the log transformation must be converging to

$$var\left[\log\left(y_{ji}\right)\right] \to \left(\mu_{ji}^2 \Lambda_\eta^2 + \sigma_\varepsilon^2\right) * \frac{1}{(\mu_{ji} + \alpha)^2}$$

as n goes to infinity. For the next step, recall that α is mostly regarded as a nuisance parameter and can frequently be ignored. However, as μ_{ji} becomes small this intercept may arguably become non-trivial. One can either make the assumption that α is ignorable or adjust the transformation $\log\left(y_{ji} - \alpha\right)$. In either case, the result follows:

$$var\left[\log\left(y_{ji} - \alpha\right)\right] \to \left(\mu_{ji}^2 \Lambda_\eta^2 + \sigma_\varepsilon^2\right) * \frac{1}{\mu_{ji}^2}$$

$$\to \Lambda_\eta^2 + \frac{\sigma_\varepsilon^2}{\mu_{ji}^2}$$

Thus as μ_{ji} approaches 0 the variance will increase without bound. $\mu_{ji} \to 0$ will obviously occur when $y_{ji} \to 0$, suggesting the variance will be high for features in which μ_{ji} is small. This implies

low metabolites with low concentration will tend have higher variance; however, recall the ion count is dependent on the product of both the concentration and ionization efficiency of the molecule. A small concentration of a metabolite with a high ionization efficiency may produce more ions than a large concentration of a metabolite that is poorly ionized.

Inflation of variances at low ion counts is problematic as it can hinder the ability to identify biomarkers that fall in this range. One solution is to only log transform metabolites with higher abundances and leave the low abundance features alone. This can work well at the extremes, i.e. ranges where $\mu_{ji}e^{\eta_{ji}}$ or ε_{ji} dominate, but ranges in which both terms meaningfully contribute remains problematic. The generalized log transformation (GLOG) offered a solution to this [69].

Although utilized for microarrays, the concept is drawn directly from the model first presented by Rocke and Lorenzato and natural to use with MS data [70]. As with Power/Box-Cox, the GLOG involves estimation of a parameter, which can also be found using Maximum Likelihood, and is based on the ratio of $\sigma_\varepsilon^2/\Lambda_\eta^2$ [69]. Note that with the power family the transformation parameter is estimated on each individual feature whereas in the GLOG a single parameter set is typically used for all features. Further detail on the advantages and disadvantages of these transformation can be found in van den Berg *et al*. [71].

2.4. Correlation

The final item of interest in this chapter is that of correlation. For $\boldsymbol{y_A}$ and $\boldsymbol{y_B}$ any two metabolites with means μ_A, μ_B and variances σ_A, σ_B the correlation is defined as

$$\rho_{\boldsymbol{x_A x_B}} = \frac{E[\boldsymbol{y_A} - \mu_A]E[\boldsymbol{y_B} - \mu_B]}{\sqrt{E[\boldsymbol{y_A} - \mu_A]^2 E[\boldsymbol{y_B} - \mu_B]^2}}$$

The numerator is simply the covariance between $\boldsymbol{y_A}$ and $\boldsymbol{y_B}$ while the denominator is the square root of the product of the two standard deviations. Hence, correlation is the covariance

normalized against the variation of two variables and is bounded between -1 and 1. A value of ρ equal to 0 implies independence between the two variables, meaning that knowledge of one variable provides no information about the other variable. When $\rho > 0$ the two are said to be positively correlated with higher values in one variable being associated with higher values in the other. Negative correlations occur when $\rho < 0$ and higher values in one variable tend to indicate lower values in the other. A value of $\rho = -1$ or $\rho = 1$ will be achieved when the two variables are an exact linear transformation of each other. The sample correlation is most frequently described using the Pearson *r* correlation coefficient [72], which replaces the population parameters with the usual sample estimates:

$$\text{Pearson's } r_{\boldsymbol{x_A},\boldsymbol{x_B}} = \frac{(\sum \boldsymbol{y_A} - \bar{x}_A)(\sum \boldsymbol{y_B} - \bar{x}_B)}{\sqrt{\sum(\boldsymbol{y_A} - \bar{x}_A)^2 \sum(\boldsymbol{y_B} - \bar{x}_B)^2}}$$

Note that in the above formula the $(n\text{-}1)^{-1}$ terms cancel out. Pearson's *r* shares the same boundary space as ρ with the interpretation also being the same. As one might expect with ρ signifying a linear relationship, this statistic is closely related to simple linear regression (SLR). In fact, *r* is the square root of the R^2 statistic, which measures the proportion of variation explained by the linear fit. It is therefore unsurprisingly that, like SLR, r is sensitive to extreme outliers. A non-parametric version of the correlation coefficient was developed by Spearman [72] and is in fact often recommended for metabolomic data due to the propensity for extremely large outliers [11]. The calculation of Spearman's rank sum correlation coefficient follows exactly from Pearson with the exception being that values of both variables are first translated into ranks:

$$\boldsymbol{y}_{\boldsymbol{A}}^{*} = \{rank(y_{A1}), \ldots, rank(y_{An})\}$$

$$\boldsymbol{y}_{\boldsymbol{B}}^{*} = \{rank(y_{B1}), \ldots, rank(y_{Bn})\}$$

and so

$$\text{Spearman's } r_{y_A y_B} = 1 - \frac{6\sum_1^n (y_{A1}^* - y_{B1}^*)}{n^3 - n}$$

Whereas as Pearson indicates a linear relationship between $\boldsymbol{y_A}$ and $\boldsymbol{y_B}$, Spearman indicates a monotonic relationship. Spearman is therefore more general and able to capture a wider range of relationships in the two variables and is also invariant to any monotonic transformation. However, as an estimator Spearman is less statistically efficient than Pearson with hypothesis tests for $\rho = 0$ using Pearson being more powerful. However, in this paper interest is more on the coefficients rather than p-values. Correlation coefficient estimates are tabulated and examined for each metabolite using both Pearson and Spearman. The availability of pathway information for the metabolites allows for an additional layer of assessment based upon functional class. Details on these pathways are given in section 2.6.

2.5. Methods

As metabolite distributions and correlations are examined it is important to consider the potential impact that transformations may have. In each of the normality and correlation methods used, results are provided for the untransformed data, as well as data transformed by log, GLOG and Box-Cox. These transformations are selected given their ability to alter the distribution of the metabolite and the interest of this paper. It is our belief that a simple log transformation is effective at inducing normality within metabolites. The GLOG is a closely related function and is rooted in MS based analytical chemistry, making it a natural choice for consideration. The Box-Cox transformation provides a gold standard by which to assess both log based transformations.

Correlation is assessed by pairwise comparisons using both Pearson and Spearman statistics. Three approaches are used to assess normality: Shapiro Wilk, Normal Quantile-Quantile plots and Box-Cox λ. The Shapiro-Wilk test is among the most recognizable tests for normality. Among the first correlation tests, it was first described in 1965 [73]. It is regarded for being

among the most powerful and having high power across a large spectrum of non-normal behavior [33]. These reasons led to its selection here as theoretically metabolites are not constrained to any specific type of non-normality. Additionally, as a correlation-based test Shapiro-Wilk can be adapted to censored samples allowing its application to metabolites in the dataset that contain missing values. Additional detail for these methods follow.

2.5.1. Shapiro-Wilk

For a given metabolite $\boldsymbol{y_i} = \{y_{1i}, \ldots, y_{ni}\}'$, the test statistic is:

$$W_i = \frac{\left(\sum_{j=1}^{n} a_i y_{[j]i}\right)^2}{\sum_{j=1}^{n} (y_{ji} - \bar{y})^2}$$

where $y_{[j]i}$ is the j^{th} ordered value of the vector $\boldsymbol{y_i}$

$$\{a_1, \ldots, a_n\} = \frac{\boldsymbol{m}^{\mathrm{T}} V^{-1}}{\left(\boldsymbol{m}^{\mathrm{T}} V^{-1} V^{-1} \boldsymbol{m}\right)^{1/2}}$$

with

$$V = cov(\boldsymbol{m})$$

$$\boldsymbol{m} = (m_1, \ldots, m_n)$$

such that

$$m_j = E\left[y_{[j]i} \mid \boldsymbol{y}_i \sim N(0, I)\right]$$

Due to the symmetry of the normal distribution $\sum m_j$, and by extension $\sum a_i$, is equal to 0. W_j is bounded between 0 and 1 and under H_0 will be closer to 1. Censored samples are handled by restricting the range to cover only the observed order statistics. That is

$$W_j = \frac{\left(\sum_{\Omega} a_i y_{[j]i}\right)^2}{\sum_{\Omega} (\mathrm{y}_{ji} - \bar{\mathrm{y}})^2}$$

where $\Omega = \{\text{all } j : y_{[j]i} \text{ is not missing}\}$.

2.5.2. Normal Quantile-Quantile plots

The normal quantile-quantile (QQ) plot is the most commonly used and recognized graphical approach. For $\boldsymbol{y_j}$ and $\boldsymbol{m}$ as before, the QQ plot displays the pairs

$$\{y_{[j]i}\,,\, m_j\}$$

Samples that are normal should follow closely to a straight line while non-normal behavior will manifest with non-linear behavior. For example, pairs of a sample from a right skewed distribution will tend to show an upward curve for larger values of i. Although QQ plots are easy to construct and straightforward to interpret, the lack of a formal decision-making structure causes the conclusions to be subjective. High dimensional data presents an additional challenge as it is infeasible to examine hundreds or thousands of features in this manner. The approach taken here is to consider all possible y_{ji} for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$ simultaneously. Each metabolite is mean centered and scaled by the sample standard deviation:

$$y_{ji}^{*} = \frac{y_{ji} - \bar{y}_j}{s_{y_j}}$$

These resulting y_{ji}^{*}'s are then combined into a single vector and all resulting pairs $\{y_{[k]}^{*}\,,\, m_k\}$ for $k \in \{1, \ldots, j*i\}$ are plotted.

The QQ plot will be used to give a graphical assessment. To accommodate the large number of variables each compound will be median scaled and standardized. A single QQ plot will then be made for all the values of all the compounds. Because missing values are likely due to limit of detection and not missing at random only, compounds that are observed in all samples will be used. Next, normality will also be assessed using the Shapiro-Wilk test. P-values can easily be tabulated across all compounds and categorized based on standard significance at the 0.05 level.

This test is the gold standard for univariate normality and its use here is consistent with previous studies.

2.5.3. Box-Cox λ

For a variable of interest $\boldsymbol{y_j}$, the family of power transformations introduced by Tukey [74] produces a variable $\boldsymbol{y_j}(\lambda_j)$ that is optimally normal based on:

$$\boldsymbol{y_j}(\lambda_j)=\begin{cases} \boldsymbol{y_j}^{\lambda_j} & \text{for } \lambda_j \neq 0 \\ \ln\left(\boldsymbol{y_j}\right) & \text{for } \lambda_j = 0 \end{cases}$$

which is monotonic for all λ_j. Box and Cox amended this to account for the discontinuity that occurs at $\lambda_j = 0$, yielding the following definition:

$$\boldsymbol{y_j}(\lambda_j)=\begin{cases} \dfrac{\boldsymbol{y_j}^{\lambda_j}-1}{\lambda_j} & \text{for } \lambda_j \neq 0 \\ \ln\left(\boldsymbol{y_j}\right) & \text{for } \lambda_j = 0 \end{cases}$$

λ_j is estimated using a profile likelihood based on

$$\left(\sigma\sqrt{2\pi}\right)^{-n} * exp\left\{\frac{\left(\boldsymbol{y_j}^{\lambda_j}-\boldsymbol{A\theta}\right)'\left(\boldsymbol{y_j}^{\lambda_j}-\boldsymbol{A\theta}\right)}{2\sigma^2}\right\}$$

in which $\boldsymbol{A}$ is a matrix of predictor variables and $\boldsymbol{\theta}$ is an unknown vector of associated parameters based on the model

$$E\left(\boldsymbol{y_j}\right)=\boldsymbol{A\theta}$$

Hence Box-Cox is appropriate for transforming dependent variables of a linear model, but can be generalized to other variables by setting $\boldsymbol{A}$ = **1.** Some definitions incorporate the geometric mean of untransformed variable $GM\left(\boldsymbol{y_j}\right)=\sqrt[n]{y_{j1} * y_{j2} * \ldots * y_{jn}}$ as this allows for simplification of the likelihood. These transformations are only appropriate for variables strictly greater than zero, and

so the two parameter Box-Cox substitutes $\boldsymbol{y}_j' = (\boldsymbol{y}_j + \alpha_j)$ in place of $\boldsymbol{y}_j$ in order to ensure that all values are positive. For MS metabolomic data the single parameter version given should suffice as detectable ion-counts are strictly greater than zero. Occasionally, missing values may be treated as zero, but this is a very strong assumption implying the metabolite is completely absent from the sample. This may hold for some metabolites, such as pharmacological agents or xenobiotics, but unlikely to be true for the vast majority of missing values. For the most part the components of metabolic processes are expected to be present in all subjects to at least some degree (though maybe too low to be observed). Finally, notice that $\lambda_j = 1$ indicates the original $\boldsymbol{y}_j$ is the optimal choice for normality and thus λ_j itself can serve as assessment of normality in the same manner as skewness and kurtosis.

2.6. Datasets

Three separate cohorts of human plasma, urine and CSF were profiled using LC-MS. These sets were part of a collaborative work between Metabolon and the Department of Molecular and Human Genetics at Baylor College of Medicine and were previously described in Elsea *et al.* [64]. Samples were drawn from pediatric subjects who were found to be free of any IEMs. Hence, these samples form a control group for which to compare suspected cases of IEMs and, as demonstrated in Miller, Kennedy, Eckhart, Burrage, Wulff *et al.* {Miller, 2015, Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism, have led to significant advancements in the diagnosis of such diseases. As a single, homogenous group these sample sets provide an excellent opportunity to assess metabolite properties in the absence of any experimental treatment or factor. Further, metabolites can be influenced by demographic factors in adults, such as differences in gender, race or age, these sets being a pediatric cohort implies such factors prevalent after puberty should be less important here. Samples were collected and

Table 2.2: Summary of metabolite data sets.

Matrix	# Samples	# Metabolites					# with missing values (%)
		Neg	Pos Ear	Pos Lat	Polar	Total	
Plasma	31	522	218	187	79	1006	464 (46.2)
CSF	31	171	203	93	56	523	266 (50.8)
Urine	40	715	376	22	125	1238	748 (60.4)

stored by the Department of Molecular and Human Genetics at Baylor College of Medicine and have previously been described in Elsea *et al*. [75]. Metabolomic analysis was conducted utilizing a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI) source [76]. For more specific details see Elsea *et al*. [75], DeHaven *et al*. [12], Evans *et al*. [77] and Evans *et al*. [76]. In addition to the tier 1 and tier 2 [78] metabolites that were previously reported, the data here also include those features with unique mass spectral signatures but did not match with any available chemical standard in the library.

Table 2.2 summarizes the major characteristics of the three datasets. Identified features are matched against an in-house library. Those features that had a match are referred to as “named”. Those features having a distinct MS/MS signature but no available standard in the library are referred to as “unknown”. Missing values are observed in around half of the metabolites within each set, which presents a challenge. Missing values have previously been shown to be associated with lower ion abundance and are thus often assumed to be related to limit of detection [79]. Propensity for missing values in these data are shown to be greater for lower abundant metabolites. *Figure 2.1* plots the proportion of observed values by the median of the observed abundance level of the observed values. Because median abundance levels range from

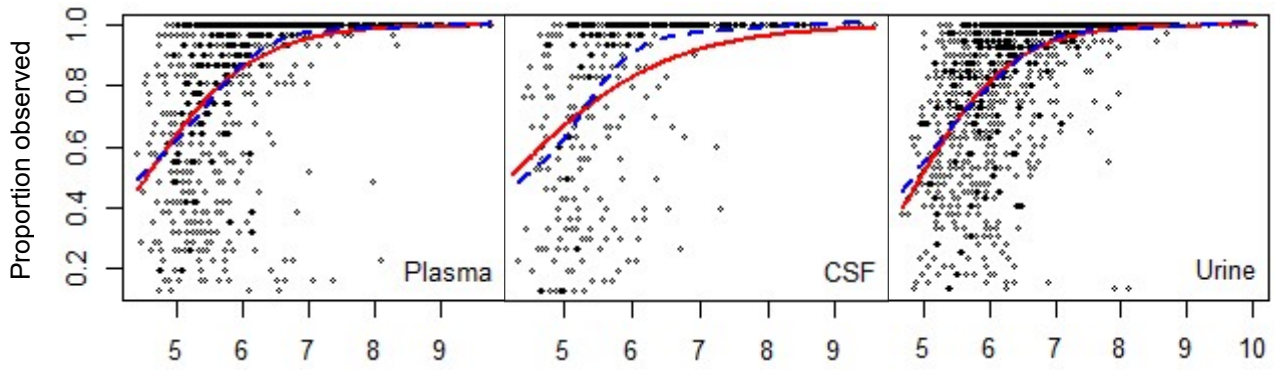


Figure 2.1: Proportion of observed samples by abundance level in plasma, CSF and urine. *X*-axis is $\log_{10}$ median abundance level. *Y*-axis is the proportion of samples observed. Solid red line represents logistic regression fit. Dotted blue line is a splinal fit using 10 knots. Biochemicals in which less than 10% of the samples are observed are excluded.

25 thousand up to 10 billion, it is convenient to $\log_{10}$ transform the medians for plotting.

Transformed abundance level is used to model the proportion of observed value using logistic regression and smooth splines. Both indicate that rate of observation increases as abundance level increases and the two fits overlap considerably in plasma and urine. For the most part, almost all biochemicals with medians ion count above 10^7 are fully observed. There are a few aberrant compounds in which less than 10% of the samples are observed but the median abundance level is well above the 10^7. Predominantly these are pharmacological metabolites and xenobiotics. For example, in the urine plot there are two compounds with a $\log_{10}$ median abundance around 8 but were observed in less than 20% of samples. These compounds are lidocaine and dexopanthenol. In fact, upon further inspection almost all of the biochemicals observed in less than 10% of samples were found to be related to pharmacological agents. For this reason, the plots have been restricted to show only those metabolites in which more than 10% of samples were observed. With missing values being associated with lower abundance level, the observed value for a metabolite then constitute a left censored sample, which makes it challenging to assess normality. For simplicity, only the metabolites that were observed in every sample are used here, except for Shapiro-Wilk analysis which can deal with censored samples.

Metabolites in which a standard is available in the library can be grouped according to their biochemical functions. This enables pathway associations using predetermined functional classes, which are shown in Table 2.3. In general pathway assignments can be somewhat arbitrary due to various roles metabolites play. Arginine, for example, is critical in urea cycle but also an important component in cellular division and also the regulation of blood pressure, among other functions. While the assignments given here could arguably be different in some cases, the overall portrait given is accurate. As the total number of biochemicals varies across the three sets, the overall counts can be somewhat misleading. CSF for example has the highest amount of amino acids as a proportion of its total, even though the raw number is less than either of the other two sets. *Figure 2.2* shows the relative proportion of each pathway. The three most dominant classes are Amino Acids, Lipids and Unknown and these serve to differentiate the three datasets. Lipids, or fatty acids, are highest in plasma, though this class is also a large contribution to CSF as well. In fact, the overall the pathway distribution is rather similar between plasma and CSF with the former having a bit higher proportion of lipids and the latter having a higher proportion of amino acids. Still, the main difference between these two matrices is density, with plasma having nearly twice as many biochemicals as CSF. Urine is actually the densest of the three, with the dominant class being unknowns, the features without an existing

Table 2.2: Pathway Designations by Matrix

Pathway Category	Plasma	CSF	Urine
Amino Acid	172	139	212
Carbohydrate	25	20	36
Cofactors and Vitamins	26	20	37
Energy	9	10	12
Lipid	331	136	96
Nucleotide	33	32	48
Peptide	34	17	35
Xenobiotics	95	48	149
Unknown[†]	281	101	613

[†] Pathway information not available for molecules with unknown structure

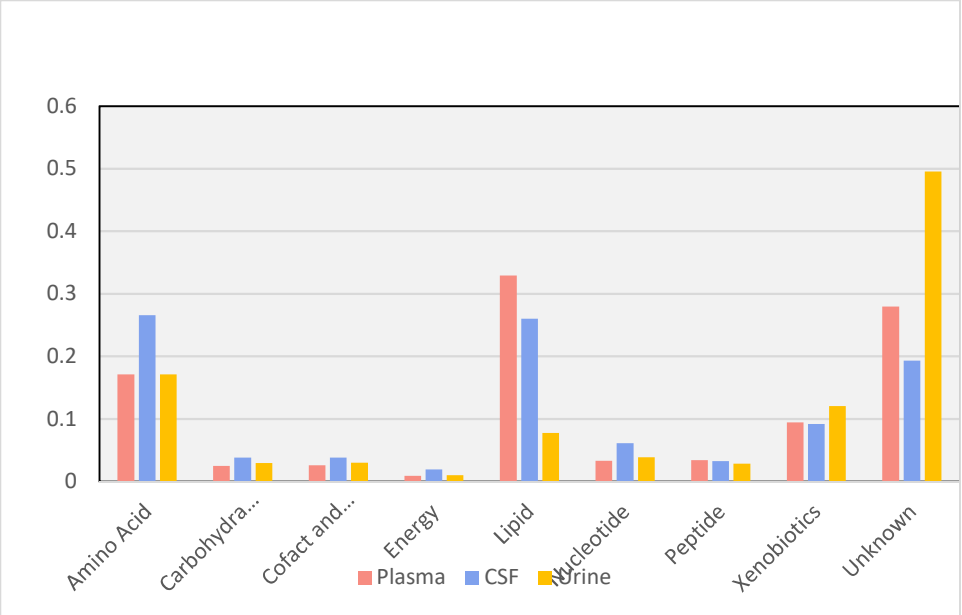


Figure 2.2: Distribution of the eight major pathways plus unknown in plasma, CSF and urine. Unknowns represent distinct mass spectrum signatures but did not matching library entry preventing classification. *Y*-axis is percentage of biochemicals per matrix.

library match, which makes up about half of all the metabolites found in this matrix. Though the lack of available standard to provide a chemical name is somewhat limiting, it is still useful when comparing against the other two matrices suggesting that nearly half of the compounds measured in urine are not found in the other two. It is noteworthy that such a large percentage of the urine biochemicals did not have an available standard in the library. At the very least this further reinforces that the composition of urine is quite different, while CSF and plasma are very similar with the latter almost being a subset of the former. To illustrate this point, *Figure 2.3*, courtesy of BioVenn [80], shows a Venn diagram of the individual biochemicals. This illustrates the significant overlap in biochemicals between plasma and CSF. 81.8% (428 of 523) of the compounds identified in CSF are also found in plasma, whereas the overlap with urine is 72.1% (377 of 523). Meanwhile only 56.7% (570 of 1006) of the plasma metabolites overlap with urine. All together this shows that the composition of plasma and CSF is extremely similar. CSF also has quite a bit in common with urine, while urine and plasma only share about half of their

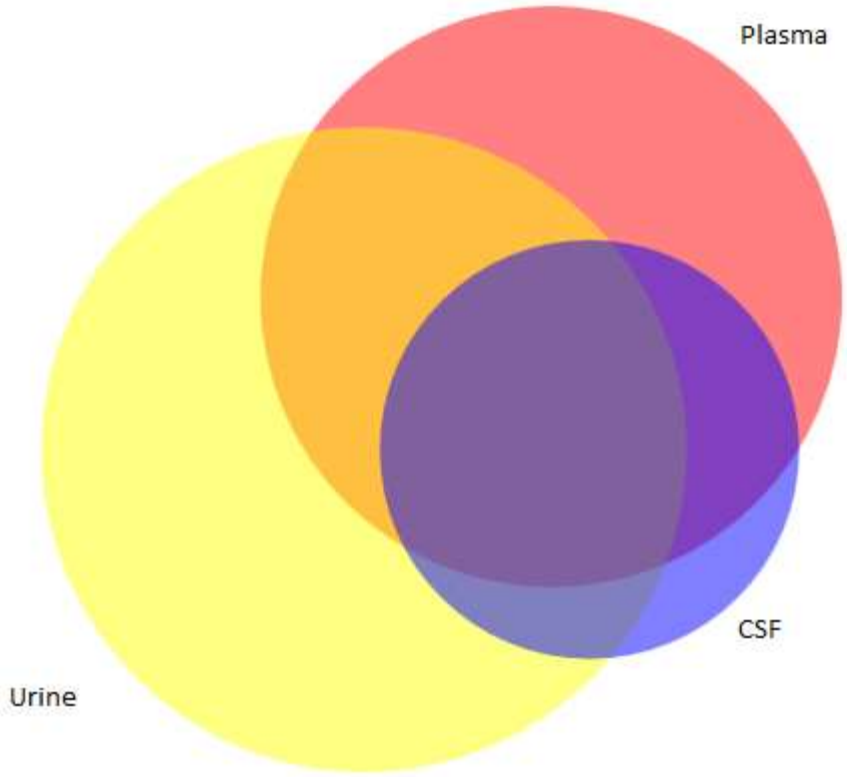


Figure 2.3: Venn Diagram of observed biochemicals in plasma, CSF and urine.

makeup. Note that while this shows the composition, in terms of metabolites present, is similar, it does imply that the concentrations of those metabolites are similar. Answering this completely would require running plasma and CSF samples from the same subject concurrently and using isotopically labeled standards to account for matrix effects. However, from the ion abundance levels shown in *Figure 2.1*, abundance levels for plasma tend to cluster more around $10^{5.5}$ - 10^{6} whereas CSF biochemicals cluster more around 10^{5}. This suggests that concentrations in CSF are likely lower than that seen in plasma.

Finally, while plasma and CSF are subject to homeostatic regulation, the concentrations of metabolites in urine are known to be strongly influenced by endogenous factors such as water intake, exercise, diet and disease state [81-83]. Creatinine normalization is often used to correct for such effects, though the practice of normalizing across compounds of such vastly different

biochemical functions and physical properties is not well understood. Normalizations differ from transformations in that these are data manipulations within samples while transformations manipulate metabolite features. In a data frame in which the rows are the samples and the columns are the metabolites, normalizations operate on the rows while transformation act on the columns. Results are given for urine both with creatinine normalization and without (un-normalized). For the most part normalization does not appear to be change the overall conclusions.

2.7. Results

2.7.1. Normality

Summary measures in all three sets suggests clear non-normality across the matrices. *Figure 2.4* shows the histograms of the sample skewness and kurtosis as well as Box-Cox λ across all of the biochemicals fully observed. The first notable observation is the surprising consistency of behavior across the four data versions. Urine is present twice, once without any normalization and also with a creatinine normalization. Skewness is consistently greater than 0 and frequently with a magnitude greater than 1, suggesting metabolite distributions have a significantly heavy right tail. Metabolites are also consistently lepto-kurtic, in many cases with excess kurtosis of 5 or more. Estimated $\lambda_i's$ for Box-Cox transformation cluster tightly and nearly symmetrically about 0. Since $\lambda_i = 0$ indicates a natural log transformation, seeing the metabolites centered on this value supports the natural log as a good candidate transformation for achieving normality. It's useful to point out that there is no other clear peak in the λ_i values. One might have expected a significant portion of metabolites to be normal, which would cause clustering around 1. However, the uni-modal and symmetric shape to the $\lambda_i's$ suggests metabolites are roughly log-normal with no favored alternative to this behavior.

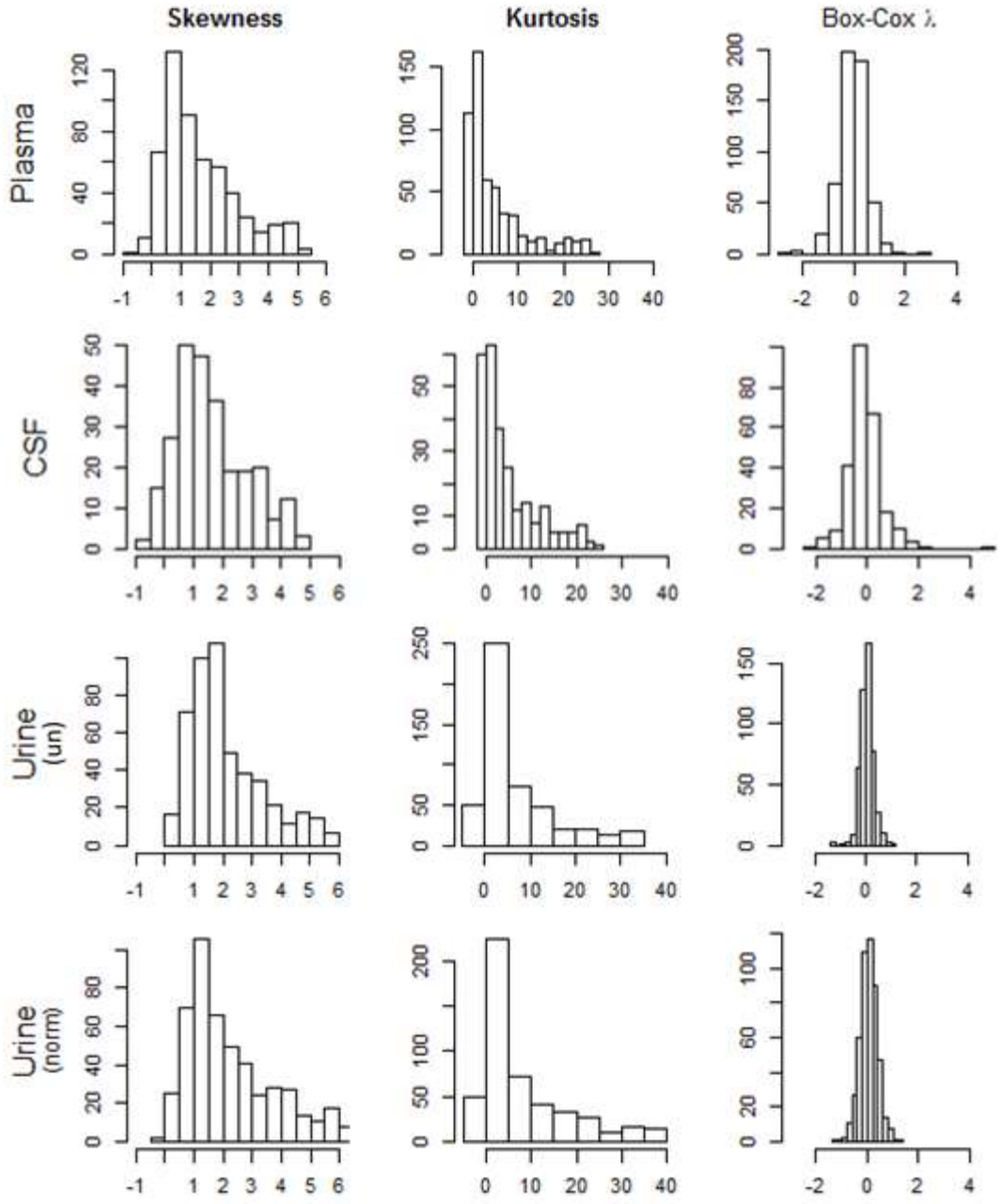


Figure 2.4: Summary normality measures in raw data. Statistics are calculated on each biochemical that is fully observed in that respective matrix.

Results for Shapiro Wilk support the raw data being non-normal and that log transforming does indeed lead to more normal behavior. Table 2.4 shows the rejection rate of the fully observed compounds in the three matrix versions. Rates are given for the raw data as well as using a log, GLOG or Box-Cox transformation. Using $p < 0.05$ the null rejection rate is 5%. No

Table 2.4: Shapiro-Wilk rejection rate in fully observed metabolites. Rejection is based on $p < .05$

		Plasma		CSF		Urine (un)		Urine (norm)	
# Compounds		542		257		490		489	
# Rejected	Raw	416	76.8%	189	73.5%	466	95.1%	452	92.4%
	Log	135	24.9%	57	22.2%	101	20.6%	199	40.9%
	glog	136	25.1%	56	21.8%	99	20.2%	306	62.6%
	Box-Cox	23	4.2%	11	4.3%	15	3.1%	45	9.8%

fewer than 70% of metabolites are rejected in the raw data across the four datasets, strongly supporting non-normal behavior. Proportions are highest in urine, where both un-normalized and creatinine normalized versions are over 90%. The log transformation reduces the rejection rate greatly, down into the mid to high 20s in plasma, CSF and un-normalized urine. So, while the log transform significantly improves deviation from normality, roughly a quarter of the metabolites are still displaying statistically significant non-normal behavior. Rejection rate for GLOG transformation is overall similar to the log with the log being slightly lower in plasma and the GLOG being a little lower in CSF and un-normalized urine. Box-Cox reduces the rejection rate the furthest further. In plasma the rejection rate is below the null rate. Although rejection in CSF and urine under Box Cox remains slightly higher than the null rate, it is still much lower than either log or GLOG. Rejections rates in the normalized version of urine are much higher than the other matrices across all transformation versions. Log transformation is rejected at nearly twice the rate of the other data versions and GLOG is rejected almost three times more. This dataset also has the highest Box-Cox rejection rate which, at nearly 10%, is twice the rejection rate in plasma. This is somewhat unexpected as, from *Figure 2.4*, the lambda values for both versions of urine were more closely clustered around 0 than plasma and CSF. However, *Figure 2.4* shows that the urine data contains more skewness and kurtosis than the other two

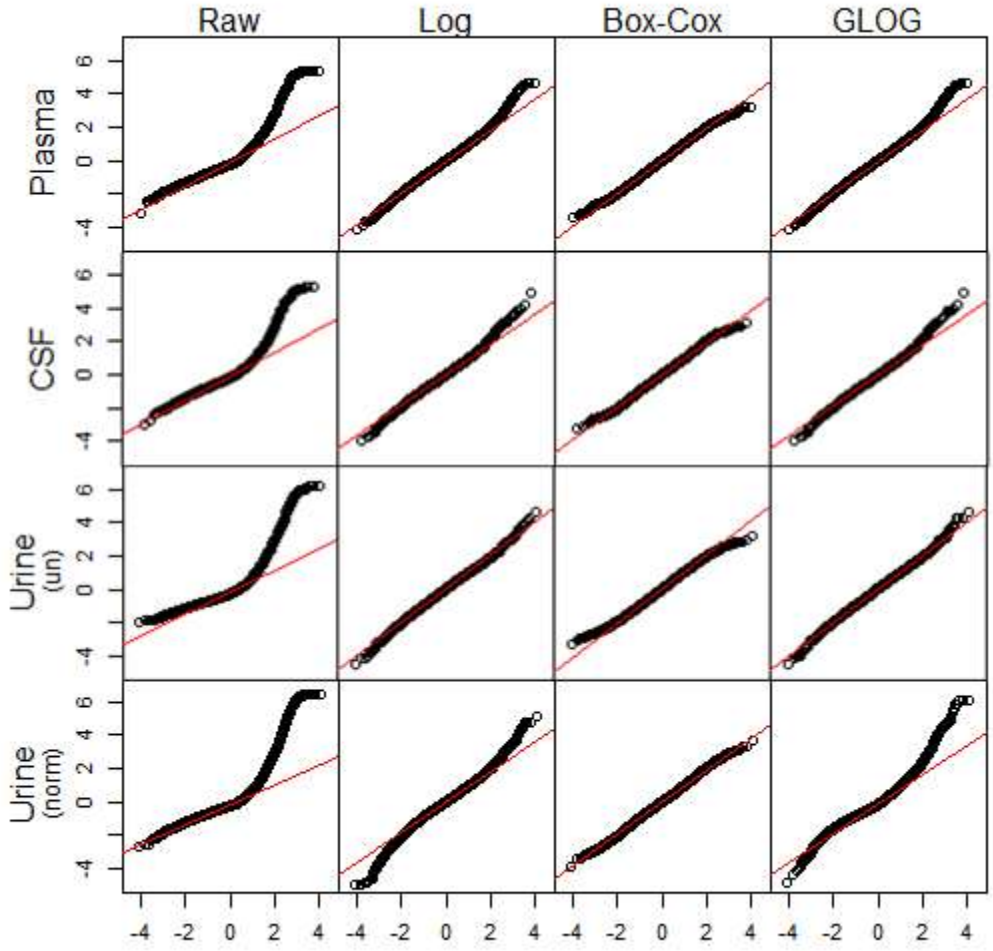


Figure 2.5: Combined quantile-quantile plots. Each point represents an observed ion count of a biochemical. Biochemical are independently centered and scaled prior to being combined.

matrices and that creatinine transformation actually increases both metrics.

As a visual approach to normality assessment, so called "combined" QQ plots are shown in *Figure 2.5*. These are made by taking each fully observed metabolite within the dataset, center and standardizing, merging together into a single variable and then plotting against the expected normal quantiles. The raw data demonstrates a severe right skewed distribution in all four data types as one might expect. Application of the log transformation significantly reduces skeweness in all four data types, producing a more symmetric set of values. Discrepancies from normality persist, mostly in the tails with the effect being strongest in plasma. These departures occur roughly in theoretical quantiles below -2 and above 2 and in the form of more extreme values

than expected. Together this implies longer tails in the lowest 5th and highest 5th percentiles. GLOG is extremely similar to the log, with the only noticeable deviations from log occurring in the normalized urine set in the form of GLOG having more extreme tails. As one would expect based on the Shapiro-Wilk p-values, a power transformation based on the Box-Cox induces almost perfect agreement with the normal theoretical quantiles throughout.

There is one particular oddity in the QQ plots regarding Box-Cox and normalized urine. The quantiles for normalized urine have the strongest agreement with the theoretical quantiles for Box-Cox of any of the four data sets, surprising since the Shapiro-Wilk rejection (Table 2.4) was highest in this data type. Generally speaking, these QQ plots do not change the impression but serve to further reinforce information from the summary measures in visual way. However, note that in the QQ plots, Box-Cox generally produces less extreme tails than expected under the normal distribution or leptokurtosis. Recall, from *Figure 2.4*, that urine tends to be more platykurtic than the other sample types and that normalization enhances this property. Indeed, extreme tailing is most present in the normalized data for all transformation types. Most likely there are a subset of biochemicals which remain platykurtic even after transformation, but since Box-Cox tends to move the other features toward leptokurtosis, this dichotomy is masked when all the observations are plotted together.

Next, we move on to biochemicals with missing values. The assumption here is that missing values are limit of detection, which would make the observed set of values for a metabolite with missing values a left-censored sample. Using Shapiro-Wilk, the observed values are tested for being a left-censored Gaussian sample. A minimum observation rate of at least 20% was imposed. Table 2.5 shows the Shapiro-Wilk rejection rate for these biochemicals. Box-Cox requires the entire sample to be observed and is not shown here. Estimation of the $\lambda_i's$ in a

Table 2.5: Shapiro-Wilk rejection rate in metabolites with missing values. Includes biochemicals with $20\% \leq$ observed proportion $< 100\%$.

		Plasma		CSF		Urine (un)		Urine (norm)	
# Compounds		421		179		695		695	
# Rejected	Raw	260	61.8%	179	43.6%	568	81.7%	551	79.3%
	Log	117	27.8%	53	29.6%	259	37.3%	284	40.9%
	glog	120	28.5%	52	29.1%	202	29.1%	492	70.8%

censored sample would attempt to transform the observed values to a normal distribution rather than fitting to the portion of the normal curve observable through the sample. Log and GLOG, when estimating a single pair (α_{gl} , λ_{gl}) across the entire dataset, do not have this limitation. The rejection rates in the biochemicals with missing values have a similar profile to the fully observed biochemicals for both plasma and CSF. The raw data performs a little better here, especially in CSF, but both a log and GLOG transformation reduce the rejection rate down into the mid- to high 20%s. Profile in the normalized urine is also broadly similar – very high rejection in the raw data with the log bringing it down only to 40% while GLOG offers only marginal improvement over the raw data. The un-normalized urine is a bit different in these missing biochemicals with log and GLOG doing worse than in the fully observed biochemicals.

The consensus from the results so far suggests that using a log-family transformation is largely effective at inducing normality, but some metabolites persist in being non-normal. In an attempt to understand why, a few variables immediately come forth as plausible candidates: proportion of missing values, abundance level, molecular mass and biochemical function. The influence of these factors on normality is examined via the Shapiro-Wilk p-values in the raw data with the first three examined in plots of *Figure 2.6*. Abundance is taken as the average of the (observable) ion counts. Due to the large spread in both mean abundance and p-value, a $\log_{10}$

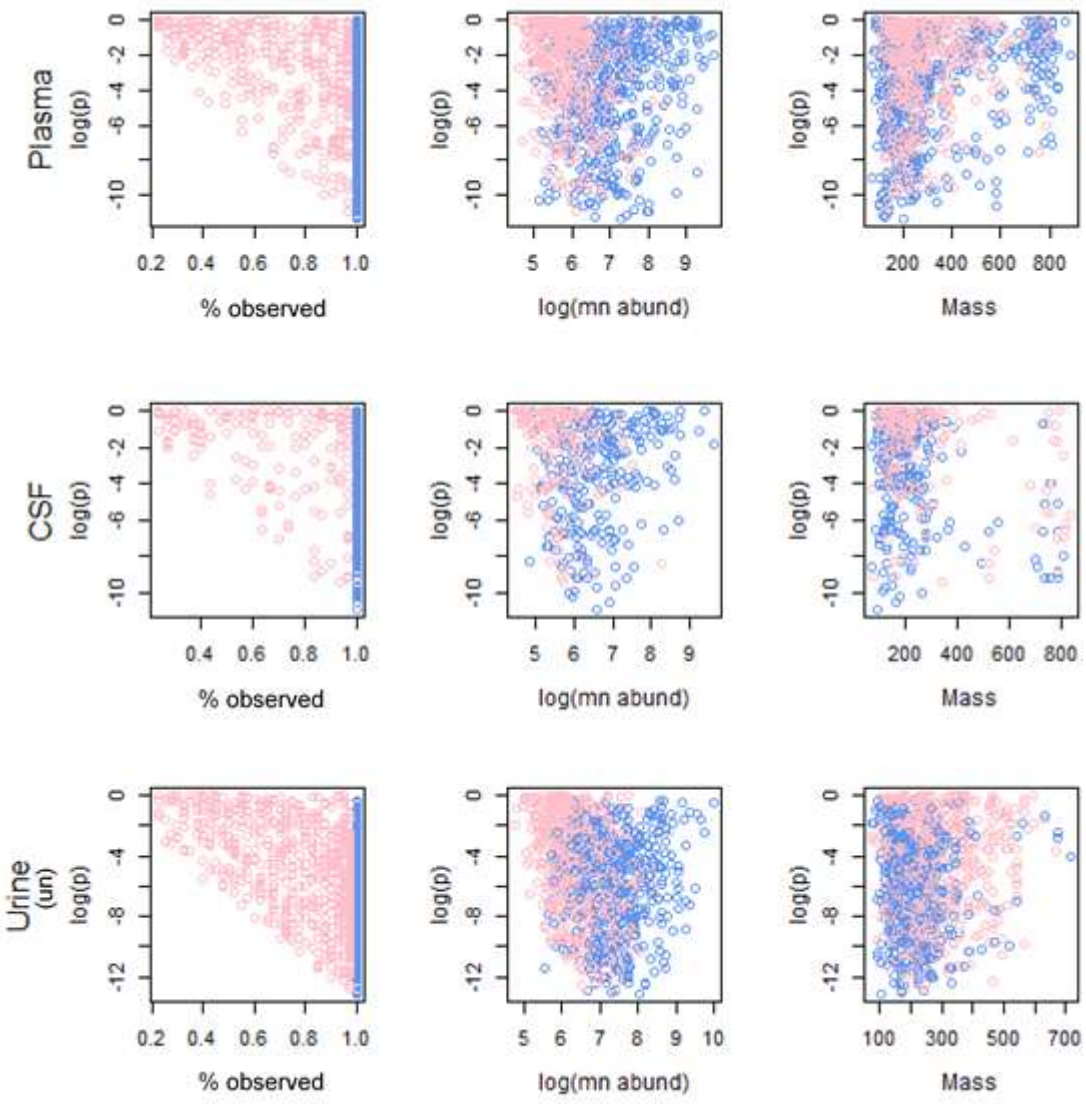


Figure 2.6: P-value by various biochemical characteristics. Blue indicates biochemicals fully observed while pink indicates the presence of missing values. Both p-value and mean abundance are on $\log_{10}$ scale. Abundance is based upon observed values. Chemical mass is measured in grams / mol.

transformation is used in the plots. With the rejection rate profiles being so similar between compounds with and without missing values, it suggests that detection does not appear to be a major factor in the distributional behavior. The plots, however, reveal that as the observable proportion decreases as the p-value increases. This negative correlation is likely the result of fewer samples leading to lower power, and hence explaining why the rejection rate for the raw data tended to be a little bit lower for the metabolites with missing values.

There does appear to be some slight connections with mean abundance, based on observed values, and the p-values. In plasma and CSF, the highest abundant compounds (mean abundance 10^8 and up) tend towards higher p-values. Below this level the p-values are quite scattered though, and no pattern is apparent in the urine even at the highest abundance levels. It's worth pointing out that in the abundance plots the blue points tend to be on the right and the pink points are on the left. This fits with that observed in *Figure 2.1* showing that more abundant compounds tend to have fewer missing values.

The final plot frame profiling the p-value against the molecular mass of the metabolite does not reveal much of a relationship either. There is perhaps some evidence that p-values are higher at mass levels around 600 Da and up, but certainly no relationship below 500 Da. One item these plots do show is that, unlike abundance, the proportion of missing values is not related to molecular mass. The blue and pink points are dispersed evenly throughout these plots in all three matrices. It is also apparent from the CSF plot that this matrix is made up of lower molecular weight compounds than either plasma or urine.

Finally, Table 2.6 gives the proportion of Shapiro Wilk rejected biochemicals by pathway designation using the biochemicals that are fully observed. The idea here is to see if any pathway or subset of pathways behaves differently from the others. Are some pathways more normal others, or are do the non-normal biochemicals tend to come from certain classes? Unfortunately, certain pathways are rather small in these sets, such as Energy metabolism which only has 6 members in plasma and 9 in both CSF and urine. So, the table certainly lacks for precise estimates of the proportions and makes formal testing challenging. But in many cases, it is still useful for observational purposes if nothing else. The proportion of rejected compounds, for example, is consistently high across all pathways and all matrix types in the raw data. There may

Table 2.6: Shapiro-Wilk rejection rate by Pathway. Rejection based on $p < .05$.

Matrix	Pathway	M	Raw		Log		Box-Cox		Glog	
Plasma	Amino Acid	117	103	88.0%	40	34.2%	10	8.5%	40	34.2%
	Carbohydrate	16	13	81.3%	5	31.3%	1	6.3%	5	31.3%
	Cofact and Vit	15	11	73.3%	5	33.3%	0	0.0%	5	33.3%
	Energy	6	6	100.0%	1	16.7%	0	0.0%	1	16.7%
	Lipid	222	153	68.9%	31	14.0%	6	2.7%	31	14.0%
	Nucleotide	18	13	72.2%	6	33.3%	0	0.0%	6	33.3%
	Peptide	16	13	81.3%	3	18.8%	1	6.3%	3	18.8%
	Xenobiotics	16	12	75.0%	5	31.3%	0	0.0%	6	37.5%
	Unknown	116	92	79.3%	39	33.6%	5	4.3%	39	33.6%
CSF	Amino Acid	93	59	63.4%	29	31.2%	10	10.8%	15	16.1%
	Carbohydrate	15	7	46.7%	3	20.0%	1	6.7%	4	26.7%
	Cofact and Vit	8	6	75.0%	3	37.5%	1	12.5%	3	37.5%
	Energy	9	7	77.8%	3	33.3%	1	11.1%	1	11.1%
	Lipid	41	34	82.9%	12	29.3%	2	4.9%	14	34.1%
	Nucleotide	23	15	65.2%	5	21.7%	0	0.0%	1	4.3%
	Peptide	8	8	100.0%	1	12.5%	0	0.0%	1	12.5%
	Xenobiotics	10	8	80.0%	4	40.0%	2	20.0%	5	50.0%
	Unknown	50	37	74.0%	15	30.0%	4	8.0%	12	24.0%
Urine (un)	Amino Acid	139	132	95.0%	45	32.4%	5	3.6%	31	22.3%
	Carbohydrate	20	17	85.0%	1	5.0%	0	0.0%	0	0.0%
	Cofact and Vit	17	17	100.0%	5	29.4%	4	23.5%	5	29.4%
	Energy	9	6	66.7%	4	44.4%	1	11.1%	4	44.4%
	Lipid	36	33	91.7%	11	30.6%	1	2.8%	9	25.0%
	Nucleotide	33	31	93.9%	7	21.2%	0	0.0%	6	18.2%
	Peptide	17	17	100.0%	3	17.6%	1	5.9%	3	17.6%
	Xenobiotics	30	29	96.7%	10	33.3%	3	10.0%	9	30.0%
	Unknown	189	188	99.5%	46	24.3%	12	6.3%	32	16.9%
Urine (norm)	Amino Acid	138	125	90.6%	69	50.0%	16	11.6%	74	53.6%
	Carbohydrate	20	17	85.0%	4	20.0%	1	5.0%	11	55.0%
	Cofact and Vit	17	17	100.0%	7	41.2%	4	23.5%	9	52.9%
	Energy	9	5	55.6%	3	33.3%	0	0.0%	3	33.3%
	Lipid	36	36	100.0%	15	41.7%	1	2.8%	24	66.7%
	Nucleotide	33	26	78.8%	17	51.5%	3	9.1%	19	57.6%
	Peptide	17	14	82.4%	8	47.1%	2	11.8%	10	58.8%
	Xenobiotics	30	29	96.7%	12	40.0%	5	16.7%	23	76.7%
	Unknown	189	183	96.8%	64	33.9%	15	7.9%	133	70.4%

some specific differences, like in plasma where the difference between amino acids and lipids, both of whom are large members in the matrix, is nearly 20% (88.0% vs 68.9%). Similarly, in CSF carbohydrates (46.7%) and nucleotides (65.2%) are quite a bit lower than lipids (82.9%). This not only suggests some differences between the metabolite pathways but also between the matrices themselves, and further demonstrates that while many biochemicals may be present in multiple matrices, their behavior can be quite different. See the lipids, which in the raw data has the lowest rejection rate in plasma but the second highest rejection rate in CSF. Regardless of the specific pathway comparisons, the proportion of rejected compounds is quite high across all the raw data. In every single case, a natural log transformation reduces the proportion of rejected biochemicals from the raw, and Box-Cox always reduces the proportion even further. Overall improvement with either transformation is therefore not attributed to any specific class of biochemicals. All pathways see improvement and the response is consistent across all pathways.

2.7.2. Correlation

For the four data types, correlations were examined between each pairwise combination of metabolites. Three different versions of correlations are presented: (1) Pearson R using raw data, (2) Pearson R using natural log transform and (3) Spearman Rho. While all three versions have been used, (2) and (3) are recommended due to the propensity for large outliers in the data [9]. Data is filtered in two ways based on proportion of missing values. The first way is to restrict to only those metabolites that are fully observed. This still leads to a myriad of pairings with 146,611 distinct pairs in plasma, 32,896 pairs in CSF and 119,805 pairs in urine (119,316 when normalizing to creatinine). The second way is to take all compounds that are at least 10% observed. This leads to a total number of distinct pairs in which at least two samples are observed in both pairs as 484,182 in plasma, 105,706 in CSF, 735,357 in un-normalized urine

Table 2.7: Summaries of pairwise correlations based on Pearson r.

Data	% Fill	Type	Percentile						
			Min	10	25	50	75	90	Max
Plasma	100	raw	-0.774	-0.205	-0.100	0.036	0.211	0.395	0.999
		log	-0.785	-0.214	-0.068	0.091	0.256	0.407	0.996
		spearman	-0.803	-0.209	-0.063	0.099	0.259	0.404	0.986
	>10	raw	-1.000	-0.281	-0.151	0.002	0.206	0.439	1.000
CSF	100	raw	-0.669	-0.205	-0.062	0.128	0.356	0.575	0.998
		log	-0.710	-0.232	-0.052	0.167	0.375	0.558	0.987
		spearman	-0.729	-0.215	-0.042	0.163	0.366	0.549	0.981
	>10	raw	-1.000	-0.297	-0.111	0.117	0.398	0.688	1.000
Urine (un)	100	raw	-0.658	-0.078	0.036	0.225	0.436	0.601	0.999
		log	-0.717	0.002	0.168	0.345	0.503	0.622	0.984
		spearman	-0.710	0.036	0.194	0.368	0.525	0.641	0.975
	>10	raw	-1.000	-0.163	-0.048	0.135	0.364	0.563	1.000
Urine (norm)	100	raw	-0.663	-0.143	-0.051	0.088	0.263	0.430	0.999
		log	-0.732	-0.145	0.008	0.177	0.336	0.471	0.985
		spearman	-0.700	-0.128	0.020	0.180	0.334	0.464	0.971
	>10	raw	-1.000	-0.211	-0.105	0.035	0.229	0.432	1.000

and 734,143 in normalized urine.

For biochemicals with missing values, correlation is based on the pairs that are present. Results are displayed in Table 2.7. For simplicity, only Pearson R values on the raw data are displayed for compounds observed in at least 10% of samples because, as seen in the table, correlations are very similar regardless of whether Pearson or Spearman is used or whether a log transformation is applied or not. Across the three datasets correlations are centered around 0.1. The exception to this is un-normalized urine in which a moderate correlation around 0.3 is more typical. Plasma and CSF are, once again, very similar and creatinine normalization causes the urine to look more like those two. The inner quartile range of correlation coefficients for these datasets is rough between -0.1 and 0.35 with urine and CSF being a little higher while plasma is a little lower. This shows the metabolites tend to be more positively correlated than negatively, and that the range of correlation for most pairs would be considered low to moderate in most

situations. Instances of extreme association do occur, however. In plasma, for example, the two isomers of bilirubin, the Z, Z and E, E forms, are highly correlated with each other (Pearson R = 0.958). Both are also correlated with biliverdin (ZZ Pearson R = 0.990, EE Pearson R = 0.927), which is the precursor metabolite to bilirubin in the breakdown of macrophages. And while negative correlations are less prevalent, there are still instances of strong negative associations as well. Note that when considering any metabolite with an observation rate of at least 10% there are several pairs with correlation of +/-1. This is because through the pairings some metabolites will have just two observed values in common, which results in a perfect positive or negative correlation depending on how the rankings match up. For this reason, it is more useful to focus on compounds that are completely observed, but from the table it is clear that considering metabolites with missing observations produces fairly similar values.

Figures 2.7-2.10 visually integrate Pearson R with pathway information in the form of a heatmap. This is essentially a matrix in which each cell corresponds to a pair of biochemicals, given by the row and column intersection, and the value is a color translated from the Pearson r value. A value of $r = 0$ is white. As correlations move towards 1 the color moves toward red. The diagonals are thus bright red reflecting the perfectly positive correlation a compound has with itself. Similarly, as the correlation moves towards -1 the color becomes more and more blue. The biochemicals were grouped according to pathway and the cells were shaded such that increasingly negative correlation is more and more blue, increasingly positive.

These maps show that though there are instances of very strong correlations, the strongest of which are almost exclusively positive, in general biochemicals are not strongly associated with each other. Furthermore, instances of strong correlations appear to be somewhat matrix specific. In plasma, strong, positive correlations tend to cluster around the diagonal. This is an indication

Figure 2.7: Plasma correlation heatmap. Coloring based on Pearson r coefficient with blue for negative values and red for positive. Each row and column represent a specific metabolite, which are ordered left to right, top to bottom according to pathway assignment.

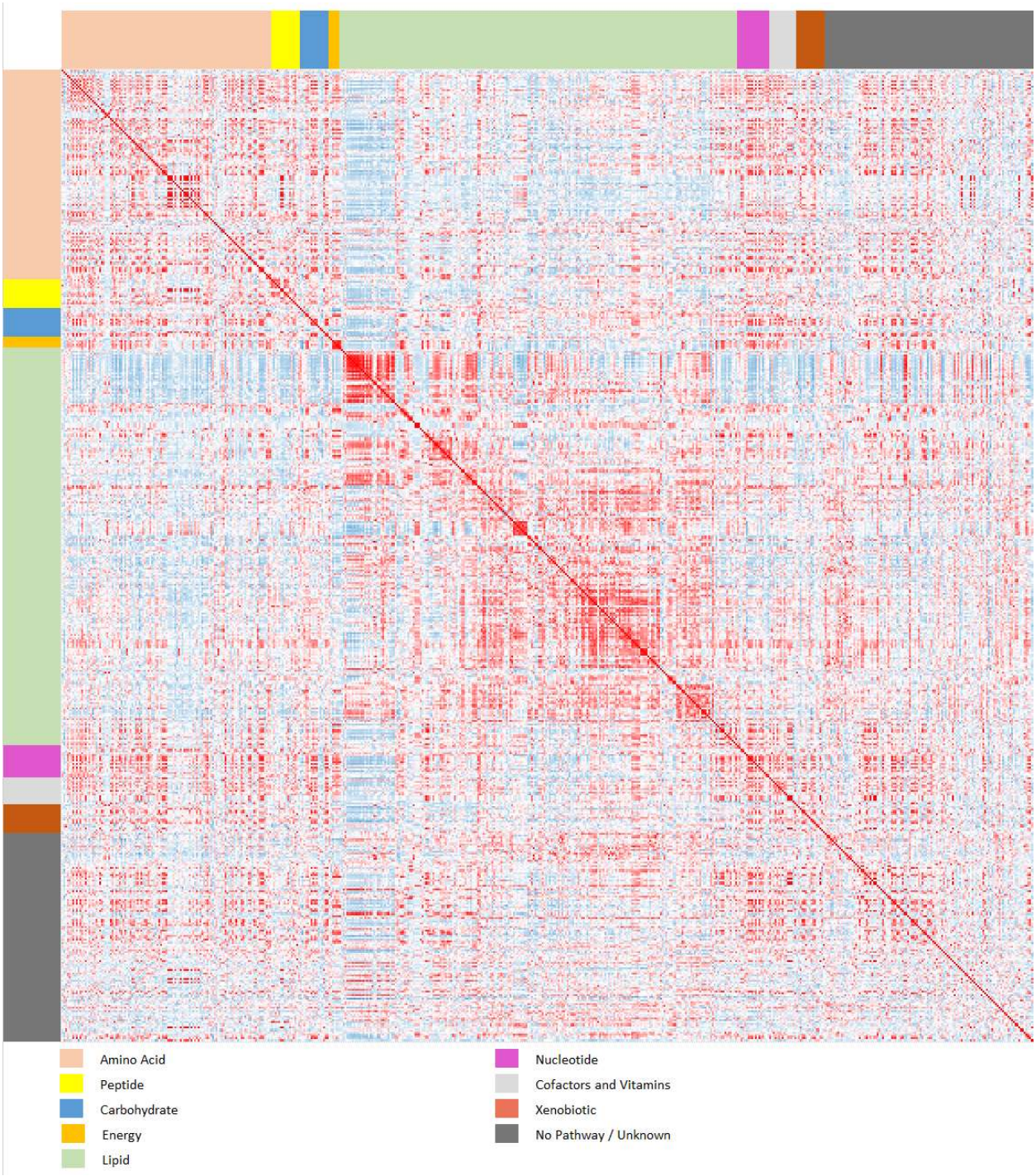


Figure 2.8: CSF correlation heatmap. Coloring based on Pearson r coefficient with blue for negative values and red for positive. Each row and column represent a specific metabolite, which are ordered left to right, top to bottom according to pathway assignment.

Figure 2.9: Urine (un) correlation heatmap. Coloring based on Pearson r coefficient with blue for negative values and red for positive. Each row and column represent a specific metabolite, which are ordered left to right, top to bottom according to pathway assignment.

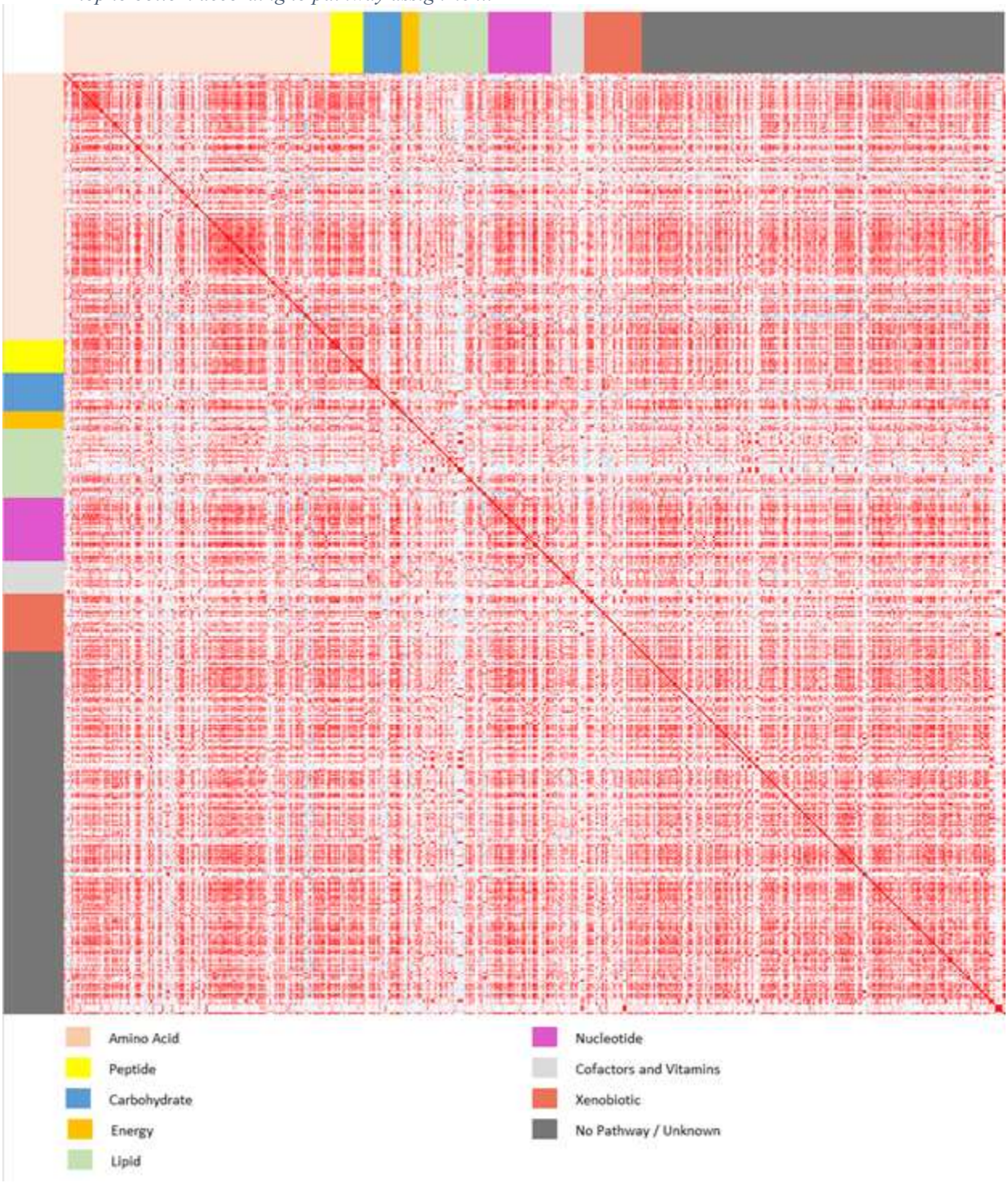
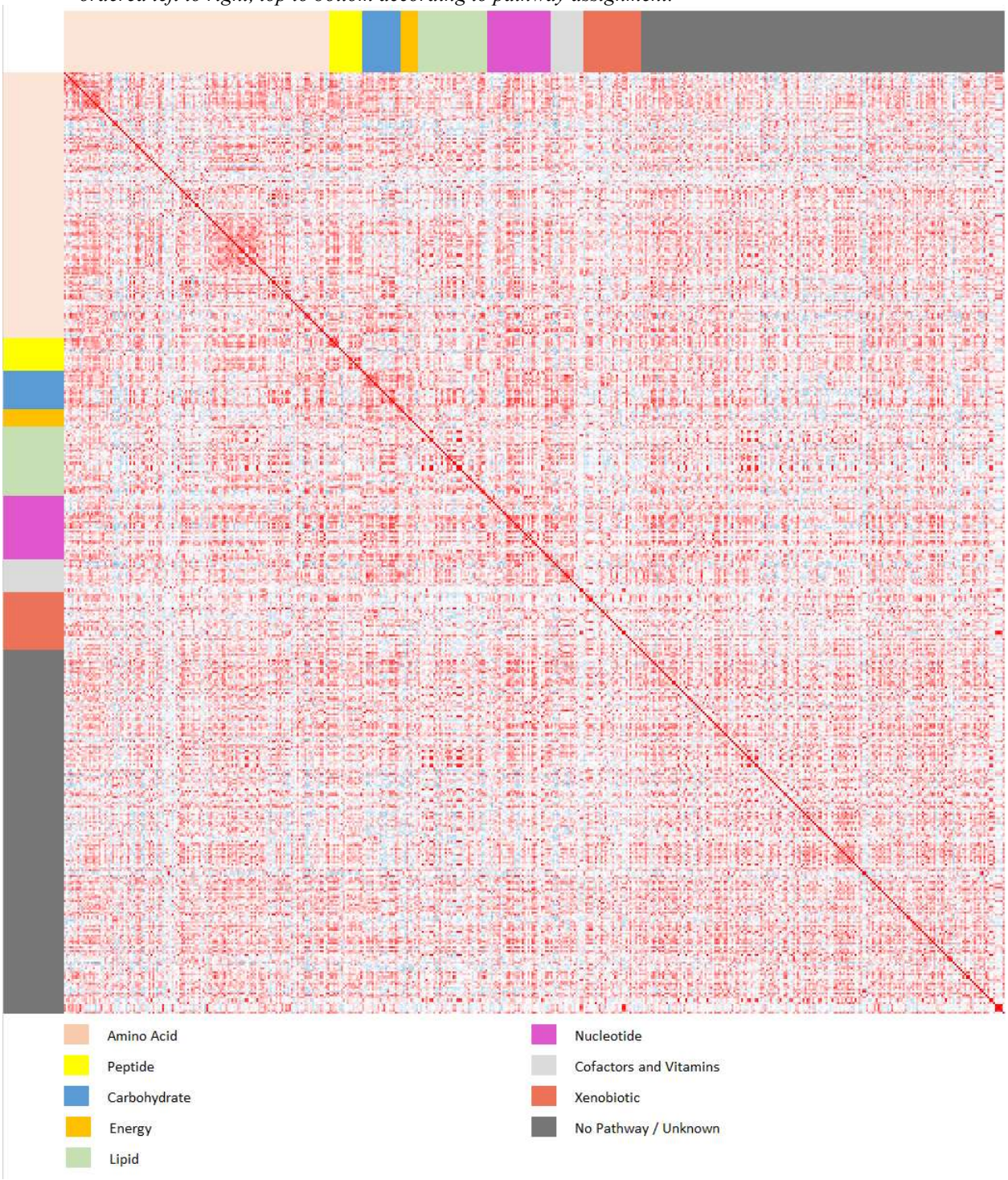


Figure 2.10: Urine (normalized) correlation heatmap. Coloring based on Pearson r coefficient with blue for negative values and red for positive. Each row and column represent a specific metabolite, which are ordered left to right, top to bottom according to pathway assignment.



that strong correlations are more likely to occur between metabolites within a pathway versus metabolites in different pathways. Case in point, in plasma the lipids do not correlate very strongly with non-lipids, and in fact several have a slight negative correlation with most other compounds, but some of the largest and strongest intensity blocks are within the lipid class. Yet, in the CSF and the un-normalized urine rather strong correlations are observed throughout the map. Creatinine normalization noticeably lowers the correlations across the entire urine map, but it remains almost exclusively red with noticeable patches spread throughout the map as opposed to clustering within pathway.

There is also some evidence of within pathway associations. This sub pathway behavior is most strongly seen in plasma where a subset of the lipids correlates strongly with each other but negatively correlated with most other biochemicals, including other lipids. Inspection of this group finds that members are comprised of medium and long chain fatty acids, branched chain fatty acids and polyunsaturated fatty acids. Lipid categories not found in this group include plasmalogens, phospholipids, monoacylglycerols and di-acyglycerols to name a few. This shows that even though metabolites may be chemically related, they need not share the same behavior. Sub setting is also seen within a group of amino acids in the top left most corner of normalized urine map.

Figures 2.11-2.14 provide another pathway view for correlations. Each node is an individual metabolite, clustered together in a circle by pathway and the pathway circles then arranged in a circular pattern like the numbers on a clock face. Two metabolites are connected when the Pearson R between the two metabolites exceeds the minimum threshold, which here is set to 0.7 in order emphasis the strongest relationships. Because negative correlations are rare, pathway figures for $r < 0$ are not provided. To further aid in interpretability, the unknowns and named

Figure 2.11: Plasma correlation network graph. Nodes are metabolites. Connecting lines indicate pairwise Pearson r coefficient ≥ .7. Pathways are presented a clockwise manner based on the order listed in the legend.

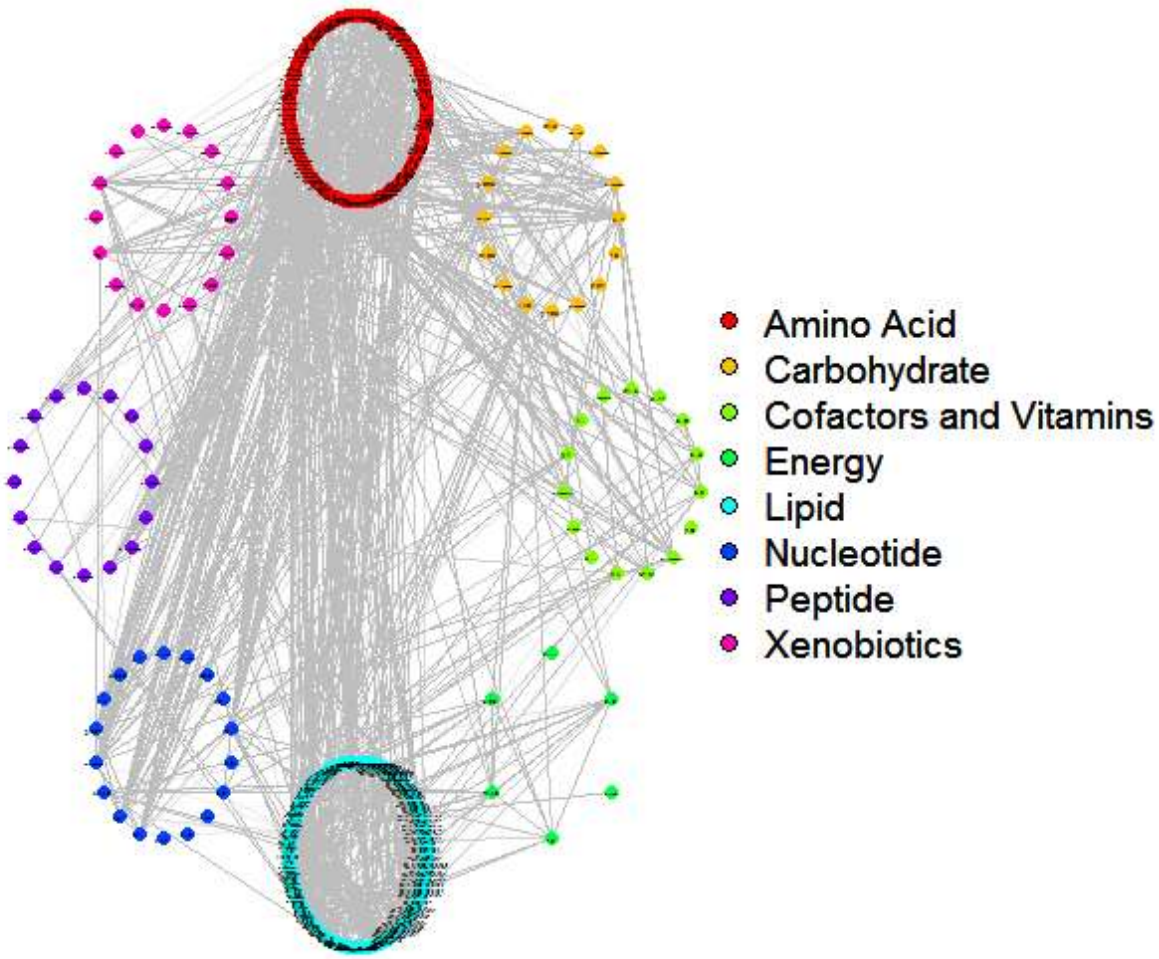


Figure 12: CSF correlation network graph. Nodes are metabolites. Connecting lines indicate pairwise Pearson r coefficient ≥ .7. Pathways are presented a clockwise manner based on the order listed in the legend.

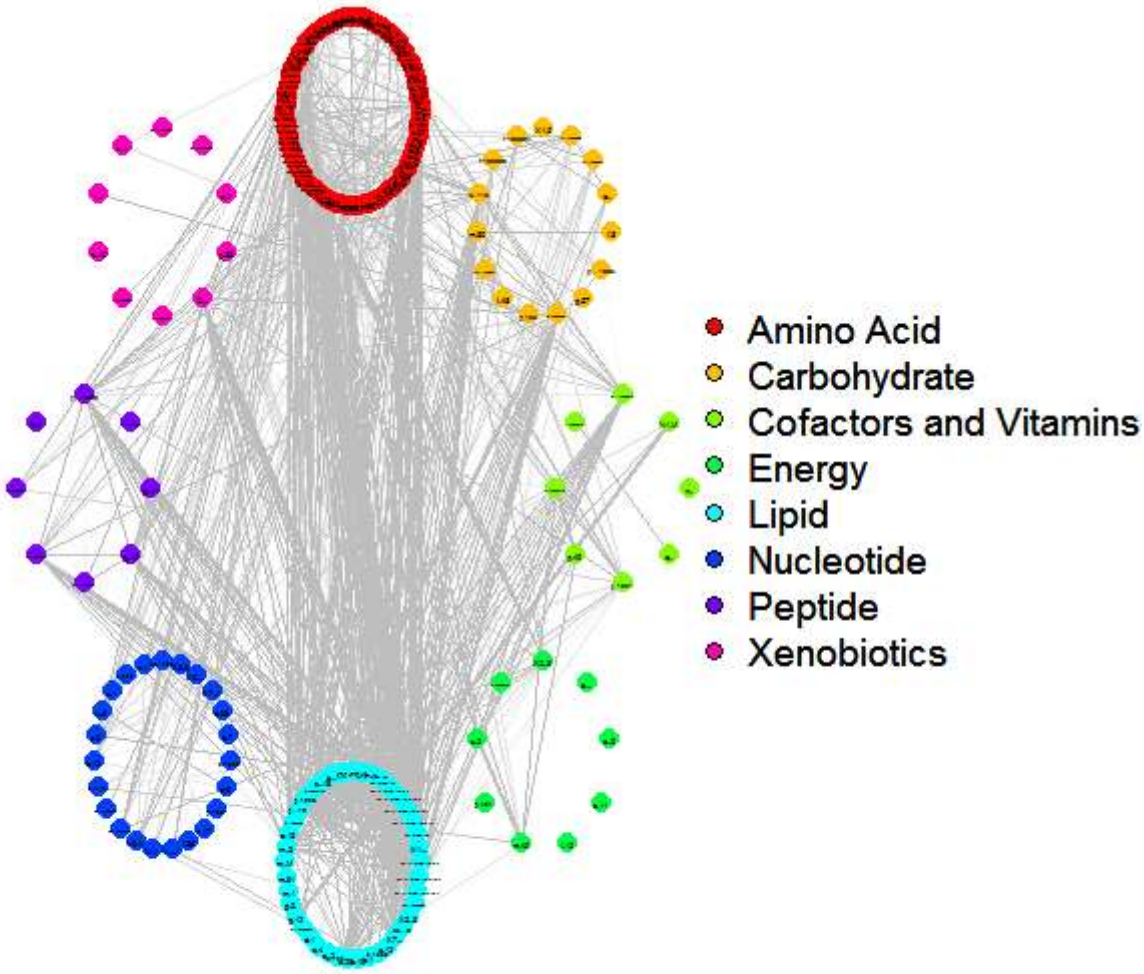


Figure 2.13: Urine (un) correlation network graph. Nodes are metabolites. Connecting lines indicate pairwise Pearson r coefficient ≥ .7. Pathways are presented a clockwise manner based on the order listed in the legend.

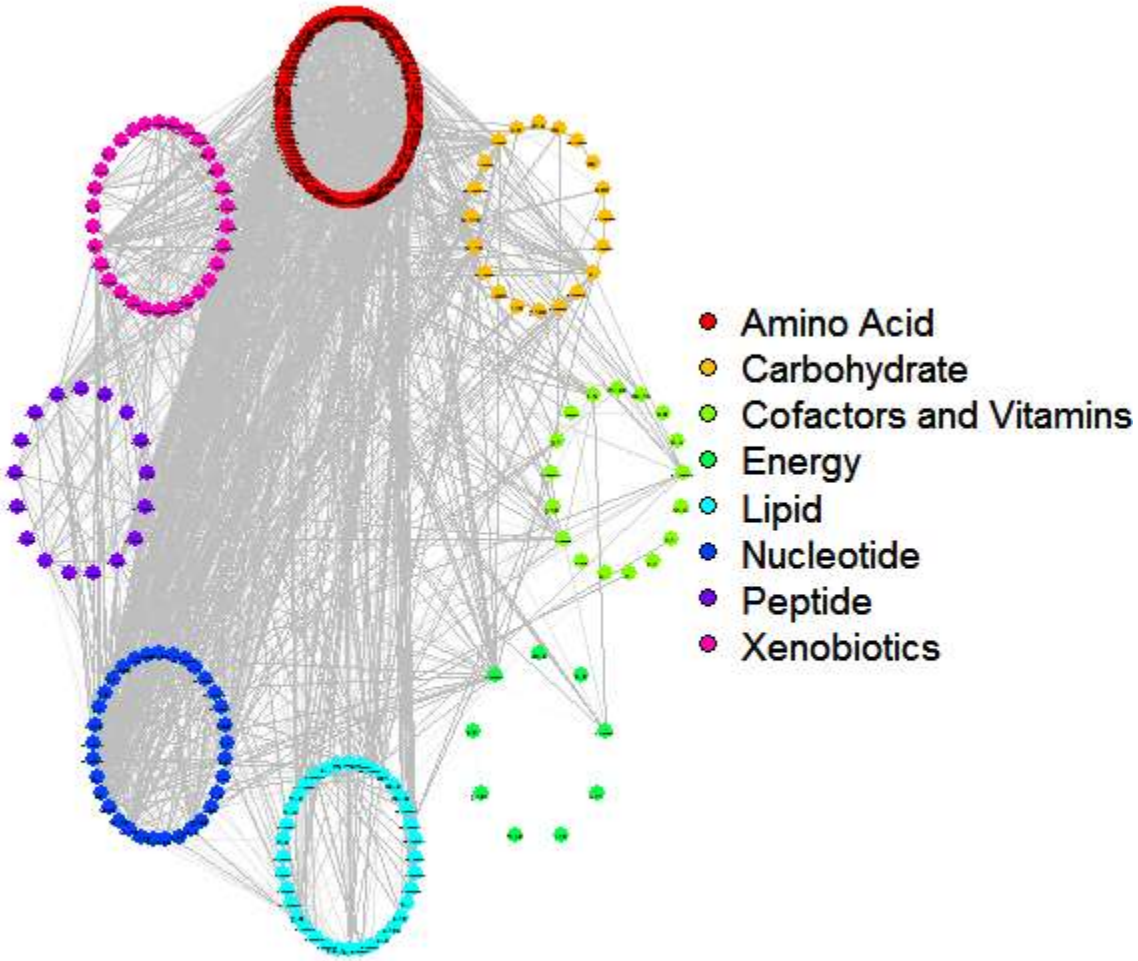
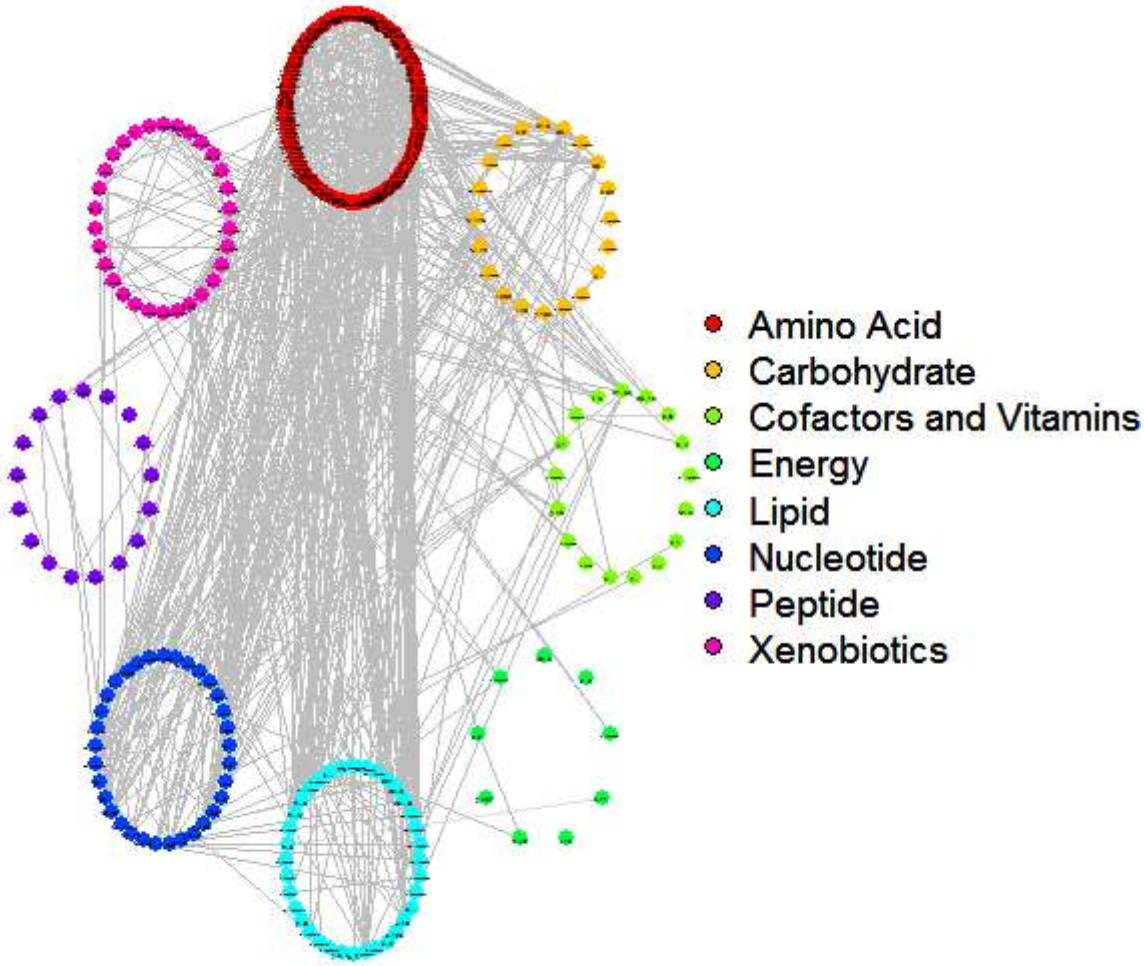


Figure 2.14: Urine (normalized) correlation network graph. Nodes are metabolites. Connecting lines indicate pairwise Pearson r coefficient ≥ .7. Pathways are presented a clockwise manner based on the order listed in the legend.



metabolites without a pathway are removed. A total of 426 metabolites are available in plasma for this graph. Although the number of lines may seem large at first, recall there are 90,525 pairwise combination and only 1,343 (1.5%) with correlation above 0.7. Amino acids as a whole, are strongly associated with lipids and nucleotides, and there is strong relationship within the amino acid and lipid pathways. But members of the other six pathways are not consistently related, and their relationships to other pathways are largely determined by individual members. This is probably most relevant to xenobiotics where behavior of any member is an amalgamation of diet, environment, microbiome, pharmacological behavior, etc. Associations with the metabolites may provide useful insight to the physiological function these biochemicals belong to.

The pathway plot for CSF is similar to plasma. Restricting to the pathway associated members gives 207 metabolites with a total of 21,321 pairs, of which 945 (4.4%) are above 0.7. Amino acids and lipids are again strongly associated both within and between pathway classifications, but as with plasma the other six pathways are not that correlated internally and their relationships with other pathways are largely member specific. Un-normalized urine, with 301 metabolites yielding a total of 45,150 pairs of which 1,730 (3.8%) are above 0.7, shows really strong association between amino acids and nucleotides. However, after creatinine normalization (300 metabolites – 44,850 pairs – 1,124 (2.5%) $\geq$ 0.7), the graph very much resembles plasma as far as pathway relationships are concerned.

2.8. Conclusions

This chapter demonstrates that ion counts of metabolites have a left skewed distribution for an overwhelming majority of features. The behavior is observed in three separate types of human material commonly used for clinical testing. A natural log transformation is largely effective at

removing skewness and inducing normality; however, it is not perfect with around a quarter of metabolites still exhibit statistically significant departures from normality. The power transformation can reduce this the proportion down to near the null level but requires estimating the $\lambda_i's$ for each metabolite. This is somewhat of a nuisance as the semi-quantitative nature of the instrumentation means that these parameters are likely to change from run to run, and is complicated by metabolites with missing values, which may be non-ignorable. The natural log is a static transformation and avoids these sorts of issues. Aggregate quantile-quantile plots point to this departure being mostly in the extreme tails. This may be due to the reference population being composed of non-IEM suspects that initially suspected of having a metabolic disorder, implying some level of poor health and the label of "healthy" being somewhat misleading. It is possible that some subjects were mis-diagnosed, given the low diagnosis rate for IEMs, or that they have non-IEM disease which still impacts their metabolic profile. Since the subjects are independent with likely different diseases, only a small fraction of any given metabolite would be affected. Hence, it may be that metabolites truly are log-normal and intentionally censoring the upper and lower 5% is sufficient to adjust for diseased subjects when the diseases are not concentrated on a specific disorder.

The second point of this chapter also shows that when the data are resolved in a chemo-centric fashion, global metabolomics is not overly correlated. In general features are at most mildly correlated with an average Pearson correlation coefficient around 0.3 and instances of moderate to high correlations, consider here to be values of 0.5 and up, are rare. An examination of biochemical pathways found that certain pathways are more correlated than others, most notably the amino acids and lipids. Thus, metabolomic datasets that concentrate on certain classes of biochemicals may exhibit higher correlation, but in global approach doesn't appear to

elevate correlations much overall.

These results are useful for general understanding of global LC-MS metabolomics and also have significant implications for statistical analysis. Because many studies employ a low sample size, parametric hypothesis testing is preferable to maximize power. Using a log transformation or other similar approach is largely helpful at satisfying this pre-requisite. Correlation is an important characteristic for many multivariate tests including Hotelling's T Square and Principle Component Analysis as well as network methods such as Gaussian Graphical Models. Knowledge of how metabolomic sets structure is useful for maximizing statistical analysis.

REFERENCES

[1] M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart, and O. Yanes, "A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data," *Metabolites,* vol. 2, no. 4, pp. 775-95, 2012.

[2] T. Obata and A. R. Fernie, "The use of metabolomics to dissect plant responses to abiotic stresses," *Cell Mol Life Sci,* vol. 69, no. 19, pp. 3225-43, Oct 2012.

[3] E. G. Armitage and C. Barbas, "Metabolomics in cancer biomarker discovery: current trends and future perspectives," *J Pharm Biomed Anal,* vol. 87, pp. 1-11, Jan 2014.

[4] N. T. Quan *et al.*, "Involvement of Secondary Metabolites in Response to Drought Stress of Rice," *Agriculture,* vol. 6, no. 2, p. 23, 2016.

[5] P. Manini, G. De Palma, R. Andreoli, M. Goldoni, and A. Mutti, "Determination of urinary styrene metabolites in the general Italian population by liquid chromatography-tandem mass spectrometry," *Int Arch Occup Environ Health,* vol. 77, no. 6, pp. 433-6, Aug 2004.

[6] T. H. Herdt, J. B. Stevens, W. G. Olson, and V. Larson, "Blood concentrations of beta hydroxybutyrate in clinically normal Holstein-Friesian herds and in those with a high prevalence of clinical ketosis," *Am J Vet Res,* vol. 42, no. 3, pp. 503-6, Mar 1981.

[7] J. D. Clarke *et al.*, "Assessment of genetically modified soybean in relation to natural variation in the soybean seed metabolome," *Sci Rep,* vol. 3, p. 3082, Oct 2013.

[8] A. M. Evans, M. W. Mitchell, H. Dai, and C. D. DeHaven, "Categorizing Ion-Features in Liquid Chromatography/Mass Spectrometry Metabolomics Data," *Journal of Metabolomics,* vol. 2, no. 3, 2012.

[9] A. V. Frane, "Power and Type I Error Control for Univariate Comparisons in Multivariate Two-Group Designs," *Multivariate Behav Res,* vol. 50, no. 2, pp. 233-47, 2015 Mar-Apr 2015.

[10] H. Zou and T. Hastie, "Regularization and variable selection via the Elastic Net," vol. 67, no. 2, pp. 301-320, 2005.

[11] D. Camacho, A. de la Fuente, and P. Mendes, "The origin of correlations in metabolomics data," *Metabolomics,* vol. 1, no. 1, pp. 53-63, 2005.

[12] C. D. Dehaven, A. M. Evans, H. Dai, and K. A. Lawton, "Organization of GC/MS and LC/MS metabolomics data into chemical libraries," *J Cheminform,* vol. 2, no. 1, p. 9, Oct 2010.

[13] A. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *G. Ist. Ital. Attuari,* vol. 4, pp. 83-91, 1930.

[14] A. K. Bera, A. F. Galvao, L. Wang, and Z. Xiao, "A new characterization of the normal distribution and test for normality," *Econometric Theory,* pp. 1-37, 2015.

[15] L. T. DeCarlo, "On the Meaning and Use of Kurtosis," *Psychological Methods,* vol. 2, no. 3, pp. 292-307, 1997.

[16] Z. Liang *et al.*, "The Statistical Meaning of Kurtosis and Its New Application to Identification of Persons Based on Seismic Signals," *Sensors (Basel),* vol. 8, no. 8, pp. 5106-5119, Aug 2008.

[17] A. L. Bowley, *Elements of statistics*, 2d ed. London, New York,: P. S. King & son; C. Scribner's sons, 1902, pp. viii p., 2 l.,.

[18] R. A. Groeneveld and G. Meeden, "Measuring Skewness and Kurtosis," *Journal of the Royal Statistical Society. Series D,* vol. 33, no. 4, pp. 391-399, 1984.

[19] M. K. Cain, Z. Zhang, and K. H. Yuan, "Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation," *Behav Res Methods,* vol. 49, no. 5, pp. 1716-1735, Oct 2017.

[20] J. C. W. Rayner, D. J. Best, and K. L. Mathews, "INTERPRETING THE SKEWNESS COEFFICIENT," *Communications in statistics. Theory and methods,* vol. 24, no. 3, pp. 593-600, 1995.

[21] D. Doric, E. Nikolic-Doric, V. Jevremovic, and J. Malisic, "On measuring skewness and kurtosis," *Quality and Quantity,* vol. 43, no. 3, pp. 481-493, 2009.

[22] R. Horswell and S. Looney, "Diagnostic limitations of skewness coefficients in assessing departures from univariate and multivariate normality," *Communications in statistics. Simulation and computation,* vol. 22, no. 2, pp. 437-459, 01/1993 1993.

[23] G. Abel, "Is formal normality testing a waste of time?," in *Methods and Methodology*, ed: Cambridge Centre for Health Services Research, 2013.

[24] M. J. Campbell and T. D. V. Swinscow, *Statistics at square one*, 11th ed. Chichester, UK ; Hoboken, NJ: Wiley-Blackwell/BMJ Books, 2009, pp. iv, 188 p.

[25] L. Baringhaus, R. Danschke, and N. Henze, "Recent and classical tests for normality - a comparative study," *Communications in Statistics - Simulation and Computation,* vol. 18, no. 1, pp. 363-379, 1989.

[26] H. Cramér, "On the composition of elementary errors," *Scandinavian Actuarial Journal,* vol. 1928, no. 1, 1928.

[27] R. von Mises, "Wahrscheinlichkeitsrechnung und ihre Anwendungen," *der Statistik und theoretischen Physik,* 1931.

[28] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *G. Ist. Ital. Attuari,* vol. 4, pp. 83-91, 1933.

[29] A. N. Kolmogoroff, "Confidence limits for an unknown distribution function," *Annals of Mathematical Statistics,* vol. 12, pp. 461-463, 1941.

[30] N. V. Smirnoff, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Bul. Math. de l'Univ. de Moscou,* vol. 2, pp. 3-14, 1939.

[31] N. V. Smirnoff, "Approximate laws of distribution of random variables from empirical data," *Uspekhi Mat. Nauk,* vol. 10, pp. 179-206, 1944.

[32] G. J. Filion, "The signed Kolmogorov-Smirnov test: why it should not be used," *Gigascience,* vol. 4, p. 9, 2015.

[33] N. M. Razali and Y. B. Way, "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," *Journal of Statistical Modeling and Analytics,* vol. 2, no. 1, pp. 21-33, 2011.

[34] H. W. Lilliefors, "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association,* vol. 62, no. 318, pp. 399-402, 1967.

[35] T. W. Anderson and D. A. Darling, "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes," *Annals of Mathematical Statistics,* vol. 23, pp. 193-212, 1952.

[36] O. Vasicek, "A Test for Normality Based on Sample Entropy," *Journal of the Royal Statistical Society. Series B (Methodological_,* vol. 38, no. 1, pp. 54-59, 1976.

[37] T. Nafee *et al.* WikiDoc The Living Textbook of Medicine [Online]. Available: https://www.wikidoc.org/index.php/Anderson-Darling_test

[38] M. de Smith, *Statistical Analysis Handbook: a comprehensive handbook of statistical concepts, techniques and software tools*. 2015.

[39] T. P. Royston, "An Extension of Shapiro and Wilk W Test for Normality to Large Samples," *Journal of the Royal Statistical Society. Series C (Applied Statistics),* vol. 31, no. 2, pp. 115-124, 1982.

[40] J. P. Royston, "Some Techniques for Assessing Multivariate Normality Based on the Shapiro-Wilk W," *Journal of the Royal Statistical Society. Series C (Applied Statistics,* vol. 32, no. 2, pp. 121-133, 1983.

[41] M. M. Rahman, "A modification of the test of Shapiro and Wilk for normality," *Journal of Applied Statistics,* vol. 24, no. 2, pp. 219-236, 1997.

[42] S. Verrill and R. A. Johnson, "Tables and Large-Sample Distribution Theory for Censored Data Correlation Statistics for Testing Normality," *Journal of the American Statistical Association,* vol. 83, no. 404, pp. 1192-1197, 1988.

[43] E. Seier, "Comparison of tests of univariate normality," *InterStat Statistical Journal,* vol. 1, pp. 1-17, 2002.

[44] K. Pearson, "Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material," *Philosophical Transactions of the Royal Society of London,* vol. 186, pp. 343-414, 1895.

[45] R. B. D'Agostino, "Transformation to Normality of the Null Distribution of g1," *Biometrika,* vol. 57, no. 3, pp. 679-681, 1970.

[46] F. J. Anscombe and W. J. Glynn, "Distribution of the Kurtosis Statistic b2 for Normal Samples," *Biometrika,* vol. 70, no. 1, pp. 227-234, 1983.

[47] J. Hosking, "L-moments: Analysis and estimation of distributions using linear combinations of order statistics," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 52, no. 1, pp. 105-124, 1990.

[48] C. Lin and G. Mudholkar, *Biometrika,* vol. 67, no. 2, pp. 455-461, 1980.

[49] G. S. Mudholkar, C. E. Marchetti, and C. T. Lin, "Independence characterizations and testing normality against restricted skewness–kurtosis alternatives," *Journal of statistical planning and inference,* vol. 104, no. 2, pp. 485-501, 2002.

[50] R. D'Agostino and E. S. Pearson, "Tests for Departure from Normality. Empirical Results for the Distributions of b2 and √ b1," *Biometrika,* vol. 60, no. 3, pp. 613-622, 1973.

[51] K. Bowman and L. Shenton, "Omnibus Test Contours for Departures from Normality Based on and b2," *Biometrika,* vol. 62, no. 2, 1975.

[52] A. K. Bera and C. M. Jarque, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo Evidence," *Economics Letters,* vol. 7, no. 4, pp. 313-318, 1981.

[53] J. Bai and S. Ng, "Tests for Skewness, Kurtosis, and Normality for Time Series Data," *Journal of Business & Economic Statistics,* vol. 23, no. 1, pp. 49-60, 2005.

[54] R. B. D'Agostino, A. Belanger, and R. B. D'Agostino, Jr., "A Suggestion for Using Powerful and Informative Tests of Normality," *The American Statistician,* vol. 44, no. 4, pp. 316-321, 1990.

[55] I. Koutrouvelis and J. Kellermeier, "A Goodness-of-Fit Test of Simple Hypotheses Based on the Empirical Characteristic Function," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 43, no. 2, pp. 173-176, 1981.

[56] K. Murota and K. Takeuchi, "The studentized empirical characteristic function and its application to test for the shape of a distribution," *Biometrika,* vol. 68, no. 1, pp. 55-65, 1980.

[57] C. J. Mecklin and D. J. Mundfrom, "An Appraisal and Bibliography of Tests for Multivariate Normality," *International Statistical Review,* vol. 72, no. 1, pp. 123-138, 2004.

[58] S. Csorgo, "Testing for Normality in Arbitrary Dimension," *The Annals of Statistics,* vol. 14, no. 2, pp. 708-723, 1986.

[59] L. Baringhaus and N. Henze, "A consistent test for multivariate normality based on the empirical characteristic function," *Metrika,* vol. 35, no. 1, pp. 339-348, 1988.

[60] M. Arcones, "Two tests for multivariate normality based on the characteristic function," *Mathematical Methods of Statistics,* vol. 16, no. 3, pp. 177-201, 2007.

[61] N. M. Keifer and M. Salmon, "Testing normality in econometric models," *Economics Letters,* vol. 11, no. 1-2, pp. 123-127, 1983.

[62] A. Hall, "Lagrange Multiplier Tests for Normality against Seminonparametric Alternatives," *Journal of Business & Economic Statistics,* vol. 8, no. 4, pp. 417-426, 1990.

[63] R. H. Koning and B. van der Klaauw, "Testing the normality assumption in the sample selection model with an application to travel demand," *Journal of Business & Economic Statistics,* vol. 21, no. 1, pp. 31-42, 2003.

[64] T. W. Epps, K. J. Singleton, and L. B. Pulley, "A test of separate families of distributions based on the empirical moment generating function," *Biometrika,* vol. 69, no. 2, pp. 391-399, 1982.

[65] J. Sigut, J. Pineiro, L. Moreno, J. Estevez, R. Aguilar, and R. Marichal, "A large deviation approach to normality testing," *Computational statistics & data analysis,* vol. 49, no. 3, pp. 741-756, 2005 2004.

[66] G. Wainrib, M. Bachar, J. Batzel, and S. Ditlevsen, Eds. *Stochastic Biomathematical Models: With Applications to Neuronal Modeling* (Lecture Notes in Mathematics). Springer, 2013.

[67] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrom Rev,* vol. 26, no. 1, pp. 51-78, 2007 Jan-Feb 2007.

[68] J. W. Locasale *et al.*, "Metabolomics of human cerebrospinal fluid identifies signatures of malignant glioma," *Mol Cell Proteomics,* vol. 11, no. 6, p. M111.014688, Jun 2012.

[69] B. Durbin and D. M. Rocke, "Estimation of transformation parameters for microarray data," *Bioinformatics,* vol. 19, no. 11, pp. 1360-7, Jul 2003.

[70] D. M. Rocke and S. Lorenzato, "A Two-Component Model for Measurement Error in Analytical Chemistry," *Technometrics,* vol. 37, no. 2, pp. 176-184, 1995.

[71] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: improving the biological information content of metabolomics data," *BMC Genomics,* vol. 7, p. 142, 2006.

[72] R. R. Pagano, *Understanding statistics in the behavioral sciences*, 5th ed. Pacific Grove: Brooks/Cole Pub. Co., 1998, pp. xxviii, 548 p.

[73] S. S. Shapiro and M. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika,* vol. 52, no. 3-4, pp. 591-611, 1965.

[74] J. W. Tukey, "The comparative anatomy of transformations," *The Annals of Mathematical Statistics,* vol. 28, no. 602-632, 1957.

[75] S. H. Elsea *et al.*, "Elucidation of the complex metabolic profile of cerebrospinal fluid using an untargeted biochemical profiling assay," *Mol Genet Metab,* Apr 2017.

[76] A. Evans *et al.*, "High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics," *Metabolomics,* vol. 4, 2014.

[77] A. M. Evans, C. D. DeHaven, T. Barrett, M. Mitchell, and E. Milgram, "Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems," *Anal Chem,* vol. 81, no. 16, pp. 6656-67, Aug 2009.

[78] L. W. Summer *et al.*, "Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI)," *Metabolomics,* vol. 3, no. 3, pp. 211-221, 2007.

[79] O. Hrydziuszko and M. R. Viant, "Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline," *Metabolomics,* vol. 8, no. 1, pp. 161-174, 2012.

[80] T. Hulsen, J. de Vlieg, and W. Alkema, "BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams," *BMC Genomics,* vol. 9, p. 488, Oct 2008.

[81] J. Calles-Escandon *et al.*, "Influence of exercise on urea, creatinine, and 3-methylhistidine excretion in normal human subjects," *Am J Physiol,* vol. 246, no. 4 Pt 1, pp. E334-8, Apr 1984.

[82] W. E. Mitch *et al.*, "Effects of Diet and Antihypertensive Therapy on Creatinine Clearance and Serum Creatinine Concentration in the Modification of Diet in Renal Disease Study," *American Society of Nephrology,* vol. 7, no. 4, pp. 556-565, 1996.

[83] R. T. Rubin, "Urine creatinine excretion: variability and volume dependency during sleep deprivation," *Psychosom Med,* vol. 33, no. 6, pp. 539-43, 1971 Nov-Dec 1971.

CHAPTER 3: MISSING VALUES

2.9. Introduction

Much of this dissertation is a result of collaboration between Metabolon and the Molecular and Human Genetics Dept. at Baylor College of Medicine. Initially this partnership explored the diagnostic value of high throughput metabolomics for the identification of IEMs [1]. This work demonstrated the power of global MS-based metabolomics as a screen for IEMs detailed in Miller, Kennedy, Eckhart, Burrage, Wulff *et al.* [2] and Kennedy, Miller, Wulff, *et al*. [3] and Elsea, Kennedy, Pappan, Donti, Evans, Wulff, *et al*. [4]. The technology has also demonstrated diagnostic potential in other disease settings. Donti *et al*. [5] details the ability of metabolomics to differentiate between adenylosuccinate lyase deficiency versus other neurometabolomic disorders sharing a similar genotypic profile. Similarly, metabolomics sphingomyelin levels were found to be depressed in subjects with peroxisomal biogenesis disorder (PBD) compared to disease mimics [6]. Recent research offers the potential for improved detection and/or staging of pancreatic [7], ovarian [8], colorectal [9] cancers through metabolomics. Evidence also supports the use of metabolomics in motor neuron diseases, including detection of Amyotrophic Lateral Sclerosis (ALS) [10], monitor progression of Alzheimer's [11], or informing treatment of Parkinson's [12].

The implications for these results and others is that metabolomics represents not just a tool for biomarker discovery, but offers the potential inform on disease progression and effectiveness of treatment at the individual patient level [13, 14]. Customization of treatment, more formally known as Personalized Medicine [15], Precision Medicine [16] or N-of-1 [17], has become

popular in recent years. For metabolomics, which has traditionally operated as a screening platform, the application to the clinical environment represents a significant shift for the technology.

2.10. The Z-Score

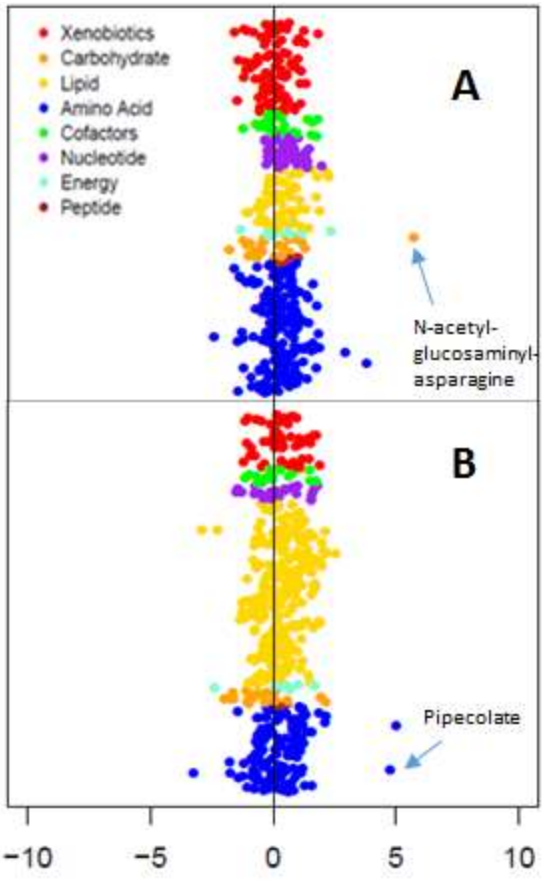
One approach to assessment at the individual level is to profile the individual's metabolic profile against the overall population at large. This is particularly effective in IEMs because the disease generally inhibits the body's ability to convert intermediates of important metabolic pathways, leading to either a buildup or depletion of metabolites in biochemical pathway affected. Identification of these cases amounts to outlier detection. The initial statistical analysis in Miller *et al*. [2] relied mainly on z-scores, which is equivalent to autoscaling (section 2.XX) and is used in metabolomic data [18, 19]. Autoscaling is attractive in metabolomics because all features on the same scale and hence given the same weight in analysis. The downside is that unit scaling may mask useful information about technical variation. For this reason, pareto scaling, which divides by $\sqrt{\sigma_i}$ rather than σ_i, is sometimes preferred as it keeps a better sense of the original scale while still making metabolite feature more homogenous [20]. In a direct comparison between auto and pareto scaling, Masson *et al*. identified autoscaling as more effective for GC-MS [21]. Conversely, Gromski *et al*. using NMR data found kNN classification to be highest with pareto scaling compared to various other scaling methods, including autoscaling [22]. The most appropriate transformation is likely to depend on the data and analysis in question.

The z-score translates the patient's biochemical level into the number of standard deviations from the average value of that compound. That is, for a given subject *j* and biochemical *i*, taking y_{ji} as the observed metabolite value for the subject. The z-score is then

$$z_{ji} = \frac{y_{ji} - \mu_i}{\sigma_i}$$

with μ_i and σ_i being the mean and standard among healthy subjects for compound i. Cases of IEMs present as outliers with very high or low values in the biochemical(s) related to their specific dysfunction. To illustrate this outlier behavior, the z-score plots of two subjects are shown in *Figure 3.1*. In these plots the individual metabolites are plotted with the x-axis being the z-scores and the y-axis being the individual metabolites. The metabolites are colored according to major biochemical pathway and fall into 1 of 8 classes. The first plot, subject A, is from a urine sample and is shown to have a high level of N-acetylglucosaminylasparagine, a compound signifying aspartylglucosaminuria [23]. The second plot, subject B, is from a plasma sample and shows two high outliers. One of these is pipecolate, elevated levels of which are related to Zellweger syndrome, Refsum Disease and others peroxisomal disorders [24, 25]. The other outlier, which is technically the most extreme hit in the plot for this subject, is S-methylcysteine. This biochemical is known to be associated with certain diets [26] and may be associated with cardiotoxicity [27], but no connection between this molecule and IEMs were found in a literature search.

Figure 3.1: Z Score plots of suspected IEM cases.



The presence of S-methylcysteine highlights an important reality of using z-score analysis. In extreme situations ion suppression may lead to erroneously low values whereas very high levels

in one sample may carry over to subsequent samples in run [28]. In cases of instrument error the biochemical may be regarded as a false positive, but note that the z-score for this subject may be accurate and not necessarily the result of instrument error. Genetic outliers are known to exist in perfectly normal, healthy subjects, with the average individual having anywhere from 60 to 200 genetic mutations [29]. The same phenomenon could easily occur in metabolomic components as well, leading to natural outliers that do not inform on the disease state. One may wish to refer to such outliers as non-informative markers instead of false positive. However, for diagnosis purposes the distinction appears to be moot. Regardless of whether the values of this biochemical is truly high in the subject or made artificially high by the instrument, either way these points are analogous to type I errors and create points of investigation that require follow up. Additional input from a clinician familiar with the patient's history is therefore always necessary to accurately diagnose a subject. Thus z-scores have a couple desirable traits. First, they are computationally easy. Second, they are easily interpretable by a physician. However, due to the natural outliers present in high through put metabolomics it is imperative to limit process induced outliers as much as possible.

2.11. Alternatives to Z-Scores

Other approaches besides z-scores are available both in clinical practice and statistical modeling strategies have been used extensively in the field of metabolomics. In fact, metabolomics is closely related to chemometrics [30-32]. Chemometrics involves a rich array of statistical methodology [33]. For example, partial least squares regression and principal component regression are have often been used in metabolomics [34]. It is, therefore, unsurprising to see multi-analyte or multivariate methods applied to metabolomics. Some examples of these are given next.

2.11.1. Linear Models

Impaired Glucose Tolerance (IGT) is a hyperglycemic state associated with higher blood glucose levels than normal but not high enough to qualify as type 2 diabetes [35, 36]. Generally, someone with IGT is considered to be pre-diabetic meaning that they have difficulty absorbing glucose but not to the same degree as someone with diabetes. Standard diagnosis for IGT involves the oral glucose tolerance test (OGTT) which involves fasting for at least 8 hours, taking a baseline blood draw, consuming a 75 gram oral glucose solution and measuring glucose levels after 2 hours. Blood glucose between 140 mg/dL and 199 mg/dL after 2 hours is indicative of IGT as are certain level of increase over baseline.

Demonstrating the ability of multiple metabolites to model impaired glucose tolerance (IGT) without the need for invasive procedures, Cobb, Eckhart, Perichon, Wulff *et al.* [37] used a logistic regression model to predict the probability of a subject being IGT. Letting p stand for the probability of a subject being IGT, the logit of IGT is modeled as linear combination four metabolites:

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 * \alpha\text{–}hydroxybutrate + b_2 * linoleoyl\text{–}glycerolphosphocholine +$$

$$b_3 * 4\text{–}methy\text{–}2\text{–}oxopentanoate + b_4 * Oleate$$

The model was developed using a total 1,623 available plasma samples from two separate cohorts: Relationship between Insulin Sensitivity and Cardiovascular Disease Study (RISC; n=955)[38] and The Diabetes Mellitus and Vascular health initiative (DMVhi; n=668) [39]. Model development was done in RISC (AUC=.82) and then validated in DMVhi (AUC=.83). This model was found to perform significantly better ($p < .05$) than the standard OGTT in both the training and test cohort.

In the realm of continuous outcomes, glomerular filtration rate (GFR) measures renal function by estimating the flow of fluid through the kidney. Calculating the true or measured GFR (mGFR) requires an intensive regiment consisting of multiple blood draws and urine collections plus a continuous intravenous infusion. In practice GFR is typically estimated (eGFR), most commonly with creatinine levels in serum (sCR). One such formulation, the Modification of Diet and Renal Disease (MDRD) study equation, is given by

$$eGFR_{MDRD}=175*sCR^{-1.154}*Age^{-0.203}*0.742_{\text{if female}}*1.212_{\text{if black}}$$

which is a linear regression equation (or four separate regression equations) depending on race and sex, on the log scale:

$$\log(eGFR_{MDRD})=\log(175)\ -\ 1.154*\log(sCR)\ -\ 0.203*\log(Age)\ +$$

$$\log(0.742)*I(\text{female})+\log(1.212)*I(\text{black})$$

The Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation, which is the current recommended standard of care for estimating GFR, involves a more complicated linear model:

$$\log(eGFR_{CKD\text{-}EPI})=\log(144)\ +a*\min\left(\frac{sCR}{k},1\right)\ -\ 1.209*\max\left(\frac{sCR}{k},1\right)\ +$$

$$\log(0.993)*Age+\log(1.018)*I(\text{female})+1.156*I(\text{black})$$

where

$$a=\begin{cases}-0.329 \text{ if female}\\ -0.411 \text{ if male}\end{cases}$$

$$k=\begin{cases}0.7 \text{ if female}\\ 0.9 \text{ if male}\end{cases}$$

The sex and race differences in the models are due to creatinine which is known to be higher in men than women and higher in African Americans than non-African Americans. Recently, new algorithms for estimating GFR have been proposed based on metabolites beside creatinine in

addition to offering the potential to remove sex and race which are becoming increasingly undesirable [40].

2.11.2. Metabolite Ratios

There are biological reasons to support the use of multiple analytes. First, individual metabolites could enhance one another. For example, both urea and creatinine are known to be related to kidney function with poor function leading to elevations of each. Therefore, high levels in both urea and creatinine help strengthen a conclusion of poor kidney function. When assessing quality of plasma samples, increases in lysophosphatidylcholines with decreases in phosphatidylcholines signal prolonged storage at room temperature [41]. Because these two classes move in different directions, taking a ratio between members of each classes provides better diagnostic ability for poor storage conditions. Similarly, creatinine and creatine are both known to be different on average between adult men and adult women with men tending to have higher levels of creatinine while women tend to have higher levels of creatine. Either of these markers are individually good discriminator of gender; however, in certain conditions the ratio of the two can provide an even stronger discriminator. The second advantage to using multiple metabolites is robustness. The number of metabolic processes and interactions in the human body is vast [42], allowing any given metabolite to be affected by multiple conditions. Focusing on either $eGFR_{MDRD}$ or $eGFR_{CKD\text{-}EPI}$, these models associate low levels of kidney function with high levels of creatinine. Yet creatinine can clearly be affected by other conditions. In fact, high levels of creatinine are also associated with diabetes and heart disease [43]. But since creatinine is also associated with exercise, kidney function and other conditions, reliance on this metabolite can lead to increases in the false positive and false negative rates when a person's behavior is outside "normal" behavior.

Association with multiple conditions is not unique to creatinine. Tryptophan is known to be positively associated with GFR [37], but as it is not synthesized by the body its levels are strongly related to diet [44]. For example, a person with even the poorest of kidney function could in theory appear healthy if they ate enough poultry or flax seeds [45].

Vitamins are obviously important to healthy human function but are frequently taken as supplements for various reasons: ascorbic acid is believed to help improve immune function, folic acid is recommended during pregnancy to boost fetal brain development, and so on. Vitamin pills and other supplements often contain ingredient levels that are well above the recommended daily intake level. The popular energy drink Red Bull for example contains 2.5 times the daily amount of pyridoxal (vitamin B6). The point here is that any single marker could easily be affected by a person's behavior. Having multiple metabolites in the model helps to mitigate these problems.

2.11.3. Decision Trees

Another way to address the complicated nature of metabolism is with decision trees. Returning to the IEM scenario, pipecolate is an example of this. Peroxisomal Biogenesis Disorders are related to the inability to breakdown certain amino acids, particularly very long chain fatty acids. Presence of these diseases are characterized by elevated levels of cerotic acid (C26:0) and hexocosenoic acid (C26:1) as well as the ratios of long chain fatty acids [46]. Within this family of disorders, pipecolate is absent from single gene mutation variations but present in the Zellweger syndrome spectrum. As a simple example this could be described with the decision tree shown in *Figure 3.2*. The advantage to decision trees is that pipecolate does not factor in until after sufficiently high levels of cerotic and hexocosenoic acid have been found to warrant a diagnosis of PBD. Thus, an extremely high level of pipecolate, which can also occur in

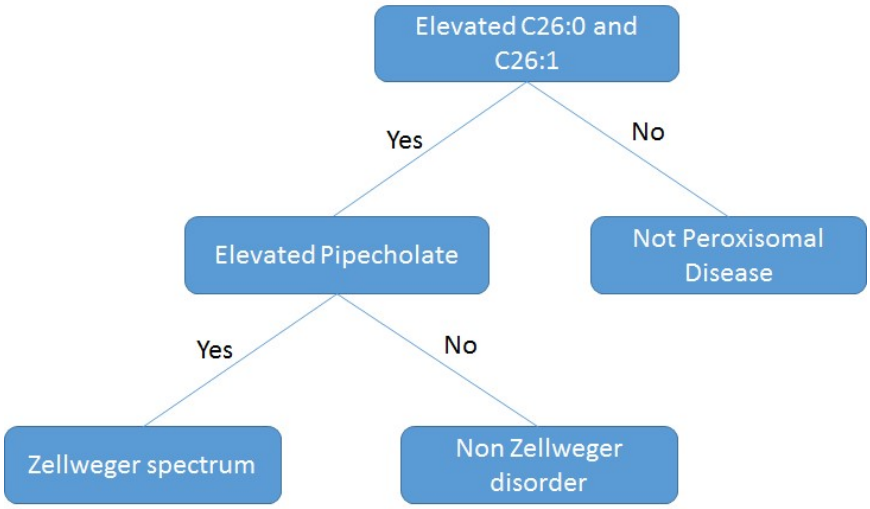


Figure 3.2: Hypothetical Decision Tree for Peroxisomal Disorders

pyridoxine-dependent epilepsy and other non PBD disorders [47, 48] would not by itself trigger a false identification of PBD. This decision tree-based algorithm is possible because of the unique relationship between pipecolate and disorders in the Zellweger spectrum. Such relationships are not always present and even when they are there is still the matter of determining what constitutes sufficiently elevated levels of the long chain fatty acids in the first split and the same for pipecolate in the second split, not to mention if there are other diseases related to the fatty acids in question. Ultimately, any algorithm used to classify a broad range of IEMs could incorporate a combination of decision trees and linear models. But developing such a model would require a greater understanding of the relationships of the hundreds of metabolites measured by global MS, many of which have received little to no attention.

2.11.4. Challenges to Modeling Rare Diseases

Developing extensive knowledge of global metabolic interactions, particularly the disease-metabolite relationships, may require large cohort studies. Particularly in the case of IEMs and other rare diseases. Consider collectively IEMs occur in roughly 1 out of 1,000 births [49-52], though incidence may vary considerably by region. Individually, incidence of specific disorders

within this family will be much less frequent. Alkaptonuria, which was discussed in Chapter 2, occurs once in every 250,000 to 1,000,000 births [53]. MCADD, the other IEM discussed in Chapter 2, occurs about once every 50,000 births [54]. This makes it challenging to acquire a reasonably sized cohort for specific disorder, let alone all disorders encompassing IEMs. On the other hand, a simple z-score calculated against a reference group of healthy individuals is sufficient to indicate outliers. Under normal theory, a z-score above 2 or below -2 indicates an extreme 2.5^{th} percent. In the previous chapter it was shown that a natural log transformation is reasonably effective at achieving a normal distribution for most metabolites. Combination of the log transformation followed by z-scoring leads to a straightforward way of identifying potential metabolomic outliers as the first step toward rare disease identification and can also be applied to the emerging discipline of *n* of 1 or personalized/precision medicine as well. Therefore, this dissertation is concerned with the general detection of biochemicals with extreme levels through the use of univariate z-scores.

2.12. Missing Values

As introduced in 1.7.2, The large amount of missing values in global MS metabolomics require some method of missing data analysis. As missing values play a central role it is appropriate to review the broader topic of missing data. Missing data are ubiquitous in scientific research and the topic has received much attention from both the statistical community and researchers within individual fields and there is a vast amount of literature covering the subject [55-57]. Conventionally, missing data are classified into three categories based on the mechanism behind the missing data. The strongest of these is Missing Completely at Random (MCAR) under which the probability of being missing is unrelated to the true value as well as any level of any covariate. A step down from this assumption is Missing at Random (MAR) in

which the missing value is related to a covariate, but otherwise unrelated to its true value. Lastly, if the missing value depends on some function of the value itself, with or without regard to the covariate levels, then the mechanism is termed Missing Not at Random (MNAR). In the context of metabolomics data, imagine $\boldsymbol{y_j}$ is the level of some metabolite and $\boldsymbol{A}$ is any set of covariates, including, potentially, other metabolites. Using the dummy variable $D_j = 1$ if the compound level is missing and $D_j = 0$ if the compound level is observed the three conditions can be written mathematically as

MCAR	$P\left(D_j = 1 \mid \boldsymbol{y_j}, \boldsymbol{A}\right) = P(D_j = 1)$
MAR	$P\left(D_j = 1 \mid \boldsymbol{y_j}, \boldsymbol{A}\right) = P(D_j = 1 \mid \boldsymbol{A})$
MNAR	$P\left(D_j = 1 \mid \boldsymbol{y_j}, \boldsymbol{A}\right) = P(D_j = 1 \mid g(\boldsymbol{y_j}, \boldsymbol{A}))$

To illustrate examples of each of these consider the following suppositions. First, suppose a short in the machine causes the portion of the chromatogram in which the metabolite elutes to be lost in the last third of the run. This situation would be MCAR as missingness, or the manner in which data are missing from the data, depends on neither the true value nor any other variable pertinent to the experiment. Next, suppose that the biochemical in question elutes very close to another biochemical with massive abundance so large that the peak could cover the area of both biochemicals. This scenario would be MAR as missingness is tied to the size of the neighboring compound but otherwise unrelated to the value of the biochemical in question. Last, consider the LOD scenario when a value is missing because it falls below the background level of the instrument. It is clear that this situation is MNAR since missingness is directly tied to the sample level. In this case a missing value constitutes partial information, namely that the sample is below a certain threshold but otherwise unknown. In statistics such partially known

measurements are referred to as censored, with the observed values for a metabolite constituting a left-censored sample [58].

2.12.1. Regulatory Suggested Methods

Use of global metabolomics to identify subjects with a disease, like IEMs, is a diagnostic screening test, and the instrumentation involved is most likely to be classified as a Laboratory Developed Test (LDT) [59]. At present the US Food and Drug Administration (FDA) does not monitor LDTs in the same fashion as pharmaceutical drugs, though FDA approval is required for Medicare and Medicaid reimbursement of LDTs and FDA's current level of monitoring is under internal review. But for now, Federal oversight of LDTs is limited to Clinical Laboratory Improvement Amendments CLIA [60], which are a series of regulations to "establish quality standards for laboratory testing performed on specimens from humans, such as blood, body fluid and tissue, for the purpose of diagnosis, prevention, or treatment of disease, or assessment of health" [61]. The intent is to ensure a level of technical proficiency as well as a consistent process for generating and interpreting results. The process itself is unspecified as the law is intended to include all laboratories covering a broad industry of diagnostic and treatment activities. Missing data is mentioned only in the context of reporting results in the presence of missing patient records. Therefore, these do not provide suggestions for handling of missing (output) data. Outside of the Federal agencies, independent laboratory licensing is provided by the College of American Pathologists (CAP) [62]. Similar to CLIA, CAP accreditation focuses mainly on laboratory procedure, including things like consistency and accuracy of instrumentation as well as proper maintenance of patient records. It too does not offer any guidance on handling of missing values.

The FDA does provide guidance for missing values with regards to clinical trials. These methods include Complete Case, Single Imputation, Multiple Imputation, Full Information Maximum Likelihood and the EM algorithm. In the research role, metabolomics itself has come to rely mainly on imputation methods to handle missing values [63-65]. Various forms of single imputation have been used. Additional imputations include K Nearest Neighbors [22, 66], which is a popular machine learning algorithm and has been employed in various settings [67]. Imputations are also borrowed from other omic fields, including a family of imputation methods based on Principle Component Analysis. These methods for handling missing values are detailed next.

2.12.1.1. Complete Case

Complete Case Analysis involves removing observations in which missing values occur. This approach can be applied either to the samples, retaining only those subjects which experience no missing value in any of the compounds, or on the variables, removing any compound feature that is not completely observed. It is equivalent to completely filtering out features or samples with missing proportion greater than 0. After the data has been trimmed to the fully observed subset, routine analysis can be performed. Statistical analysis by complete cases is only unbiased if the missing data is MCAR, a condition that is rarely true in practice let alone for high throughput metabolomics. Although very simple to execute, when applied to metabolomics the loss of information can be staggering and often impractical. As observed by Steinfath *et al* [68]:

> *"The following simple calculation shows that such omissions would be extremely costly and may render the experiment worthless. Let us assume that only 1% of all measurements, i.e. of the components in the data matrix, are missing. Let us also assume that we have 100 genotypes and 100 metabolites. Then, according to the binomial distribution, on average, 63% of the rows or columns must be omitted, while in the case of pairwise correlation, generally only one or two values must be omitted. Therefore, in the case of multivariate methods, an estimation of missing values is reasonable."*

Here, Steinfath and his colleagues recognize the fundamental problem with removing all trace of missing values from the data, yet the individual metabolites may contain only a few missing observations. The assumption posited assumes that missing values are MAR, which is likely not the case. However, the total amount of missing data proposed by Steinfath is extremely conservative [22, 66]. In the experience of this author a general rule of thumb is that roughly half the metabolites in a large, global analysis will be fully observed while the other half will have some fraction of missing values. Overall percentage of missing values can vary depending on instrumentation and sample type, but it is generally true that complete case will result in significant, often unacceptable, reduction to the data.

2.12.1.2. Single Imputation

Single Imputation (SI) replaces missing observations with an identical value based on the observed data. As an example, a vector of 7 with four observed values is shown. In the case of mean imputation, the average of the four observed values is 1.5, which is subsequently filled in for the three unobserved values.

$$\begin{matrix} 1.2 \\ 2.2 \\ \\ 0.7 \\ \\ \\ 1.9 \end{matrix} \xrightarrow{\text{Mean Imputation}} \begin{matrix} 1.2 \\ 2.2 \\ 1.5 \\ 0.7 \\ 1.5 \\ 1.5 \\ 1.9 \end{matrix}$$

Choices of SI frequently used in metabolomics include mean or median [66], zero [69] and minimum / half-minimum [70]. Mean and median are consistent with missing at random while the other three are associated with missingness resulting from low values. As metabolomics is high dimensional it is worth noting that SI is generally applied to each individual compound or feature separately because the LOD could vary from compound to compound. One exception to this would be zero imputation. This is an extreme case of low valued assumption, implying that

compounds unobserved in sample are completely absent from that sample. This may be true in certain cases especially with drugs or other xenobiotics, but many endogenous metabolites are likely present to some degree.

In clinical trials, Last Observation Carried Forward is a form of single imputation for longitudinal data. Repeated sampling of the same biological unit is a common experimental design strategy and has been used in metabolomics [71]. However, longitudinal studies in metabolomics could be confounded by circadian rhythm. Metabolite levels fluctuate over a 24-hour period, responding to basic environmental changes such as sleeping, eating and physical activity. Understanding the relationship between metabolites and the circadian rhythm is an active area of metabolomics [72, 73]. Although repeated sampling is un-related to the IEM setting of this thesis there is diagnostic potential through monitoring of metabolite levels. This approach would be particularly applicable with patient monitoring for disease progression or treatment response. In such studies, specifying study conditions and sample collection to mitigate circadian variation is important for minimizing intra-subject variability. Therefore, it is important that the consistency of sample collection be extended, ideally, to commonalities such as the same time of day and/or the same fasted status for each sample draw.

Naturally, SI is very simple to implement. Unlike complete case though, SI does not produce universally unbiased parameter estimates even under MCAR. For proof of this, assume that values are indeed MCAR. Imputation with the observed mean will produce unbiased estimates for the mean. However, replacement with a single value will lead to an unnatural reduction in variation thus providing under estimates of the variance. Next assume that missingness follows LOD. Imputation with zero, half minimum or minimum. When estimating parameters for the z-score transformation, minimum imputation is almost guaranteed to be produce an inflated

estimate of the mean since the true will in most cases be lower than the imputed value. The only way this can be unbiased for the mean is if every single missing value happens to be right at the LOD. Through similar arguments, zero imputation is all but guaranteed to underestimate the mean. And of course, all will poorly estimate the standard deviation for the same reason mean imputation does.

2.12.1.3. Multiple Imputation

In contrast to SI, multiple imputation (MI) creates multiple sets of imputed data. Each realization contains plausible values of the missing data sampled from a Bayesian posterior distribution. The sets are analyzed as any complete dataset would be with final inference coming from a combination of the results across all the sets. MI is advantageous for a few reasons. Estimates are consistent and asymptotically normal under MAR or if the model of the missing data mechanism is correctly specified. Next, the variance of parameter estimates can be divided into portions due to between imputation and within imputation. Decomposition of the variance allows variability due to the imputation to be accounted for, unlike the single imputation case where variation is often artificially deflated, and that the impact of imputing can be assessed. Finally, MI is available in a few software packages, although most of these appear to assume MAR. One challenge to MI can be reproducibility due to the randomness of the draws. Also, there are a number of ways in which to implement MI, which may also contribute to confusion and inconsistency of results. Methods for simple missingness mechanisms include Linear Regression, Propensity Scores and Predictive Mean Matching. The first two of these are offered by PROC MI in SAS [74]. For more complex or data specific missingness structures there is the Markov Chain Monte Carlo (MCMC) approach; however, this assumes multivariate normality. Multiple (or Multivariate) Imputation by Chained Equations (MICE)[75], also known as Fully

Conditional Specification [76] or Sequential Regression [77] among others, is a popular extension of MI for multivariate data that has been used in a number of disciplines.

2.12.1.4. Maximum Likelihood

Maximum Likelihood is a statistical technique for parameter estimation involving the likelihood function:

$$L(\boldsymbol{\theta};\boldsymbol{X_n})=\prod_{j=1}^{n} \mathrm{f}\,(x_j;\,\boldsymbol{\theta})$$

The formula is equivalent to the probability density (pdf) of a random sample $\boldsymbol{X_n}$; however, the emphasis of the likelihood is to view $\boldsymbol{\theta}$ as a function of the observed data. Taking the derivative of L with respect to $\boldsymbol{\theta}$, setting the result equal to 0 and solving for $\boldsymbol{\theta}$ results in parameter estimates which maximize the likelihood function, giving the name maximum likelihood (ML). In practice a log transformation is applied along with a sign change to simplify the calculation by turning the product into a summation:

$$-\sum_{j=1}^{n} \frac{\partial \log \mathrm{f}\,(x_j;\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

This function is known as the score equation. The negative sign follows from $\log \mathrm{f}(x_i;\boldsymbol{\theta}) \leq 0$ as the pdf is always bounded between 0 and 1. Conceptually this approach may be thought of as producing the "best" estimates of θ given the observed data. In general ML tends to be efficient. However, the estimates may be biased and in many cases in which unbiased estimators can be claimed it is only done so asymptotically. Additionally, the solution to the score equation(s) may not have a closed form in certain settings, requiring numerical approximations to solve.

In relation to this dissertation, missing data due to limit of detection has two principle forms in ML. The first is censoring, in which an observation is recorded but there is no value to

associate with it due to the true value being below the detection threshold. The second is truncation in which values below the detection threshold are not just unknown but the observations themselves are unrecorded. For metabolomics, the number of missing values for a given compound is always known and when due to limit of detection it is appropriate to incorporate the partial information from these observations into the model. However, as discussed in Chapter 2, strictly speaking there are times when missing values will not be LOD. Alternative reasons for missing values are more in line with MCAR, where ignoring the missing values is acceptable. Truncation may offer a tradeoff over censoring, producing more accurate estimates when some values are actually MCAR/MAR but being less efficient when all values are due to LOD. As this paper focuses on the LOD assumption, the censoring approach is used exclusively.

2.12.1.5. Expectation-Maximization Algorithm

Broadly, the Expectation-Maximization (EM) Algorithm is a two-step, iterative approach useful for generating maximum likelihood estimates when the likelihood function is difficult to solve directly, such as with latent variables or missing data. The overall process is shown in *Figure 3.3*. The algorithm begins with the Expectation, or E, step in which creates a function of the expected log-likelihood based on current parameter estimates. This is followed by the Maximization, or M, step in which the derived function is maximized to obtain updated parameter estimates. With these updated estimates in hand the algorithm returns to the E step and continues cycling until some tolerance between the current and updated estimates is achieved.

Key here is the ability to handle missing values. In this case the EM steps essentially take current estimates of the parameters to generate estimates of the missing values based on the conditional expectations. After filling in the missing values, new parameter estimates are created.

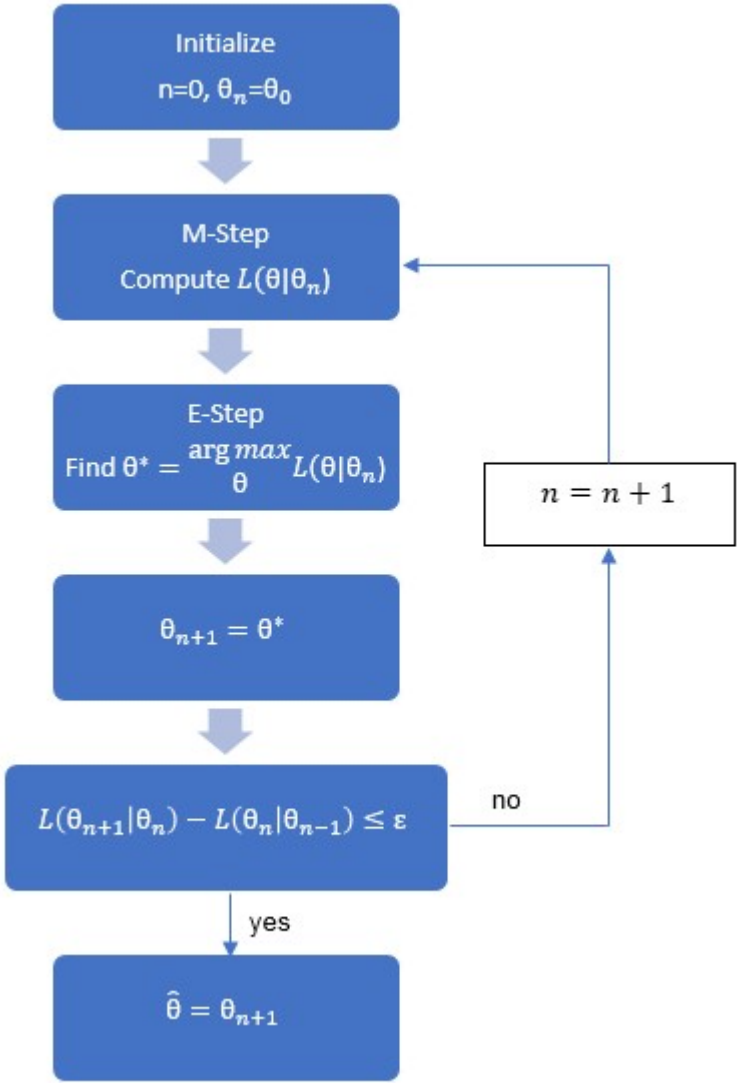


Figure 3.3: EM process flow diagram.

Primarily the purpose is to generate parameter estimates; however, a "complete" dataset can be found from the final E step. Overall, EM is regarded as being efficient and generally accurate regarding its primary goal – parameter estimation. As with ML, it does require specification of the likelihood function.

2.12.2. Common Metabolomic Imputations

In regard to missing data, metabolomic researchers are mostly interested in a complete dataset. Therefore, producing plausible estimates of the missing observations is most critical. The previous methods suggested by FDA are not necessarily meant for this purpose and many of

those methods require specifying a statistical model. Following Chapter 2 and 3, metabolite characteristics may not be known leaving researchers uneasy about specifying such behavior. Imputation in metabolomics has come to rely more on non-parametric methods. There does exist some overlap, such as SI via mean, minimum, zero, etc. But more complicated imputation strategies have largely been borrowed from other omic fields. This include variations on Principle Component Analysis, k Nearest Neighbors and Random Forest. The latter two have been found to perform well in comparison with other methods. Greater detail on given on these methods next.

2.12.2.1. Principal Component Analysis

Principal Component Analysis (PCA) is a popular multivariate technique with broad use in a number of disciplines including chemometrics and metabolomics [33, 78-81]. PCA is most simply a data reduction technique that reduces a set of correlated vectors into a smaller set of uncorrelated, or "principal", component vectors. Intuitively, even when there are a massive number of variables on n observations it should take no more than n - 1 independent pieces of information to completely separate the subjects in an experiment. Thus, PCA has useful application in high dimensional data in which $m \gg n$ and also in data that is highly correlated. More precisely, for $\boldsymbol{X}_{n\times m}$ a matrix of m metabolite features on n samples the singular value decomposition theorem states there exist matrices $\boldsymbol{U}_{n\times n}$, $\boldsymbol{V}_{m\times n}$ and $\boldsymbol{D}_{n\times n}$ (note that in practical applications $\boldsymbol{X}$ will always be of rank n) such that:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$$

$\boldsymbol{D}$ is a diagonal matrix with elements known as the singular values of X satisfying $\boldsymbol{D}_{11} \geq \boldsymbol{D}_{22} \geq \ldots \geq \boldsymbol{D}_{nn}$. $\boldsymbol{U}$ and $\boldsymbol{V}$ are the solutions associated with the eigenvalues of $\boldsymbol{X}\boldsymbol{X}'$ and $\boldsymbol{X}'\boldsymbol{X}$ respectively:

$$\boldsymbol{X}\boldsymbol{X}'\boldsymbol{U} = \boldsymbol{U}\boldsymbol{D}^2$$

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{D}^2$$

The full PCA decomposition of $\boldsymbol{X}$ is given by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{V}$. The columns of $\boldsymbol{Y}$ are orthogonal to each other and are referred to as the principle components of $\boldsymbol{X}$. A useful interpretation, relating the singular values, is that the i^{th} component of $\boldsymbol{Y}$ accounts for $\boldsymbol{D}_i \,/\, \sum \boldsymbol{D}_j$ of the observed variation in $\boldsymbol{X}$.

Pearson and Hotelling were the first to use PCA in the early 20th century [82, 83]. It was introduced to missing data by Christoffersson to handle matrices with missing values [84]. Near the turn of the century Tipping and Bishop added a probabilistic model to PCA (PPCA), which is closely related to factor analysis [85]. From this Oba developed an imputation method consisting of three parts: principal component regression, Bayesian estimation and an EM-like algorithm [86]. This approach is termed Bayesian PCA (BPCA). Finally, when suspicion that missing values are not believed to be a linear function of latent parameters Non-linear PCA (NLPCA) may be used [87]. For PPCA and NLPCA, missing values are essentially ignored during the fitting process but can be imputed on the back end.

2.12.2.2. k Nearest Neighbors

Broadly speaking, k-nearest neighbors (kNN) is a simple machine learning algorithm. In statistical application it is a tool for non-parametric prediction. It functions much the way other prediction tools do: Given a set of outcomes variables on *n* observations with m input variables from which to train, outcomes for a new observation can be made based upon the inputs. When the output variable is categorical it is often referred to as kNN classification, while continuous outputs are called kNN regression. It's worth pointing out that in metabolomic imputation, the neighbors are generally the metabolites with missing values for one metabolite of a sample filled

in based upon the observed values of the neighbors for the same sample. The alternative would be to fill in values of a metabolite in a missing sample based upon the observed values of the metabolite in another sample. Replacement by nearest sample is less likely to be successful since there is no guarantee that a group of randomly selected subjects would associate well. However, in theory certain metabolites can be very closely associated with one another based upon biochemical function or pathway relationship.

The specific algorithm for kNN involves comparing the input for a new observation (sometimes referred to as query) to the input of the training set by some dissimilarity measure. This is most commonly based on one of the following distance measures:

$$\text{Euclidean} = \sqrt{\sum_{j=1}^{n} \left(y_{ji} - y_{ji'}\right)^2}$$

$$\text{Manhattan} = \sum_{j=1}^{n} \left| y_{ji} - y_{ji'} \right|$$

$$\text{Minkowski} = \sqrt[q]{\sum_{i=1}^{n} \left| y_{ji} - y_{ji'} \right|^q}$$

$$\text{Hamming} = \sum D_i \text{ where } \begin{Bmatrix} y_{ji} = y_{ji'} : D_i = 0 \\ y_{ji} \neq y_{ji'} : D_i = 1 \end{Bmatrix}$$

Euclidean represent the shortest distance between two points in an n-Dimensional space and is the most popular distance measure for continuous values. Manhattan measures the distance between two points based on a grid structure. Minkowski is the generalization of the first two, giving Manhattan when $q = 1$ and Euclidean when $q = 2$. While these three are all appropriate for numeric values, Hamming is able to handle categorical variables by counting the number of instances in which $\boldsymbol{y_i}$ and $\boldsymbol{y_{i'}}$ disagree. Although not highly applicable for metabolomics,

metabolites could conceivably be dichotomized based on the ion counts being below versus above a certain threshold, or by an indicator variable based on presence or absence of an ion count. Regardless, once the dissimilarity is calculated, the k training observations with the smallest dissimilarity are used to formulate a prediction of the outcome variable based on some function of the k training outcomes. When the outcome is categorical this can be the mode of the k "neighbors" while continuous variables typically use the mean or median.

The nearest neighbor concept has been extended to missing data as a method to produce realistic replacements for missing values and has become popular in the omic fields [88]. As an imputation method it is non-parametric and software packages for kNN imputation are widely available; however, it is not guaranteed to produce unbiased estimates under MAR or even MCAR.

To illustrate how kNN works a classification example is given. Taurocholate (TC), glycocholate (GCC), taurochenodoexycholate (TCDC) and glycochenodeoxycholate (GCCDC) are the major bile acids in humans and are markers of liver health. Non-alcoholic fatty liver disease (NAFLD) is a form of liver disease in which excessive amounts of fat accumulate in the liver regardless of alcohol consumption resulting in impaired liver function. Non-alcoholic steatosis (NASH) is more advanced condition which can lead to liver failure. The following observations relate the biochemical levels of 7 hypothetical patients to fictional disease outcomes:

Patient	TC	GCC	TCDC	GCCDC	Condition
1	4.2	10.4	9.7	19.1	Normal
2	3.8	9.1	10.6	12.6	Normal
3	5.8	9.7	11.4	17.3	NAFLD
4	4.1	8.1	8.9	11.8	NAFLD
5	3.1	6.5	12.8	3.2	NASH
6	3.5	7.9	15.1	9.8	NASH
7	4.3	9.2	12.8	7.3	?

The first six with known disease conditions represent training samples while the seventh sample is unknown and the desire is to classify it based on the data at hand. Using Euclidean as the dissimilarity measure and setting k = 3 gives:

Patient	TC	GCC	TCDC	GCCDC	Euclidean Distance	Rank
1	0.01	1.44	9.61	139.24	12.26	6
2	0.25	0.01	4.84	28.09	5.76	3
3	2.25	0.25	1.96	100.00	10.22	5
4	0.04	1.21	15.21	20.25	6.06	4
5	1.44	7.29	0.00	16.81	5.05	2
6	0.64	1.69	5.29	6.25	3.72	1

The three smallest distances are, in ascending order, patients 6, 5 and 2. The corresponding conditions are NASH, NASH and Normal. With NASH being the condition of two of the three closest neighbors, patient 7 is predicted as being NASH. If Condition had been continuous, such as a severity measure, then the prediction would be either the mean or median (or some other function) of the three neighbors.

The main attractions to kNN are rapid processing speed, ability to handle categorical and continuous variables in both the inputs and outputs, and distributional free assumptions. The number of user defined settings are few. Dissimilarity is generally Euclidean for continuous variables and Hammering for Categorical. Resolving the prediction from the neighbors is typically done by the mean, though median may be more appropriate when severe outliers are present. The biggest hurdle is determining the value of k and the optimal value will depend upon the data. Selecting k = 1 is analogous using Nearest Neighbor or 1NN. A rough rule of thumb is k = 10 and this is the default in a number of software packages; however, in metabolomics performance has been shown to level off around k = 5.

Major drawbacks to kNN are that it can be harmed by the presence of non-informative input variables, imbalance in the output variable and ties in the prediction for classification, and

inconsistent scales among continuous inputs. For the first point, consider that neighbors are evaluated by some dissimilarity, which, for the common measures listed above, will increase in magnitude as more input variables are included. When a large number of the input variables in the training set are unrelated to those of the new variable then the dissimilarity measure becomes dominated by noise meaning that training sample association with a new observation is largely determined by chance. In essence, non-informative inputs tend to weaken the power of the dissimilarity measure to discriminate between values. Focusing on kNN classification, by increasing the number of observation of a given class the more opportunity there is to increase the classes' reach due to natural variation within a population. Thus, when one class is severely over-represented in the training set compared to other classes the more opportunity that class has to be close to any new observation. This can be overcome by condensing the training space, such as with subsampling, to produce a training set with equal representation of the class variable or alternatively by weighting the prediction so that closer neighbors have more impact on the prediction. A similar solution can be employed for instances of when two or more classes have equal representation in the k nearest neighbors. In the binary case ties can be avoided by setting k to an odd value; however, in the general multiclass setting there is no way to guarantee that a tie will be avoided. Ties must be broken somehow. Taking the value of the closest neighbor (i.e. 1NN) or defaulting to the most prevalent class are analogous to weighting. However, random selection of the tied categories is also possible.

The most relevant drawback to metabolomic data is the impact of input variables with vastly different scales. This can be seen somewhat in the example provided above. Glycochenodeoxcholate has a larger abundance than the others causing this metabolite to dominate the dissimilarity (under distance-based metrics at least). Scaling of variables is

therefore important with kNN. Section 2.1 discussed various scaling techniques for metbolomic data. Repeating the above example by dividing the levels of each metabolite by the median of the training set gives:

Patient	TC	GCC	TCDC	GCCDC	Condition
1	1.063	1.209	0.882	1.566	Normal
2	0.962	1.058	0.964	1.033	Normal
3	1.468	1.128	1.036	1.418	NAFLD
4	1.038	0.942	0.809	0.967	NAFLD
5	0.785	0.756	1.164	0.262	NASH
6	0.886	0.919	1.373	0.803	NASH
7	1.089	1.070	1.164	0.598	?

and the associated Euclidean based ranking is:

Patient	TC	GCC	TCDC	GCCDC	Euclidean Distance	Rank
1	0.0006	0.0195	0.0794	0.9355	1.02	6
2	0.0160	0.0001	0.0400	0.1887	0.49	2
3	0.1442	0.0034	0.0162	0.6719	0.91	5
4	0.0026	0.0164	0.1257	0.1361	0.53	3
5	0.0923	0.0986	0.0000	0.1129	0.55	4
6	0.0410	0.0229	0.0437	0.0420	0.39	1

The three closest neighbors are now patients 1, 2 and 4, the conditions of which are NASH, Normal and NAFLD. This indecision is arguably more reflective of the data: the new observation is similar to the NASH patients in GCCDC but more similar to the Normal or NAFLD patients in TC and GCC. Regardless of one's opinion, median scaling has altered the neighbors (though the 1NN tie breaker would keep the prediction the same).

One final point for clarity, in the examples the patients form the neighbors and the missing variable is imputed from other values of this same variable. In the metabolomics setting the compounds become the neighbors and missing values in a compound are replaced based upon the values of other compounds. This further emphasizes the importance of scaling the variables.

2.12.2.3. Random Forest

Random Forest is another machine learning algorithm useful for classification and regression built upon decision trees. Unlike kNN, Random Forest is a supervised approach that uses the entire dataset. Multiple trees are created, each by randomly sampling from both the samples and candidate predictors. For a given tree, the samples selected for training are referred to as "in bag" and the unselected samples are "out of bag" (OOB). Within each random selection a tree is created based upon the observed outcome to classify the in-bag samples with the tree grown until the nodes are pure. An additional caveat to growing the tree is that at each node separation is based on the best split from a subsampling of the predictors rather than the best predictor. This approach to splitting protects against overfitting and is shown to boost performance [89]. The Final prediction for each sample is then aggregated over all the sets in which the sample was out of bag. Forests created for continuous outcomes are referred to as regression forests, and use regression trees rather than classification trees, with the final prediction being the mean or median predicted value. Categorical outcomes are termed classification forests with final predictions based on the mode.

The usual advantages and disadvantages from such ensemble approaches hold for Random Forest. Namely, the use of several (up to several thousand) trees significantly improves accuracy but sacrifices interpretability. Additional advantages are the ability to assess predictive performance by comparing out of bag predictions against the true value as well as the ability to assess the contribution of the individual predictors. The latter feature is helpful for biomarker identification as more influential variables are good candidates for further investigation. For an excellent review of Random Forest see Liaw and Wiener [90].

2.13. Methods

This paper seeks to evaluate the performance of common imputation methods versus more clinically accepted missing data methods in regard to estimating the mean and standard deviation from a single "control" population. The common imputation methods selected are (1) None – no imputation of the data, (2) Mean – single imputation with the observed mean/average on a feature to feature basis, (3) Min – single imputation with the observed minimum, (4) kNN – k Nearest Neighbors using metabolites as the neighbors and (5) RF – Random Forest regression. The two competing clinical methods are (6) rankit regression (RR) which involves regression the observed order statistics against their corresponding expected values under a standard normal distribution and (7) Maximum Likelihood (ML) for left censored data. (1)-(3) were chosen because they are common to metabolomics and are acceptable for use in clinical trial evaluation. (4) and (5) were selected over others because previous studies have found them to generally have better performance than other methods [22, 66, 91].

These alternatives were selected based on the results of chapter 4 which found that metabolites have strong normal tendency following a natural log transformation. Invoking this property allows for the use of maximum likelihood or other parametric models which can be combined with left censoring following the LOD assumption. Both methods do this. As EM is a ML based approach, inclusion of this method felt redundant. Finally, MI has many attractive properties, but it has already been explored in metabolomic data where its performance was found to be poor [66].

2.13.1. Rankit Regression

For X a normal variate with mean μ and variance σ^2 and $\boldsymbol{Z}$ the standard normal with mean 0 and variance 1, the distribution of $\boldsymbol{X}$ is related to that of $\boldsymbol{Z}$ by the following relationship:

$$\boldsymbol{X} \sim \mu + \sigma \boldsymbol{Z}$$

This can be recognized as a linear regression problem. Given a random sample $\boldsymbol{X}_n$ drawn from an unknown normal distribution, estimates of μ and σ^2 can be found by regression the ordered values of $\boldsymbol{X}_n$ against the corresponding expected values under a standard normal. Notationally, let $x_{(i)}$ be the i^{th} ordered value of $\boldsymbol{X}_n$ and $z_{(i)}$ be the i^{th} ordered value of sample of size n drawn from the standard normal. Additionally, let $z_i^* = E(z_{(i)})$. These are known as the rankits of the normal distribution. In general, the expected value of an ordered statistic is non-trivial. However, it is possible to show the expected value of $x_{(i)}'s$ are in fact a linear transformation of the $z_{(i)}'s$:

(3.3.1-1)
$$E\left[x_{(i)}\right] = \int_{-\infty}^{\infty} x_{(i)} \frac{n!}{(n-i)!i!} \mathrm{F}\left(x_{(i)}\right)^i \left[1-\mathrm{F}\left(x_{(i)}\right)\right]^{n-i} \mathrm{f}\left(x_{(i)}\right) dx_{(i)}$$

(3.3.1-2)
$$= \int_{-\infty}^{\infty} x_{(i)} \frac{n!}{(n-i)!i!} \phi\left(\frac{x_{(i)-\mu}}{\sigma}\right)^i \left[1-\phi\left(\frac{x_{(i)-\mu}}{\sigma}\right)\right]^{n-i} \varphi\left(\frac{x_{(i)-\mu}}{\sigma}\right) * \frac{1}{\sigma} dx_{(i)}$$

(3.3.1-3)
$$= \int_{-\infty}^{\infty} x_{(i)} \frac{n!}{(n-i)!i!} \phi\left(\frac{\mathrm{x_{(i)-\mu}}}{\sigma}\right)^i \left[1-\phi\left(\frac{\mathrm{x_{(i)-\mu}}}{\sigma}\right)\right]^{n-i} \varphi\left(\frac{x_{(i)-\mu}}{\sigma}\right) * \frac{1}{\sigma} dx_{(i)}$$

(3.3.1-4)
$$= \mu \int_{-\infty}^{\infty} \frac{n!}{(n-i)!i!} \phi(u)^i [1-\phi(u)]^{n-i} \varphi(u) du +$$

$$\sigma \int_{-\infty}^{\infty} u \frac{n!}{(n-i)!i!} \phi(u)^i [1-\phi(u)]^{n-i} \varphi(u) du$$

(3.3.1-5)
$$= \mu + \sigma z_i^*$$

Step 2 follows from a property of exponential families. Step 3 comes from the change of variable $u = (x-\mu)/\sigma$. The first resulting integral in step 4 is equivalent to the probability density function of the i^{th} ordered rankit of a standard normal, which over the entire real line has probability equal

to 1. The second integral is equivalent to the expected value the i^{th} ordered rankit from a standard normal. Returning to the so called rankit regression model of interest:

$$\begin{bmatrix} x_{(n)} \\ \vdots \\ x_{(1)} \end{bmatrix} = a + b \begin{bmatrix} E\,(z_{(n)}) \\ \vdots \\ E\,(z_{(1)}) \end{bmatrix}$$

The least squares estimate of a and b are:

$$a = \bar{x} \; - \; b\bar{z}^*$$

$$b = \frac{\sum_1^n (z_i^* \; - \; \bar{z}^*)(x_{(i)} \; - \; \bar{x})}{\sum_1^n (z_i^* \; - \; \bar{z}^*)^2}$$

Next, consider any subset of $U \subseteq \{1, \ldots., n\}$ and the expected value of the linear coefficients are:

(3.3.1-6) $$E(b) = E\left[\frac{\sum_U (z_i^* \; - \; \bar{z}^*)(x_{(i)} \; - \; \bar{x})}{\sum_U (z_i^* \; - \; \bar{z}^*)^2} \right]$$

(3.3.1-7) $$= \frac{\sum_U (z_i^* \; - \; \bar{z}^*) E(x_{(i)} \; - \; \bar{x})}{\sum_U (z_i^* \; - \; \bar{z}^*)^2}$$

(3.3.1-8) $$= \frac{\sum_U (z_i^* \; - \; \bar{z}^*)(\{\mu + \sigma z_i^*\} - \{\mu + \sigma\bar{z}\})}{\sum_U (z_i^* \; - \; \bar{z}^*)^2}$$

(3.3.1-9) $$= \frac{\sigma \sum_U (z_i^* \; - \; \bar{z}^*)(z_i^* \; - \; \bar{z}\})}{\sum_U (z_i^* \; - \; \bar{z}^*)^2}$$

(3.3.1-10) $$E(b) = \sigma$$

And

(3.3.1-11) $$E[a] = E\left[\bar{x}_U \; - \; b\bar{z}_U^*\right]$$

(3.3.1-12) $$= \left\{\mu + \sigma\bar{z}_U^*\right\} - \bar{z}_U^* E[b]$$

(3.3.1-13) $$= \mu + \sigma\bar{z}_U^* - \sigma\bar{z}_U^*$$

(3.3.1-14) $$E[a]=\mu$$

The substitution made in (3.3.1-8) and (3.3.1-12) follows from the result in (3.3.1-5). Hence the regression coefficients are unbiased estimators of μ and σ for any subset of the ordered values $\{x_{(1)}, \ldots, x_{(n)}\}$, including the full set $X_{(n)}$. In fact, when all the ordered values are used $\bar{z}^* = 0$ due to the symmetry of the normal distribution implying:

$$a = \bar{x} - b\bar{z^*} = \bar{x}$$

$$b = \frac{\sum_1^n (z_i^* - \bar{z}^*)(x_{(i)} - \bar{x})}{\sum_1^n (z_i^* - \bar{z}^*)^2} = \frac{\sum_1^n (z_i^*)(x_{(i)} - \bar{x})}{\sum_1^n (z_i^*)^2}$$

So, in the complete case the estimate of the mean is identical to the sample average while the estimate of the standard deviation is the ratio between the sum of the observed ranks, after centering, and the rankits over the sum of the rankits squared.

In terms of this thesis, rankit regression is relevant as it can easily be adapted to left-censored data by ignoring the unobserved values and regressing only with the observed values, treating them as the highest ordered values in the data. This approach is straightforward as many software applications have the capability to perform simple linear regression. The biggest challenge is finding values of z_i^* as the rankits are not trivial to calculate. However, numerical approximations are readily available [92].

2.13.2. Maximum Likelihood and Left-Censoring

The general likelihood function for a sample in which there are n total observations and k values are missing due to begin below a certain value T_i is:

$$L(\boldsymbol{\theta_i}\,;\mathbf{y_i}) = \frac{n!}{k_i!(n - k_i)!} P(x < T_i)^{k_i} * \prod_{k_i+1}^{n} \mathrm{P}(x < y_{ji} \mid y_{ji} > T_i)$$

The likelihood above is in two parts. The first part, $P(x < T_i)^{k_i}$, represents the probability associated with the unobserved values, while the second part, $P(x < x_{ji} \mid x_{ji} > T_i)$, relates to the probability of the observed values. Under a normal distribution with parameters μ and σ the likelihood is:

$$L\left(\mu_j, \sigma_i ; \boldsymbol{y_i}\right) = \frac{n!}{k!(n-k_i)!}\left[(\sigma_i\sqrt{2\pi})^{-1}\int_{-\infty}^{T_i} exp\left[-(x-\mu_i)^2 \Big/ 2\sigma_i{}^2\right]dx\right]^{k_i}$$

$$*\left(\sigma_i\sqrt{2\pi}\right)^{-(n-k_i)} exp\left[-\sum_{k_i+1}^{n}\left(y_{ji}-\mu_i\right)^2 \Big/ 2\sigma_i{}^2\right]$$

Here the j subscripting of the metabolomic framework is used to emphasize that there are several features with each having their own distribution ($\boldsymbol{\theta_i}$), point of truncation (T_i), and number of missing values k_i. However, the total number of samples, n, is the same. For simplicity the i subscripting is removed in the following sections, but the emphasis on individual metabolites remains. The log-likelihood gives:

$$l = \log\left(\frac{n!}{k!(n-k)}\right) + k*\log\left((\sigma\sqrt{2\pi})^{-1}\int_{-\infty}^{T}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)dx\right) - (n-k)\log\left(\sigma\sqrt{2\pi}\right) - \sum_{k+1}^{n}\frac{\left(y_j-\mu\right)^2}{2\sigma^2}$$

Using the transformation $\xi = (T-\mu)/\sigma$ the likelihood can be simplified to:

$$l = \log\left(\frac{n!}{k!(n-k)!}\right) + k*\log[\phi(\xi)] - (n-k)\log\left(\sigma\sqrt{2\pi}\right) - \sum_{k+1}^{n}\frac{\left(y_j-\mu\right)^2}{2\sigma^2}$$

The score functions follow as:

$$\frac{dl}{d\mu} = 0 + k*\frac{1}{\phi(\xi)}*\varphi(\xi)*\frac{d\xi}{d\mu} - 0 - \sum_{k+1}^{n}\frac{2*\left(y_j-\mu\right)}{2\sigma^2}*-1$$

$$= \frac{k*\varphi(\xi)}{\phi(\xi)}*\frac{-1}{\sigma} - 0 - \sum_{k+1}^{n}\frac{\left(y_j-\mu\right)}{\sigma^2}$$

$$= \frac{-k^{*}\varphi(\xi)}{\sigma\phi(\xi)} + \sum_{k+1}^{n} \frac{\left(y_j\text{-}\mu\right)}{\sigma^2}$$

$$\frac{dl}{d\sigma} = 0 + k^{*}\frac{1}{\phi(\xi)}{*}\varphi(\xi){*}\frac{d\xi}{d\sigma} - \frac{n\text{-}k}{\sigma} - \sum_{k+1}^{n} \frac{\left(y_j\text{-}\mu\right)^2}{2\sigma^3}{*}\text{-}2$$

$$= \frac{k^{*}\varphi(\xi)}{\phi(\xi)}{*}\frac{(T\text{-}\mu)}{\sigma^2} - \frac{n\text{-}k}{\sigma} + \sum_{k+1}^{n} \frac{\left(y_j\text{-}\mu\right)^2}{\sigma^3}$$

$$= \frac{k^{*}\xi^{*}\varphi(\xi)}{\sigma^{*}\phi(\xi)} - \frac{n\text{-}k}{\sigma} + \sum_{k+1}^{n} \frac{\left(y_j\text{-}\mu\right)^2}{\sigma^3}$$

Setting these equal to 0 and solving gives:

$$\sum_{k+1}^{n} \left(y_j - \mu\right) = k\sigma^{*}\frac{\varphi(\xi)}{\phi(\xi)}$$

$$\sum_{k+1}^{n} \left(y_j\right) - (n - k)\mu = k\sigma^{*}\frac{\varphi(\xi)}{\phi(\xi)}$$

$$\bar{x} - \mu = \frac{k}{n\text{-}k}\sigma^{*}\frac{\varphi(\xi)}{\phi(\xi)}$$

and

$$\sum_{k+1}^{n} \frac{\left(y_j\text{-}\mu\right)^2}{2\sigma^3} = \frac{n\text{-}k}{\sigma} - k^{*}\frac{\varphi(\xi)}{\phi(\xi)}{*}\frac{\xi}{\sigma}$$

$$\sum_{k+1}^{n} (y_i^2) - 2(n - k)\bar{y}\mu + (n - k)\mu^2 = \sigma^2(n - k) - k\sigma^2\xi^{*}\frac{\varphi(\xi)}{\phi(\xi)}$$

$$\sigma^2\left[(n - k) - k\xi^{*}\frac{\varphi(\xi)}{\phi(\xi)}\right] = \sum_{k+1}^{n} (y_j^2) - (n - k)\bar{y}^2 + (n - k)\bar{y}^2 - 2(n - k)\bar{y}\mu + (n - k)\mu^2$$

$$\sigma^2\left[1\text{-}\frac{k}{n - k}\xi^{*}\frac{\varphi(\xi)}{\phi(\xi)}\right] = \frac{\sum_{k+1}^{n}\left(y_i^2 - \bar{x}^2\right)}{n - k} + (\bar{y} - \mu)^2$$

Due to the symmetry of the standard normal distribution $\varphi(t) = \varphi(-t)$ and $\phi(t) = 1 - \phi(-t)$. This relationship is used to maintain some consistency with the estimating equations in the truncated case:

$$\bar{y} - \mu = \frac{k}{n - k}\sigma^* \frac{\varphi(-\xi)}{1 - \phi(-\xi)}$$

$$\sigma^2\left[1 - \frac{k}{n - k}\xi^* \frac{\varphi(-\xi)}{1 - \phi(-\xi)}\right] = \frac{\sum_{k+1}^{n}\left(y_i^2 - \bar{y}^2\right)}{n - k} + (\bar{y} - \mu)^2$$

2.13.3. Maximum Likelihood and Left-Truncation

In the normal truncated case the likelihood function is

$$L(\mu,\sigma) = \prod_{j=1}^{n}\left(1 - \phi\left(\frac{T - \mu}{\sigma}\right)\right)^{-1} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{\left(y_j - \mu\right)^2}{2\sigma^2}\right)$$

T, as before, represents the point of truncation giving the probability of a value being unrecorded as $\phi((T - \mu)/\sigma)$. The term in the first part of the likelihood relates to this probability of being observed and normalizes the area over the space $[T,\infty)$ to a total probability of 1. ML estimators for μ and σ have previously been shown [93-95]. Rewriting by expanding the product and using the transformation $\xi = (T - \mu)/\sigma$ produces:

$$L(\mu, \sigma) = (1 - \phi(\xi))^{-n}(\sigma\sqrt{2\pi})^{-n} exp\left(-\sum \frac{\left(y_j - \mu\right)^2}{2\sigma^2}\right)$$

Taking the natural log gives:

$$l(\xi, \sigma) = -n^* \log\left(1 - \phi(\xi)\right) - n^* \log\left(\sigma\sqrt{2\pi}\right) - \sum \frac{\left(y_j - \mu\right)^2}{2\sigma^2}$$

The resulting score equations are:

$$\frac{dl}{d\mu} = -n^* \frac{1}{1 - \phi(\xi)} * [-\varphi(\xi)] * \frac{d\xi}{d\mu} - n^* 0 - \sum \frac{2(y_j - \mu)}{2\sigma^2} * (-1)$$

$$= \frac{-n^* \varphi(\xi)}{1 - \phi(\xi)} * \frac{-1}{\sigma} + \sum \frac{y_j - \mu}{\sigma^2}$$

$$= \frac{-n^* \varphi(\xi)}{\sigma[1 - \phi(\xi)]} + \sum \frac{y_j - \mu}{\sigma^2}$$

$$\frac{dl}{d\sigma} = \frac{n^* \varphi(\xi)}{1 - \phi(\xi)} * \frac{d\xi}{d\sigma} - n^* \frac{1}{\sigma\sqrt{2\pi}} * \sqrt{2\pi} - \sum \frac{(y_j - \mu)^2}{2\sigma^3} * -2$$

$$= \frac{-n^* \varphi(\xi)}{1 - \phi(\xi)} * \frac{(T - \mu)}{\sigma^2} * (-1) - \frac{n}{\sigma} + \sum \frac{(y_j - \mu)^2}{\sigma^3}$$

$$= \frac{-n^* \xi^* \varphi(\xi)}{\sigma[1 - \phi(\xi)]} - \frac{n}{\sigma} + \sum \frac{(y_j - \mu)^2}{\sigma^3}$$

Solving for each parameter gives:

$$\frac{-n^* \varphi(\xi)}{\sigma(1 - \phi(\xi))} + \sum \frac{y_j - \mu}{\sigma^2} = 0$$

$$\sum \frac{y_j - \mu}{\sigma^2} = \frac{n^* \varphi(\xi)}{\sigma(1 - \phi(\xi))}$$

$$\sum \frac{y_i - \mu}{n} = \frac{\sigma^* \varphi(\xi)}{(1 - \phi(\xi))}$$

$$\bar{y} - \mu = \frac{\sigma^* \varphi(\xi)}{(1 - \phi(\xi))}$$

And:

$$\frac{-n^* \xi^* \varphi(\xi)}{\sigma(1 - \phi(\xi))} - \frac{n}{\sigma} + \sum \frac{(y_j - \mu)^2}{\sigma^3} = 0$$

$$n^* \left(1 - \frac{\xi^* \varphi(\xi)}{1 - \phi(\xi)}\right) = \sum \frac{(y_j - \mu)^2}{\sigma^2}$$

$$n\sigma^2 * \left(1 - \frac{\xi^* \varphi(\xi)}{1 - \phi(\xi)}\right) = \sum (y_j^2) - 2n\bar{y}\mu + n\mu^2$$

$$n\sigma^2*(1 - \frac{\xi*\varphi(\xi)}{1 - \phi(\xi)}) = \sum(y_j^2) - n\bar{y}^2 + n\bar{y}^2 - 2n\bar{y}\mu + n\mu^2$$

$$\sigma^2*\left(1 - \frac{\xi*\varphi(\xi)}{1 - \phi(\xi)}\right) = \frac{\sum(y_j - \bar{y})^2}{n} + (\bar{y}\text{-}\mu)^2$$

These equations produce a system of two equations with two unknowns (recall ξ is a function of μ and σ) which can be approximated using the Newton Raphson method [96].

2.13.4. Maximum Likelihood vs. Rankit Regression

As stated above, one challenge with rankit regression is estimation of the rankits themselves. Much work has been devoted to the approximation of these values for items that include the normal quantile plots and Shapiro-Wilk Test [97-99]. As the x-variable in the rankit regression, any bias in the estimation of the rankits can lead to spurious conclusions. To illustrate this, a simple simulation of a normal random variable X_n with mean $\mu = 10$, standard deviation $\sigma = 3$ and $n = 30$ is conducted. The variable is censored by removing low values three at a time, beginning with three smallest and progressively increasing up until only the three largest values remain. At each level of censoring both maximum likelihood and rankit regression are used to estimate the value.

To evaluate different approaches to estimating the rankits themselves, rankit regression was conducted in three different ways. The first uses the qqnorm function from R, which is used create normal quantile plots in that software. Estimates generated by qqnorm come from the function points, which, for $n > 10$, the default is to take the quantiles corresponding to $\phi^{-1}\left(\frac{j\text{-}0.5}{n}\right)$ for $j \in \{1, \ldots, n\}$. This decent approximation to the rankits is not exact, leading to systematic bias in the estimates. To address this bias, the second approach simulates a $N(0,1)$ variable, referred to as Z_n, with X_n regressed on the ranks of Z_n. On average, the ranks of Z_n will be equal to the desired ranks, leading to less bias. But since each individual realization is not the true rankits,

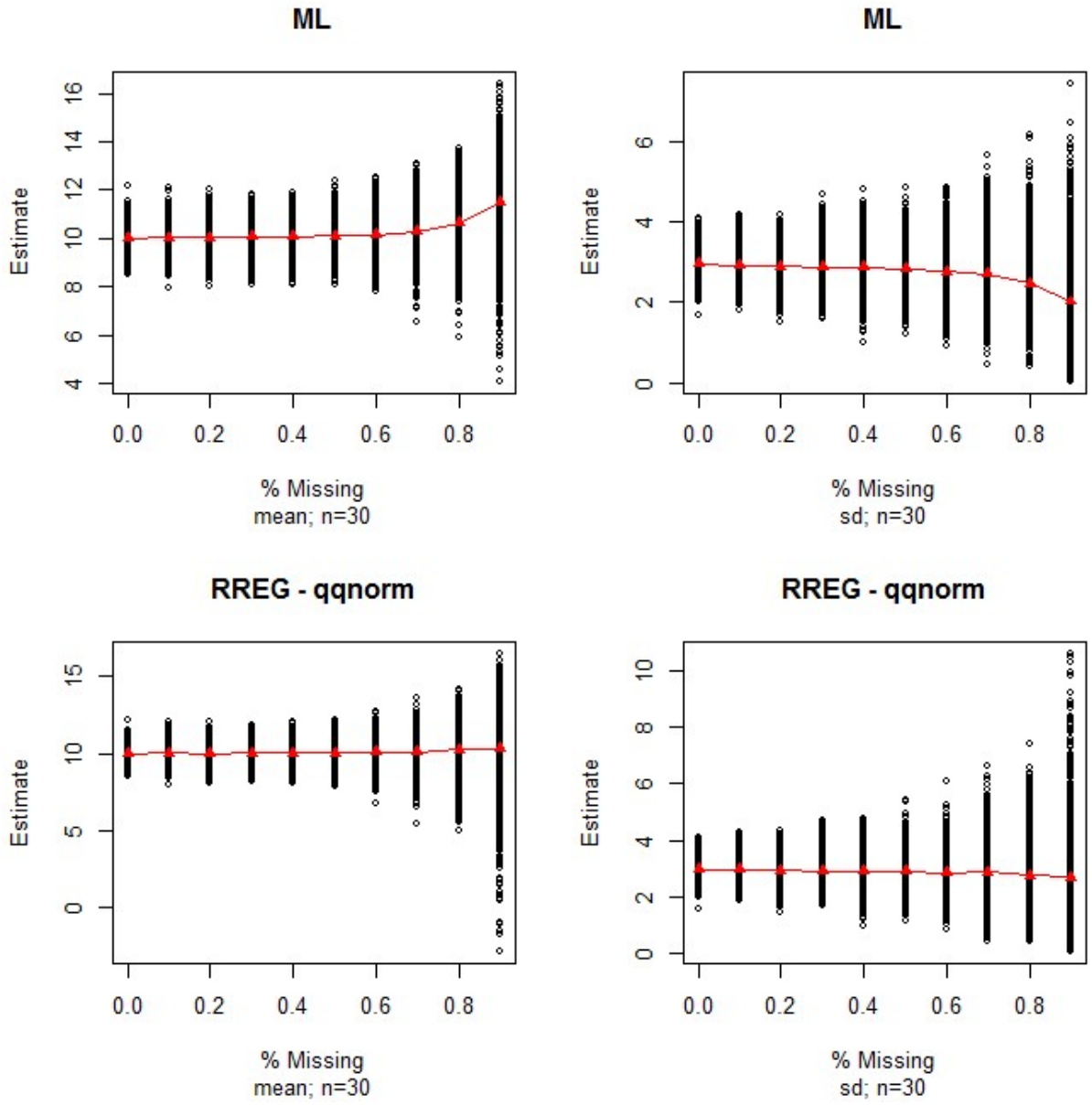


Figure 3.4: Parameter estimates from Maximum Likelihood versus Rankit Regression, part I. Individual estimates in black with average values in red. Rankits based on qqnorm function in R.

higher variance in the estimates are expected. In essence, this second approach does not remove error in rankit estimation but allows it to balance out over the simulations. Third, 100,000 realizations of Z_n are created, sorted and the ranks averaged to produce the independent variable Z_n^{avg}. This third way should have the least systematic bias due to error in the rankits while minimizing variance in the estimator.

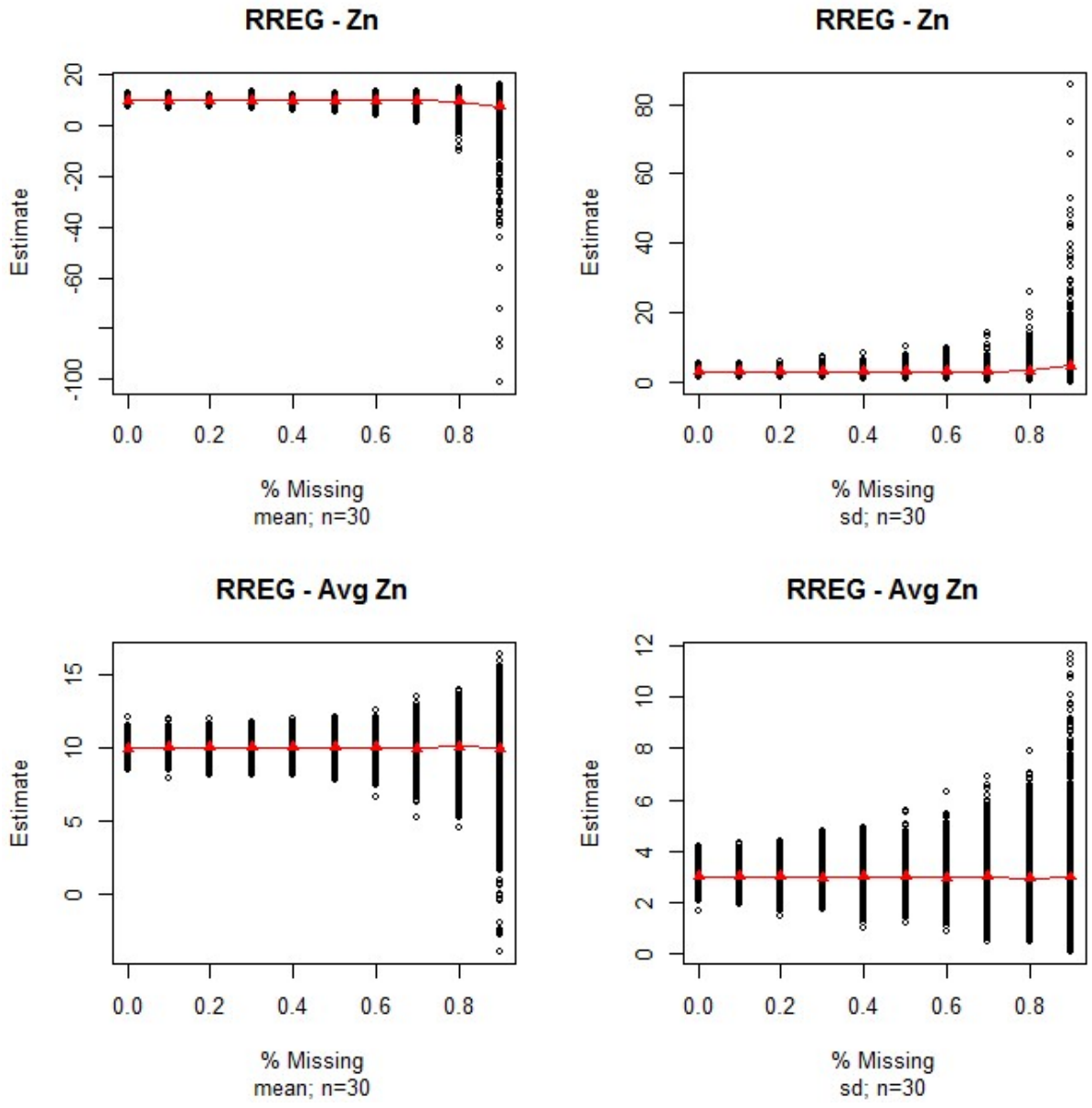


Figure 3.5: Parameter estimates from Maximum Likelihood versus Rankit Regression part II. Individual estimates in black with average values in red. Rankits based on simulated standard normal and average of 100,000 standard normal simulations.

Simulations results found that maximum likelihood does indeed exhibit bias in both the mean and standard deviation while rankit regression generally performs as expected. These results are summarized in *Figures 3.4-3.5*, which show the individual parameter estimates as a function of the proportion of missing values in the sample. Trend lines for the average parameter estimates

Table 3.1: Estimates of Population Parameters in Maximum Likelihood versus Rankit Regression. Results based on simulations of normal random variable with true mean 10 and true standard deviation 3.

		ML		RREG-qqnorm		RREG - Zn		RREG - Avg Zn	
Parameter	% Miss	Avg	Var	Avg	Var	Avg	Var	Avg	Var
MEAN	0%	9.97	0.302	9.97	0.302	9.98	0.642	9.97	0.302
	10%	10.03	0.289	10.01	0.288	10.04	0.593	10.02	0.288
	20%	10.03	0.308	10.01	0.317	10.05	0.625	10.01	0.317
	30%	10.05	0.354	10.03	0.370	10.09	0.738	10.02	0.370
	40%	10.07	0.378	10.03	0.426	10.06	0.883	10.01	0.427
	50%	10.11	0.432	10.04	0.515	10.06	1.125	10.01	0.519
	60%	10.15	0.567	10.07	0.695	10.03	1.685	10.01	0.706
	70%	10.27	0.870	10.09	1.247	9.99	2.856	9.99	1.292
	80%	10.64	1.558	10.25	2.457	9.77	8.076	10.09	2.624
	90%	11.48	3.845	10.33	8.045	7.66	102.080	9.98	9.223
SD	0%	2.95	0.137	2.96	0.141	2.94	0.289	3.02	0.147
	10%	2.92	0.162	2.96	0.180	2.97	0.392	3.01	0.186
	20%	2.89	0.192	2.94	0.224	2.97	0.489	2.99	0.232
	30%	2.86	0.261	2.91	0.301	2.94	0.636	2.98	0.313
	40%	2.86	0.304	2.93	0.368	2.96	0.817	3.00	0.385
	50%	2.82	0.344	2.91	0.448	2.99	1.078	3.00	0.472
	60%	2.75	0.457	2.86	0.583	3.01	1.575	2.96	0.622
	70%	2.69	0.636	2.87	0.923	3.07	2.364	3.00	1.002
	80%	2.46	0.850	2.76	1.339	3.22	5.185	2.94	1.499
	90%	2.01	1.454	2.70	3.064	4.62	51.104	2.98	3.692

are indicated in red. Maximum likelihood shows steadily increasing bias towards over-estimates in the mean and under-estimates in the standard deviation over the range of proportion missing. The bias appears to inflate at around 60% missing values.

Rankit regression using qqnorm is less biased, but an inspection of the average estimates, shown in Table 3.2, reveal a similar trend towards overestimates in the mean and under-estimates in the SD. Impact is most apparent in the SD parameter where rankit regression by qqnorm is seen to consistently underestimate the parameter through all proportions of missing values. While the bias in RREG is around .5-1 unit less than maximum likelihood, the variability of the estimates is much greater. This result is to be expected as maximum likelihood methods are generally very efficient. Bias in rankit regression using Z_n demonstrates more variability than qqnorm, as expected, at all levels of missing values. The variability of these estimates is tremendous in both parameters at 80-90% missing values. However, Z_n does tend to be slightly less biased than qqnorm, especially in the standard deviation, up to 70% missing values. Lastly,

Table 3.2: Difference in Rankit Estimates

Rank	qqnorm	Avg Zn	| Diff |
1	2.128	2.046	0.08230
2	1.645	1.619	0.02603
3	1.383	1.368	0.01549
4	1.192	1.181	0.01118
5	1.036	1.028	0.00861
6	0.903	0.896	0.00664
7	0.784	0.778	0.00571
8	0.674	0.670	0.00452
9	0.573	0.570	0.00310
10	0.477	0.475	0.00182
11	0.385	0.384	0.00164
12	0.297	0.296	0.00097
13	0.210	0.210	0.00053
14	0.126	0.125	0.00017
15	0.042	0.042	0.00044
16	-0.042	-0.041	0.00056
17	-0.126	-0.124	0.00118
18	-0.210	-0.209	0.00173
19	-0.297	-0.295	0.00208
20	-0.385	-0.382	0.00286
21	-0.477	-0.473	0.00365
22	-0.573	-0.568	0.00463
23	-0.674	-0.669	0.00553
24	-0.784	-0.777	0.00682
25	-0.903	-0.895	0.00816
26	-1.036	-1.026	0.01045
27	-1.192	-1.178	0.01365
28	-1.383	-1.364	0.01902
29	-1.645	-1.614	0.03046
30	-2.128	-2.041	0.08657

rankit regression by Z_n^{avg} demonstrates little evidence of bias at any parameter level while having variance that is roughly equivalent, if a little bit higher, than when using qqnorm. The specific rankit estimates for the qqnorm and Z_n^{avg} approaches are displayed in Table 3.2. Magnitudes between the two are largest in the tails, which fits with the bias under qqnorm being most noticeable at 80-90% missing. With this much of the data gone, estimation is entirely dependent on these most extreme values which happen to be the most off.

The evaluations that follow, the expectation is that maximum likelihood will generally perform well. However, being only asymptotically unbiased there is the possibility that in certain circumstances, such as very large proportions of missing data or small sample sizes, rankit regression will prove to be less biased. However, the findings here support maximum likelihood for left censored data as being clearly more efficient. The variability of rankit regression estimation is an item to pay attention to in further assessment.

2.13.5. Assessment

Selected methods will be evaluated in chemocentric metabolomic data using the same three data sets evaluated in Chapter 4. Using only those metabolites which were completely observed in all samples, the data is first log transformed, and then observed values are removed in left-censored fashion. Each method was applied to censored data to produce estimates of the mean and standard deviation, which were then compared against the corresponding sample values of the fully observed data. Error in the method can be described as the difference between the fully observed parameter value τ and the estimate of that parameter $\tilde{\tau}$. Comparison involves taking the so called relative error, which is error as a proportion of the fully observed value:

$$Relative\ Error = \frac{\tilde{\tau} - \tau}{\tau}$$

In order for relative error to be defined τ must be non-zero. This should not be a problem here since the standard deviation is always non-zero and ion counts are always positive values that are, even at the lowest level, in the thousands. The primary reasoning for relative error is that by viewing error as a percentage of the original parameter value it makes results comparable between features and between parameters. That is, we can not only equate relative error in the mean between two features, but also compare relative in the mean against that of the standard deviation for the same feature. Error is also closely related to bias, which is an important concept in statistics and one of the most, if not the most, important ways to evaluate estimators. By aggregating relative error across all features, some estimate of bias can be achieved.

Before experimenting with these data sets, simulation experiments were conducted to test the performance of the methods under ideal circumstances and their sensitivity to the normality and left censored assumption. These theoretical results will be a point of reference in which to consider the results in the real datasets and to understand the importance of the assumptions to performance. Simulations experiments are described fully in section 5.4.

Briefly, a series of relatively small datasets were created in which the columns are based on either a normal or a chosen representative non-normal distribution. Observations will then be removed from a portion of the dataset, with the methods being applied using the assessment approach outlined above. Use of kNN and RF necessitate simulating an entire dataset since both of these methods use the fully observed features to impute missing values. Naturally, correlation between the features is important to these methods and is included as part of the simulation experiments. Finally, since performance of kNN is known to suffer when the variables are on different scales, and metabolomic sets easily qualify this definition with average feature abundance ranging from the tens of thousands to the hundreds of millions, median scaling is also

including in the simulation workflow. Scaling is selectively used in the field and median scaling was chosen due to the propensity for large outliers in metabolomic data [18, 100]. Indeed, results using kNN were observed to be much poorer without any scaling. The other methods, including RF, were largely unaffected. Scaling is thus optional, for the other methods. Fortunately, median scaling has no effect on relative error for NONE, MEAN, MIN, ML and RR. It is important to realize that in practice such scaling can of course only be performed on the observed data. Thus, to best imitate the process in real data variable scaling is performed after variable censoring.

2.14. Simulation Experiments

2.14.1. Workflow

The proposed methods, ML and RR, assume normality and that missing values are due to the ion count falling below the instruments level of detection. Goals of the simulation experiments are as follows (i) evaluate performance of the methods when both assumptions are true, (*ii*) evaluate change in performance when one of the two assumptions fail and (*iii*) evaluate the impact of correlation on kNN and RF. Besides kNN and RF, the other five procedures are univariate, since they act on each feature separately, meaning it would be sufficient to asses these using simulations of a single variable. However, since kNN and RF operate from a dataset with fully observed variables it was necessary to simulate a full dataset. For this reason, these experiments involve 1,000 simulation runs of a multivariate dataset $\boldsymbol{Y}_{n\times m}=[\boldsymbol{y_1} \quad \boldsymbol{y_2} \quad \cdots \quad \boldsymbol{y_m}]$ where the $\boldsymbol{y_j}'s$ are identically distributed random variables. A value of $n = 30$ is selected as the standard since it is consistent with the sample sizes of the datasets discussed in Chapter 4. A value $m = 114$ is chosen with the variables $\boldsymbol{y_1, \ldots, y_{14}}$ censored such that the lowest $2*j$ values are removed for $j \in \{1, \ldots, 14\}$. Variables $\boldsymbol{y_{15}, \ldots, y_{114}}$ are uncensored and represent the fully observed data available to kNN and RF for imputing. The methods are then applied to produce

estimates $\{\tilde{\bar{y}}_1, \ldots, \tilde{\bar{y}}_{14}\}$ and $\{\tilde{s}_1, \ldots, \tilde{s}_{14}\}$ of, respectively, the censored sample means and standard deviations. These estimates are compared back to the usual sample mean and standard deviation without censoring the variable.

Beginning with the hypothesis that the assumptions of normality and LOD are correct, $\boldsymbol{Y}_{n\times m}$ is first simulated as multivariate normal. It is presumed that the correlation between metabolites is a highly influentially characteristic that will impact the performance of both kNN and RF. Metabolite correlation will vary depending on instrumentation. For example, ion-centric data is known to be highly correlated due to a single parent metabolite producing multiple related fragmentation features [101], but in Chapter 4 it was shown that the majority of feature pairs have at most only a moderate level of correlation. Incorporating this correlation can be accomplished by beginning with a multivariate normal $Z_{n\times m}$ having mean vector $\mathbf{0}$ and variance covariance matrix $\boldsymbol{I}_n$. Let $\boldsymbol{Y}_{n\times m}$ be as follows:

$$\boldsymbol{Y}_{n\times m} = \boldsymbol{Z}_{n\times m} * Chol(\boldsymbol{\Sigma}) + \boldsymbol{\mu}$$

where $Chol(.)$ is the Cholesky root. $\boldsymbol{Y}_{n\times m}$ will be normally distributed with mean $\boldsymbol{\mu}$ and variance-covariance $\boldsymbol{\Sigma}$. These simulations use the compound symmetry structure, namely:

$$\boldsymbol{\Sigma}_\rho = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho & \rho \\ \rho & 1 & & \rho & \rho \\ \vdots & & \ddots & & \vdots \\ \rho & \rho & & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix}$$

Let $\boldsymbol{Y}_N^\rho$ designate the multivariate dataset with mean vector $\boldsymbol{\mu}$ and variance-covariance Σ_ρ. Within each simulation a single random draw of $Z_{n\times m}$ is made from which the datasets $\boldsymbol{Y}_N^\rho$ for $\rho = \left\{0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{9}{10}\right\}$ are created. The complete workflow of this simulation is shown in *Figure 3.6*.

The two proposed methods are expected to perform well under normal and LOD simulations with performance degrading as the amount of missing values increases. One curiosity is whether

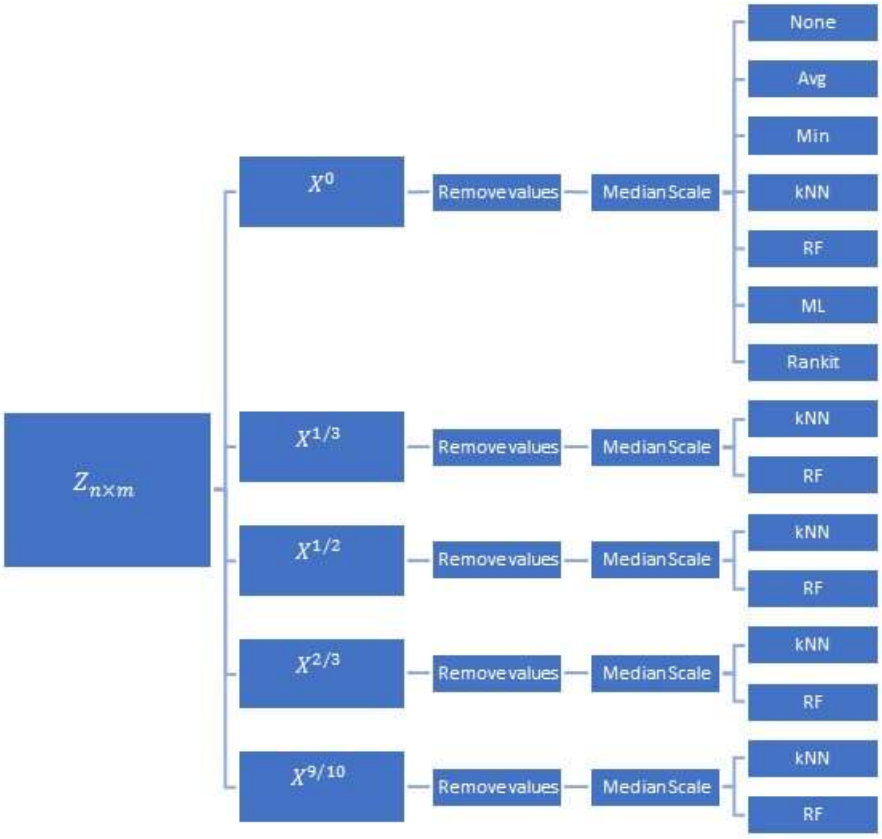


Figure 2.14.16: Feature set simulation workflow. Each experiment comprises 1,000 simulations. Z_{nxm} is a multivariate standard normal from which X^p is constructed with pairwise (spearman) correlation ρ for all vectors. Normal simulations involve X^p being multivariate normal while the non-normal simulations consist of log-normal and uniform. In all cases, values removed are done so by taking the lowest from the variables selected. For the normal simulation case, simulations are also performed by removing values at random.

performance is linked to the *number* of values that are missing or the *proportion* of values that are missing. To examine this the normal LOD simulations are repeated using $n = 10$ and $n = 600$ in addition to the standard $n = 30$. The smallest sample setting is particularly interesting with ML since this method is known to be only asymptotically unbiased.

The first assumption to be tested is the LOD assumption. Aside from LOD, potential sources of missingness are largely independent of the true value, i.e. lost scans or swamping from another peak, implying missing at random. Therefore, simulations with normal are conducted by removing values completely at random instead of in a left censored fashion. This will inform how critical the LOD assumption is and the possible impact of the methods on minor, but

plausible sources of missing values.

The next assumption to be tested is normality. There are numerous ways in which non-normality can manifest. The approach taken here is to create non-normality through skewness and kurtosis. The results of Chapter 4 showed that metabolites, particularly when untransformed, are prone to large outliers leading to a significant right skew. The log-normal distribution is useful at producing such right tailed data and has the added benefit of indicating what happens when a log transformation is not applied to a feature that does have strong log-normal behavior. Noting that metabolites can sometimes be rather evenly distributed within a tight range, we turn to the uniform distribution which is a non-skewed platykurtic distribution with excess kurtosis -6/5. In addition to both distributions having some basis for being used, data can also be simulated using the same workflow by transforming $\boldsymbol{Y}_N^{\rho}$. In the log normal case this is done by simply taking the $exp(\boldsymbol{Y}_N^{\rho}) = \boldsymbol{Y}_{LN}^{\rho}$. Similarly, any random variable can be made uniform using the cumulative distribution function from the target variable. Thus, for the uniform we have $\boldsymbol{Y}_U^{\rho} = \mathrm{F}_{\boldsymbol{Y}}(\boldsymbol{Y}_N^{\rho})$. Being able to create non-normal variables from $\boldsymbol{Y}_N^{\rho}$ allows for the correlation structure to be somewhat preserved. Since both transformations are monotonic and one-to-one, the spearman correlation will be unaffected, though the Pearson correlation will vary.

Table 3.3 summarizes all conditions of the simulation. In the real dataset the average ion count abundance level of metabolites was around 15 in the compounds that were fully observed after a log transformation (plasma = 15.09, CSF = 15.33, and urine = 16.75) while the standard deviation ranged from just less than 0.5 to over 0.9 (plasma = 0.50, CSF = 0.48, urine = 0.91). Based

Table 3.3: Conditions of simulation experiments.

Distribution	Missingness	
	LOD	MAR
Normal	n=10, 30, 600	n = 30
Log Normal	n = 30	
Uniform	n = 30	

on this, in the normal simulations $\mu_j = 15$ and $\sigma_j = 0.6$ for all $j \in \{1, \ldots, 114\}$. For the log-normal case, $\mu_j = 2.65$ and $\sigma_j = 0.3$ so that the resulting exponentiated variable has mean of 14.8, close to that in the normal case, along with skewness of 0.94. With the uniform μ_j is 0 and σ_j is 1, treating a standard normal $Y_j^{\rho}=\Phi^{-1}\left(Z_j^{\rho}\right)*2+14$. This produces a uniform distribution with mean 15 and standard deviation of 0.57.

2.14.2. Results

2.14.2.1. Normal Distribution and LOD

To begin, one simulation run from $\boldsymbol{Y}_N^0$ is presented. Tables 3.4 and 3.5 show the estimated mean and standard deviation for each method as well as the corresponding uncensored parameter value. A clear result from these tables is that median scaling considerably alters the original parameter value, shown in uncensored column. Each variable begins as $N(15,0.6^2)$ and the impact of median scaling when no values are missing move each variable towards a $N(1,0.6^2/15^2)$. Dividing the variable by an average of 15 centers the distribution around 1. However, as low values are removed from the data the observed median steadily increases

Table 3.4: Estimates of mean in one simulation. Distribution is normal with missing values censored below and ρ=0.

# Removed	ρ	Uncensored	NONE	AVG	MIN	kNN	RF	ML	RREG
2	0	0.986	0.992	0.992	0.987	0.993	0.992	0.986	0.986
4	0	0.987	0.996	0.996	0.990	0.999	0.997	0.987	0.986
6	0	0.991	1.005	1.005	0.996	1.004	1.005	0.992	0.993
8	0	0.995	1.009	1.009	1.001	1.006	1.009	0.995	0.992
10	0	0.981	1.002	1.002	0.985	1.000	1.002	0.975	0.977
12	0	0.975	0.999	0.999	0.985	1.000	0.999	0.974	0.973
14	0	0.970	1.010	1.010	0.986	1.007	1.010	0.963	0.963
16	0	0.970	1.006	1.006	0.993	1.001	1.006	0.980	0.982
18	0	0.980	1.011	1.011	1.001	1.004	1.011	0.985	0.975
20	0	0.965	1.005	1.005	0.990	1.000	1.006	0.968	0.965
22	0	0.963	0.998	0.998	0.984	1.002	0.998	0.961	0.968
24	0	0.953	1.008	1.008	0.988	0.999	1.010	0.945	0.931
26	0	0.941	1.002	1.002	0.993	1.000	1.003	0.969	0.958
28	0	0.945	1.000	1.000	0.996	1.001	0.999	0.983	0.970

Table 3.5: Estimates of SD in one simulation. Distribution is normal with missing values censored below and ρ=0.

# Removed	ρ	Uncensored	NONE	AVG	MIN	kNN	RF	ML	RREG
2	0	0.045	0.042	0.040	0.044	0.040	0.040	0.046	0.046
4	0	0.042	0.036	0.034	0.037	0.034	0.034	0.041	0.044
6	0	0.038	0.026	0.024	0.030	0.025	0.024	0.035	0.034
8	0	0.037	0.028	0.024	0.027	0.026	0.024	0.034	0.038
10	0	0.040	0.031	0.025	0.035	0.027	0.026	0.048	0.045
12	0	0.038	0.027	0.021	0.027	0.023	0.021	0.040	0.041
14	0	0.056	0.040	0.029	0.039	0.030	0.029	0.063	0.064
16	0	0.044	0.018	0.012	0.017	0.017	0.012	0.030	0.029
18	0	0.034	0.021	0.013	0.015	0.017	0.013	0.029	0.037
20	0	0.041	0.020	0.011	0.015	0.017	0.011	0.034	0.037
22	0	0.031	0.013	0.006	0.011	0.012	0.007	0.030	0.025
24	0	0.038	0.025	0.011	0.015	0.016	0.011	0.046	0.056
26	0	0.037	0.011	0.004	0.005	0.013	0.004	0.021	0.027
28	0	0.040	0.005	0.001	0.001	0.012	0.001	0.009	0.016

further and further above 15 as more and more of the lower values are removed. Thus, as the amount of missing values increases the target mean is being driven up. This effect of median scaling presents a challenge. For example, examining the eighth line of Table 3.4, in which 16 values have been censored (corresponding to $j = 8$) the five common metabolomics methods are all very close to 1. On other hand, ML estimates a mean of 0.980 while RREG gives an estimate of 0.982. These estimates are actually closer to the true uncensored level of 0.970 after scaling. Although the standard deviation does not display a systematic trend with regards to missing proportion, the uncensored values are prone to shift noticeably ranging from a high of 0.056 to a low of 0.031. From Table 3.5, when 14 values are missing NONE estimates the standard deviation as 0.04 while both ML and RREG estimate just over 0.06. NONE may seem better as the theoretical standard deviation (without censoring) is 0.6/15=0.04. However, the uncensored, scaled sample standard deviation is actually 0.056 for the variable in question. *Figures 3.7* and *3.8* plot the results of the estimated mean and standard deviation against their corresponding uncensored values for the seven methods. Each dot corresponds the estimated mean or standard

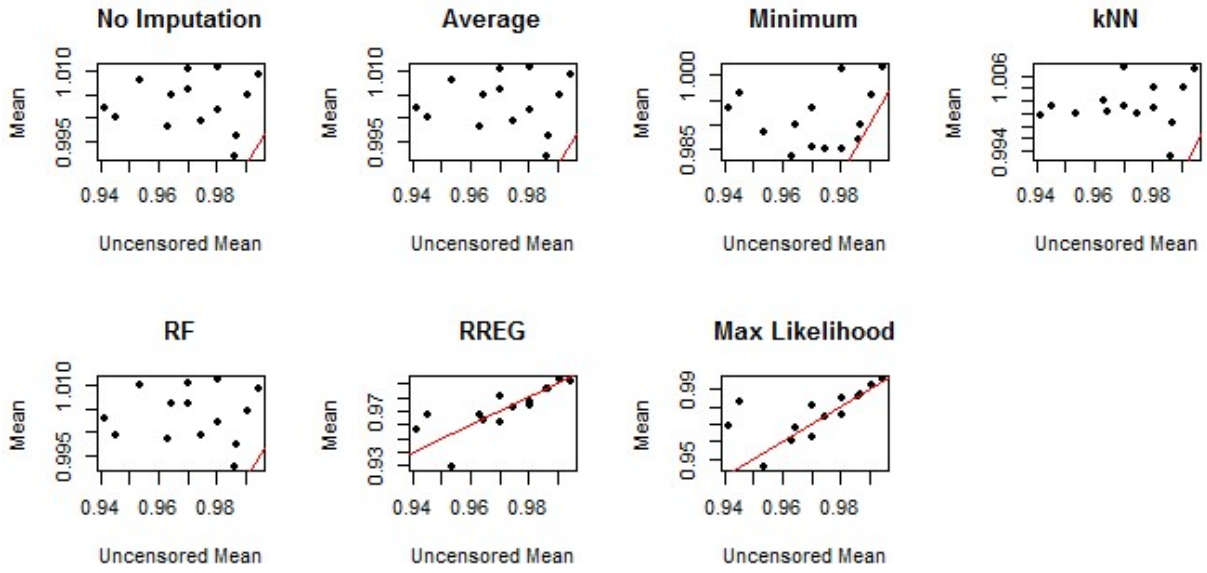


Figure 3.7: Estimated mean versus sample mean in one simulation of normal data where values are censored from below and ρ=0.

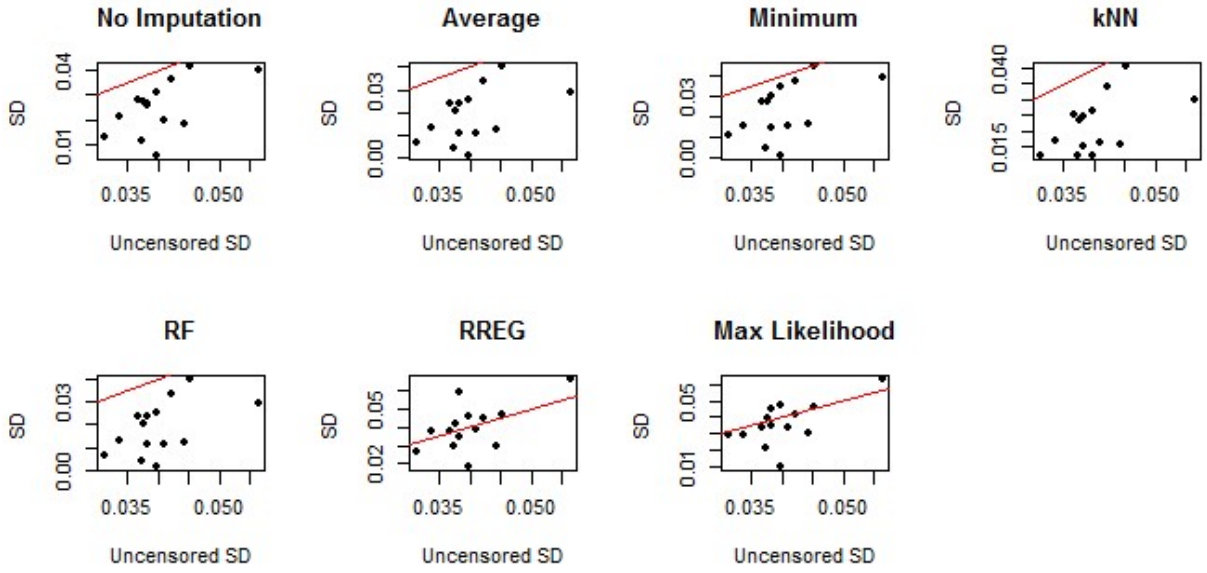


Figure 3.8: Estimated standard deviation versus sample mean in one simulation of normal data where values are censored from below and ρ=0.

deviation from one of $\{\boldsymbol{Y}_1^0, \ldots, \boldsymbol{Y}_{14}^0\}$ and the solid red line represents line y = x, indicating agreement between the uncensored value and the estimated value. All five common metabolomics methods over predict the mean parameter while under predicting the standard deviation. In the mean parameter especially, the estimated value appears to have little correlation with the uncensored value. In contrast, both of the proposed methods spread around the line y = x, particularly in the mean parameter where higher values of the uncensored parameter are approximated well. Of course, this is just one simulation with pairwise correlation 0. *Figures 3.9* and *3.10* plot the results for the mean and standard deviation of all 1,000 simulations for all seven methods at each level of ρ. As expected the value of ρ does not affect any of the five univariate procedures. Focusing on the mean, NONE, AVG and MIN always over predict. This result fits with what one would expect as the observed values are always higher than the full sample when missing values are censored from below. In fact, NONE and AVG are identical in the mean parameter since imputing with the average does not change the average. MIN demonstrates somewhat better behavior at higher values of the parameter. Again, due to the scaling after censoring higher values of the uncensored mean imply fewer missing values, so performance in this range is expected. As the proportion of censored values decreases the observed min is approaching the lowest the sample, meaning fewer values are having to be imputed and those that are being imputed are being approximated with a value closer to their true level. As with NONE, AVG and MIN, kNN and RF almost always over predict the mean. kNN sees some improvement with greater correlation between the variables, but even at $\rho = \frac{9}{10}$ there is a still a bias toward lower values. RF improves only marginally with increasing correlation. At $\rho = 0$ the profile for random forest is very similar to NONE or AVG, while at $\rho = \frac{9}{10}$ the profile is similar to MIN. RREG and ML share very similar profiles and both over and underestimate in

Figure 3.9: Error in mean under Normal and LOD simulations. X-axis is uncensored sample mean. Y-axis is predicted mean. Columns represent pairwise correlation between variables in data set. Red line is y=x.

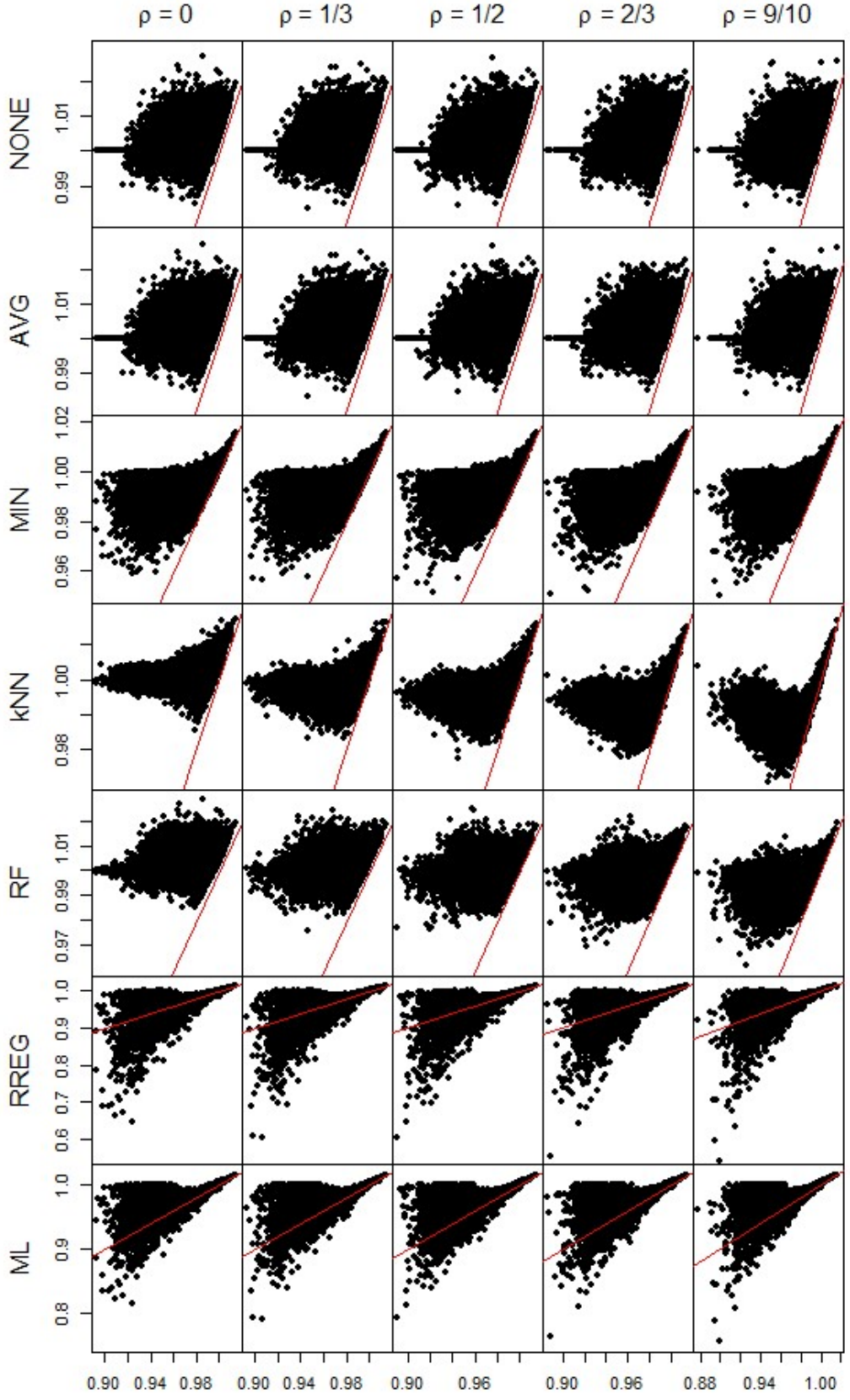
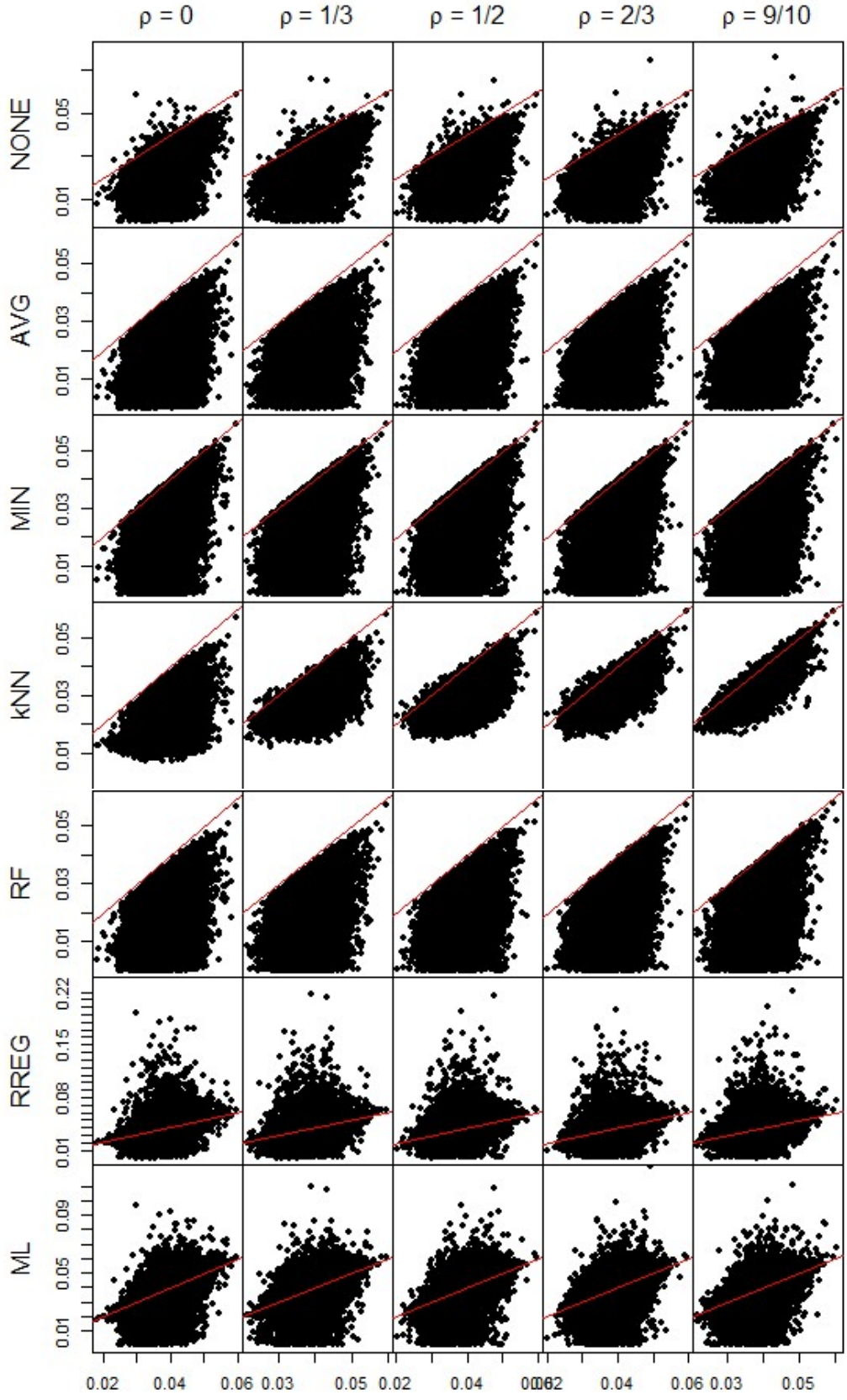


Figure 3.10: Error in SD under Normal and LOD simulations. X-axis is uncensored sample SD. Y-axis is predicted SD. Columns represent pairwise correlation between variables in data set. Red line is y=x.



roughly equal measure, suggesting less bias. Accuracy again tends to be better at the larger parameter values / lower proportions of missing values. The variation within the estimates for RREG and ML is quite large compared to the other methods as well. Across both profiles the estimates range from 0.75 to 1.01, while the other methods are largely confined to 0.95 to 1.03. The larger spread indicates that while RREG and ML are less biased, they are more variable as the amount of missing values increases. Bias in the standard deviation is present for the five common methods as well, but in this case the methods consistently under estimate the uncensored parameter. On occasions NONE does overestimate the standard deviation, but these are rare instances in which the variable is sufficiently right skewed. Notice that this behavior is absent from AVG. This is because imputed missing values with the observed average will always lower the standard deviation, effectively increasing the sample size while holding the variation constant. MIN and AVG are very similar in this regard, though MIN does tend to provide better mean estimates in variables that have higher means. Both kNN and RF behave similarly to AVG when the correlation is low. As correlation increases, RF more closely resembles MIN which is identical to the relationship observed in mean parameter. kNN displays two interesting qualities compared to the others in the standard deviation. First is a unique trait in which standard deviation estimates never decrease much below .01 while all the other methods have instances of basically reaching 0. Values of zero in the standard deviation imply imputation with a near constant value. That kNN avoids near 0 estimates of the standard deviation indicates the neighbors are producing distinct values to replace the missing observations. Second, as the correlation increases the variance in the estimates becomes tighter and less biased. At the highest levels of correlation this consistency is important, because while RREG and ML appear to be unbiased, both exhibit greater variability than the other methods. Estimated standard deviation

values for the five common methods over the entire experiment are generally between 0 and 0.06. Estimates for RREG, however, range from 0 to 0.26 – more than four times all the other methods. The spread for ML isn't as extreme as RREG with but does exhibit estimates well above 0.08.

Figures 3.11 and *3.12* plot the relative error for the mean and standard deviation, respectively, by the percentage of missing values. The results indicate a similar pattern as that seen in the previous set of plots showing the uncensored values against the estimated values; namely, that the five common metabolomics methods over predict in the mean parameter and underestimate the standard deviation. Unsurprisingly, relative error steadily increases for all five of these methods as the proportion of missing values increases, with a minimum at 100% which is, of course, the maximum percentage any estimator bounded by zero can decrease by. In the mean parameter NONE, AVG and MIN all have similar profiles with MIN showing slightly less relative error for moderate and high levels of missingness. Regardless of the value for ρ, kNN and RF are virtually indistinguishable from those methods in the mean parameter, with performance somewhat in between MIN and AVG/NONE. In contrast, error for RREG and ML remains centered around 0 for even large proportions of missingness. However, the variation in the errors increases dramatically as the proportion of missing values approaches 1. Both are prone to severe underestimates when the amount of missing values exceeds 80%, especially RREG. In the standard deviation, NONE performs better than AVG and MIN. This is the result of single imputation eliminating variation across the samples. RF continues to track closely with AVG and MIN regardless of ρ. KNN, however, shows great improvement for even low correlation of $\rho = \frac{1}{3}$ and continues to improve as the correlation increases.

Figure 3.11: Relative error in mean under Normal and LOD simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y=0.

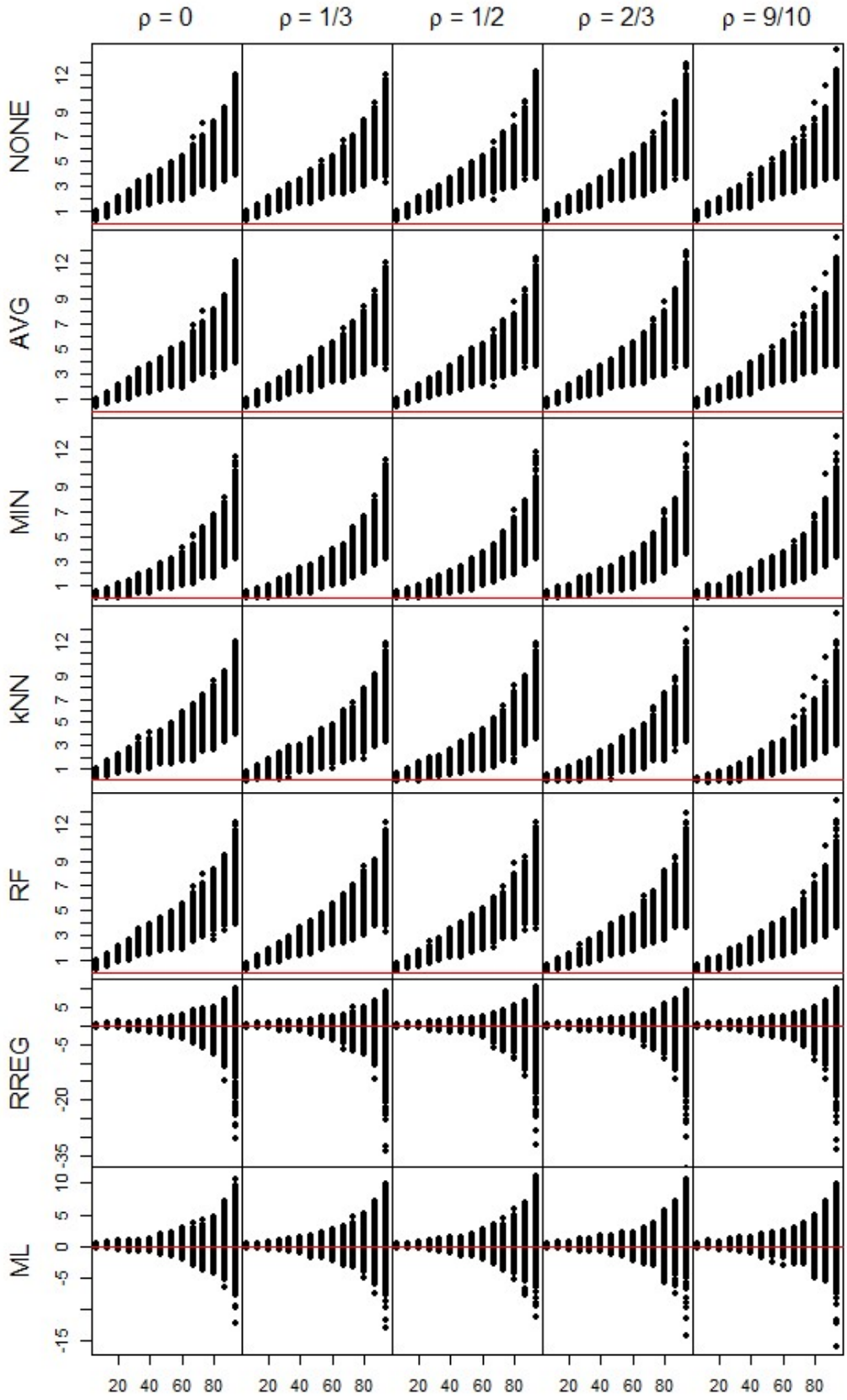
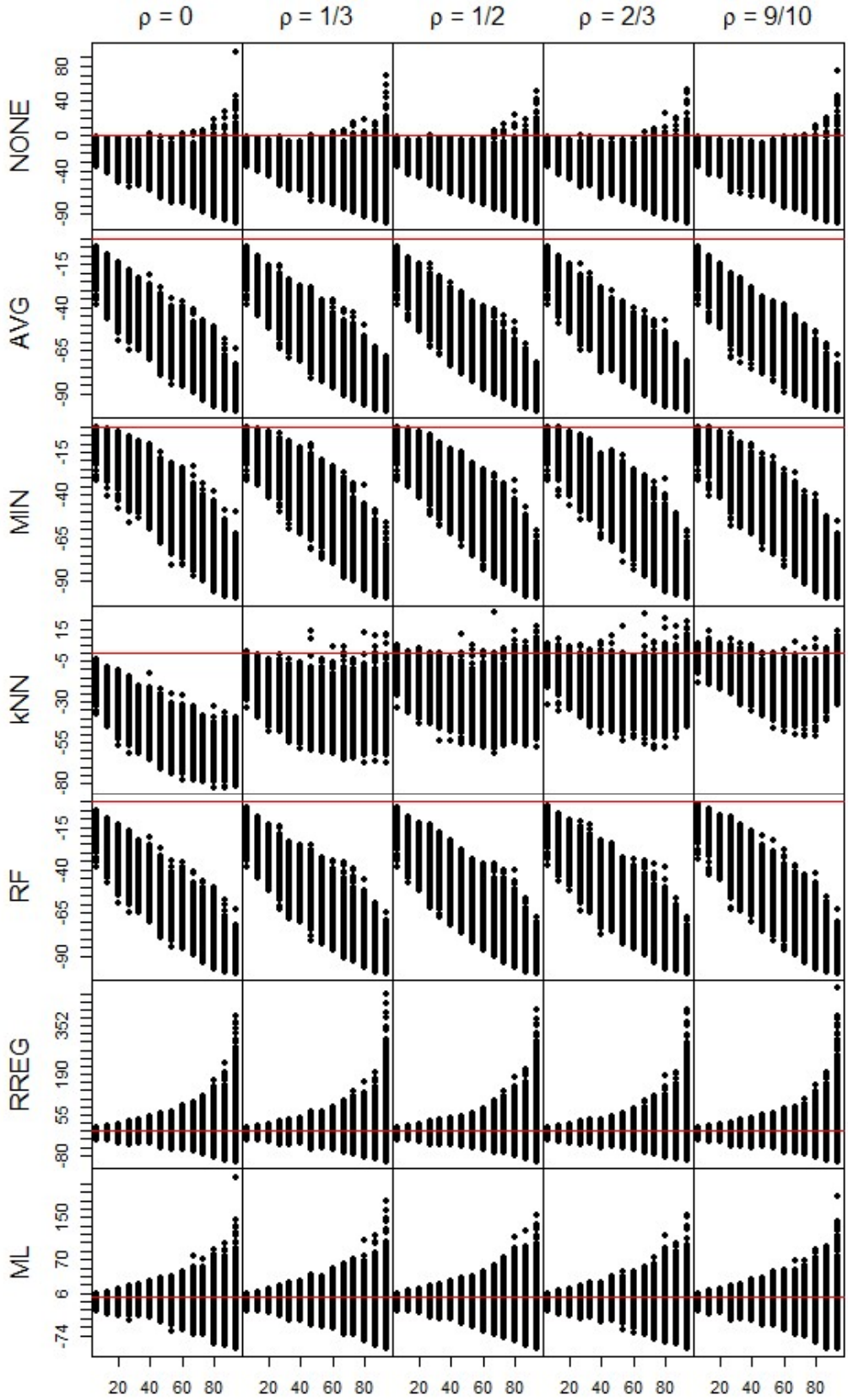


Figure 3.12: Relative error in SD under Normal and LOD simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y=0.



There is a curious behavior with kNN such that with $\rho = \frac{9}{10}$ the standard deviation estimate with the highest amounts of missing values are predicted better than those with moderate levels of missingness. As the proportion of missing values increases, the imputed variable under kNN is becoming more and more a reflection of the neighbors. Since the variables are on the same scale this results in the imputed variable becoming more and more like another random variable. However, the median scaling following variable censoring causes the neighbors to be shifted slightly upward, systematically shifting the imputed values high compared to the original values.

Simultaneously, the weighted average over the k neighbors reduces the overall variability in the imputed values. The result for high proportion of missing values is an imputed variable that is slightly higher on average and with lower variation than the original variable. In short, kNN is systematically replacing the censored variables with values that is slightly shifted due to the scaling and less variable as a result of aggregating across the neighbors. Meanwhile, RREG and ML have similar profiles with over and under estimates occur, just as they did in the mean parameter. Both of these methods display noticeably higher variance than the other methods as the amount of missing data becomes very large. However, the most dramatic errors in these two methods tend to be over-estimates rather the underestimates seen in the other five methods.

Bias plots, shown in *Figures 3.13* and *3.14*, take the average relative error as a function of the missing proportion for all ρ in $\{0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{9}{10}\}$. On average the five common methods overestimate the mean and underestimate the standard deviation, as expected. When $\rho = 0$, average relative error profiles for kNN (green) and RF (gray) are similar to that of AVG (red), though kNN does a little better with the standard deviation at higher levels of missingness. As ρ increases the profile for RF moves closer and closer towards that of MIN (yellow) in both parameters. kNN does the same in the mean parameter, but in the standard deviation it performs

Figure 3.13: Bias in mean parameter under Normal and LOD simulations. X-axis represents percentage of missing values. Y-axis is average relative error.

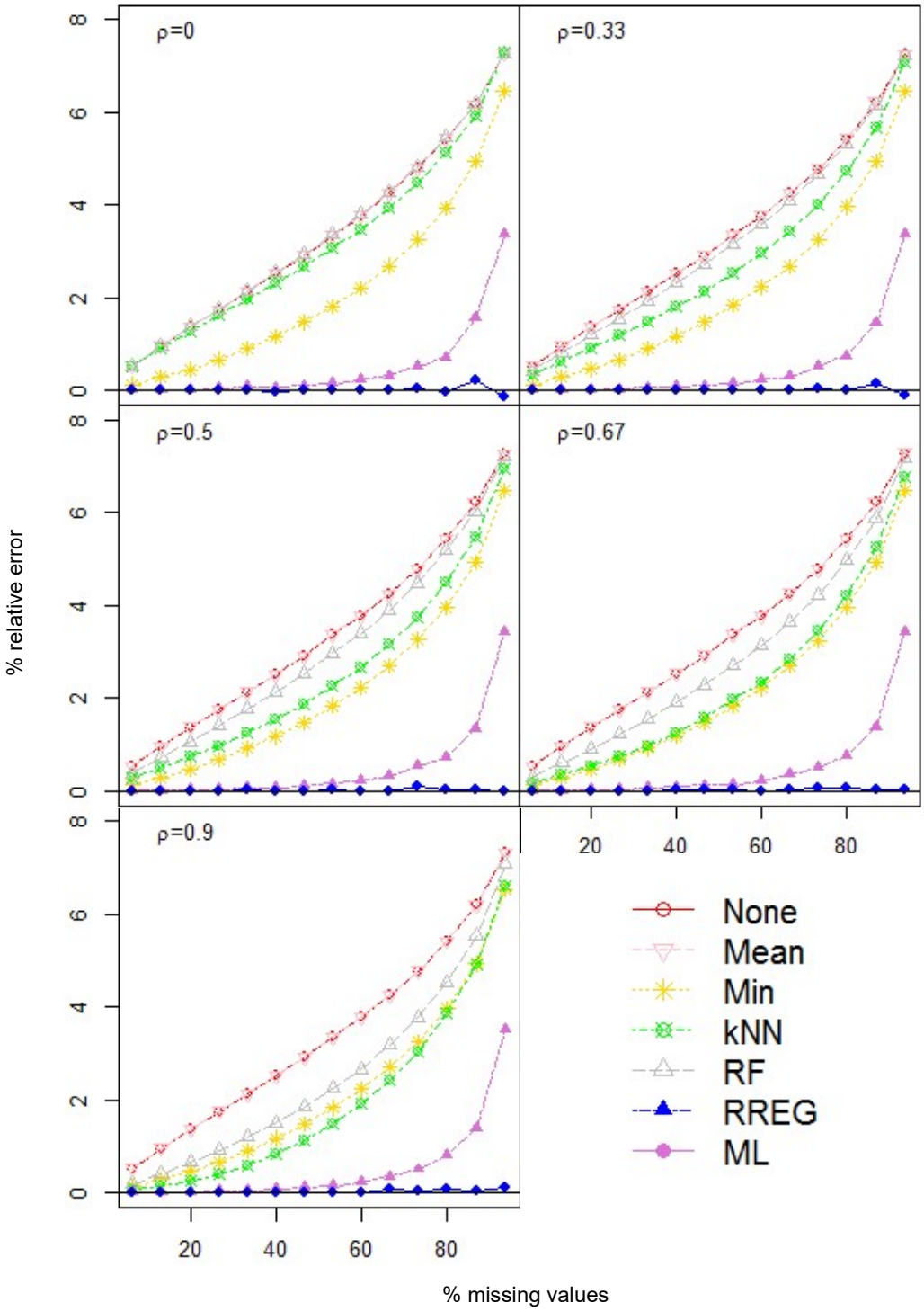
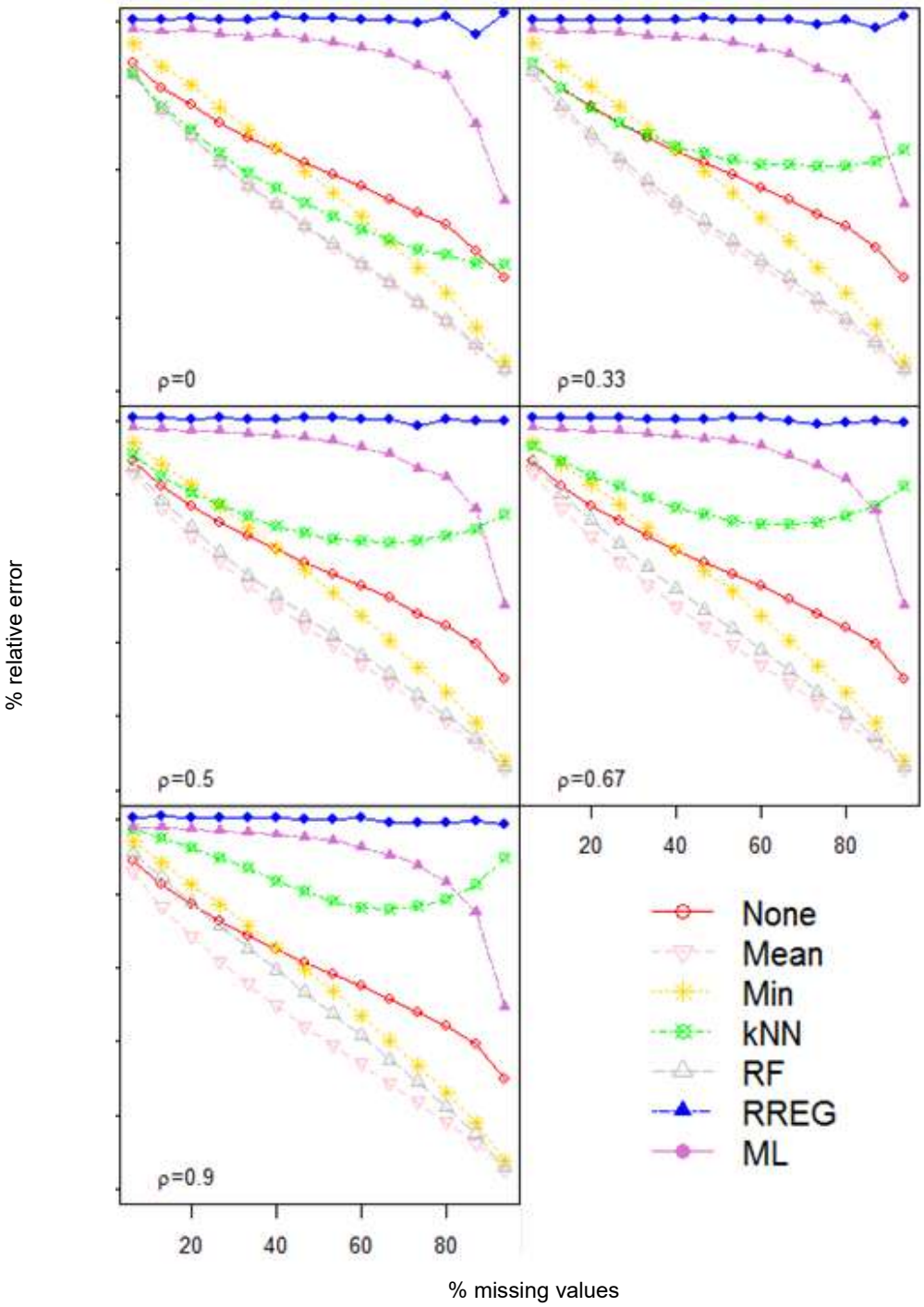


Figure 3.14: Bias in SD parameter under Normal and LOD simulations. X-axis represents percentage of missing values. Y-axis is average relative error.



better than other four common methods even when the correlation is as little as $\rho = \frac{1}{3}$. Average relative error for ML (purple) and RREG (blue) remain close to zero in both parameter up until missing rates of 80-90%. When the proportion of missing values reaches 80% and beyond, which in these simulations implies 6 or fewer observations, deviation from zero becomes noticeable with clear bias of over-estimation in the mean and under-estimation in the standard deviation. Though it is most evident at over 80% missing, the trend begins to emerge around 60%. Recall that ML is asymptotically unbiased and comes with no guarantee in small sample situations. These simulations suggest that noticeable bias begins for ML when the number of observed data points is around 10. RREG on the other hand remains unbiased in both parameter over the entire missing proportion range. Increasing variability does manifest for RREG, however, at around 80%.

An additional point, which can be extracted from the relative error plots, is that across all the seven methods used the relative error is much greater in the standard deviation than in the mean. On average increase in the mean is never more than 10% for any of the seven methods, but average decreases of 20% are rather common in the standard deviation. Such under estimates can have important consequences to outlier detection methods that rely on variance related measures to detect outliers. In the z-scores, the standard deviation is the denominator, meaning that a decrease of 50% would lead to a doubling of the z-score. Such decreases are common in the standard deviation for the five common methods when more than half of the 30 observations are missing. Such errors can lead to a large inflation of false positive rate. For example, in the normal distribution a value of 1.96 represents the upper 2.5% of the distribution. If one were to take this as a cutoff for identifying an outlier, then under estimates of 50 percent in the standard deviation would lead to any true z-score of 0.98 being classified as an outlier. Theoretically, such

an increase would lead to a false positive rate of 16.3%, more than six times the assumed rate. Technically, error will also be present in the mean parameter; however, error in the mean tends to be much lower, as a percentage of the true value, than error in the standard deviation. For example, at around 50% missing, average error in the mean is only 2-3% for the five common methods. So, it is reasonable to consider mainly the standard deviation when assessing impact on the z-score). An inflation of the false positive rate may be acceptable in certain situation. Screening tests for example may trade a higher false positive rate for increased power. However, in reality, there is a cost involved to investigating any subject flagged as a possible case. At some point the cost associated with an excessive number of false positives renders the test impractical.

Finally, the last item considered here is the variation of the methods. Tables 3.6 and 3.7 show the bias for the seven methods at each proportion of missing values as well as the variance

Table 3.6: Relative error averages and variances in Mean parameter from normal simulations with missing values left-censored. Correlation coefficient ρ=1/3.

Proportion Missing	NONE		NONE		NONE		NONE		NONE		NONE		NONE	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
6.7%	-11.06	0.01	-14.18	0.01	-5.87	0.01	-11.14	0.02	-13.42	0.01	-1.90	0.01	-0.85	0.01
13.3%	-17.83	0.02	-23.70	0.02	-11.79	0.02	-18.03	0.04	-22.67	0.02	-2.40	0.02	-1.10	0.03
20.0%	-22.96	0.05	-31.39	0.05	-17.22	0.03	-23.12	0.07	-30.14	0.04	-2.42	0.04	-1.16	0.06
26.7%	-27.39	0.06	-38.22	0.06	-22.89	0.04	-27.20	0.09	-36.89	0.06	-2.76	0.06	-1.35	0.09
33.3%	-31.40	0.09	-44.47	0.09	-28.95	0.06	-30.52	0.12	-43.08	0.08	-3.70	0.09	-1.80	0.16
40.0%	-34.94	0.12	-50.19	0.12	-34.61	0.09	-33.44	0.15	-48.69	0.11	-4.05	0.15	-1.87	0.23
46.7%	-38.05	0.16	-55.45	0.16	-40.28	0.12	-35.56	0.20	-53.96	0.15	-4.42	0.24	-1.77	0.39
53.3%	-41.40	0.23	-60.76	0.23	-46.39	0.15	-37.44	0.29	-59.22	0.21	-5.41	0.35	-2.35	0.57
60.0%	-44.84	0.28	-66.03	0.28	-52.93	0.20	-38.80	0.34	-64.48	0.24	-7.26	0.51	-3.10	0.86
66.7%	-47.89	0.38	-70.97	0.38	-59.18	0.26	-38.76	0.44	-69.34	0.36	-8.54	0.84	-3.33	1.38
73.3%	-52.10	0.51	-76.47	0.51	-66.57	0.38	-39.22	0.59	-74.98	0.48	-12.58	1.44	-5.56	2.38
80.0%	-55.36	0.74	-81.47	0.74	-73.38	0.55	-38.87	0.80	-80.20	0.71	-15.25	2.31	-5.36	3.98
86.7%	-61.36	0.97	-87.57	0.97	-82.11	0.74	-37.66	1.01	-86.60	0.92	-25.46	4.26	-9.48	7.89
93.3%	-69.40	2.00	-94.32	2.00	-92.10	1.70	-34.42	1.93	-93.94	1.96	-49.07	11.07	-10.45	34.02

Table 3.7: Relative error averages and variances in SD parameter from normal simulations with missing values left-censored. Correlation coefficient ρ = 1/3

Proportion Missing	NONE		AVG		MIN		kNN		RF		RREG		ML	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
6.7%	0.52	0.01	0.52	0.01	0.12	0.01	0.34	0.02	0.43	0.01	-0.01	0.01	0.00	0.01
13.3%	0.95	0.02	0.95	0.02	0.28	0.02	0.63	0.04	0.81	0.02	0.00	0.03	0.02	0.02
20.0%	1.37	0.05	1.37	0.05	0.46	0.03	0.92	0.07	1.19	0.04	0.00	0.06	0.02	0.04
26.7%	1.74	0.06	1.74	0.06	0.66	0.04	1.19	0.09	1.55	0.06	0.01	0.09	0.03	0.06
33.3%	2.12	0.09	2.12	0.09	0.91	0.06	1.48	0.12	1.93	0.08	0.02	0.16	0.06	0.09
40.0%	2.52	0.12	2.52	0.12	1.17	0.09	1.81	0.15	2.31	0.11	0.03	0.23	0.08	0.15
46.7%	2.91	0.16	2.91	0.16	1.47	0.12	2.13	0.20	2.71	0.15	0.03	0.39	0.10	0.24
53.3%	3.36	0.23	3.36	0.23	1.83	0.15	2.55	0.29	3.17	0.21	0.06	0.57	0.15	0.35
60.0%	3.77	0.28	3.77	0.28	2.22	0.20	2.95	0.34	3.59	0.24	0.08	0.86	0.23	0.51
66.7%	4.24	0.38	4.24	0.38	2.68	0.26	3.44	0.44	4.09	0.36	0.10	1.38	0.31	0.84
73.3%	4.78	0.51	4.78	0.51	3.26	0.38	4.00	0.59	4.65	0.48	0.21	2.38	0.52	1.44
80.0%	5.43	0.74	5.43	0.74	3.96	0.55	4.74	0.80	5.32	0.71	0.24	3.98	0.74	2.31
86.7%	6.22	0.97	6.22	0.97	4.95	0.74	5.68	1.01	6.14	0.92	0.50	7.89	1.45	4.26
93.3%	7.26	2.00	7.26	2.00	6.46	1.70	7.07	1.93	7.23	1.96	0.58	34.02	3.37	11.07

associated with the proportion. For simplicity only, the values are shown for $\rho = \frac{1}{3}$ as this level is consistent with the average correlation seen in the real metabolomic sets in Chapter 2. These tables confirm that while ML and RREG are more consistent on average with the sample parameters these methods also demonstrate more variability when missing proportion is above 60%, and this variation is much greater when the proportion is above 80%. Variance of the estimates under the five common methods steadily increase as missing proportion increases, but tend to stay around or under 1 up to the most extreme missing proportion of 93.3%. The variance in ML and RREG approaches a value of 1 around 60% missing values and increase rapidly from then on.

2.14.2.2. Large Sample Normal Simulations and LOD

To examine the influence of sample size, the same simulation was repeated using $n = 600$. The number of samples censored is increased proportionally according to $20*j$ for $j \in \{1, \ldots, 14\}$.

This way the proportion of values missing is identical to the previous simulation, but the number of available samples remaining is twenty times greater. Maintaining the same missing proportion while also having a large sample size will help answer whether results are dependent on the proportion of data available or on the number of samples available. Given that normal-LOD simulations have already been explored, only the relative error and bias plots are shown for the large sample setting (*Figures 3.15-18*). The same y-axes ranges were kept for comparison, and the general trend remains the same for all methods. However, it is easy to see that variation in the estimates is greatly reduced. The bias plots show that both ML and RREG remain almost completely unbiased in both parameters throughout the entire range, implying that the number of available samples is more important that the percentage of available values.

2.1.1.1.Small Sample Normal Simulations and LOD

Analogous to the large sample simulation, a small simulation with $n = 10$ is conducted next. Due to the number of available samples, the number of samples being censored was modified to j for $j \in \{1, \ldots, 8\}$. For this simulation only, the bias plots are shown. These are shown in *Figures 3.19* and *3.20*. There a couple notable results in this simulation. First is that in the mean parameter, bias is not found to be as extreme in any of the methods. This is because these simulations end when 8/10=80% of the values are missing as opposed to 28/30=93% in the simulations with $n = 30$. In those simulations ML was found to deteriorate quickly when more than 80% of the values were missing. The deterioration is stronger here, and more noticeably in the standard deviation with average under-estimates of 20% when 5 of 10 values are missing. Combined with the results of the large sample simulation, the accuracy of ML is tied most strongly to the number of available samples rather than proportion of missing values. RREG continues to be unbiased in both parameters. However, the estimates produced by this method

Figure 3.15: Relative error in mean under large sample Normal-LOD simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y=0.

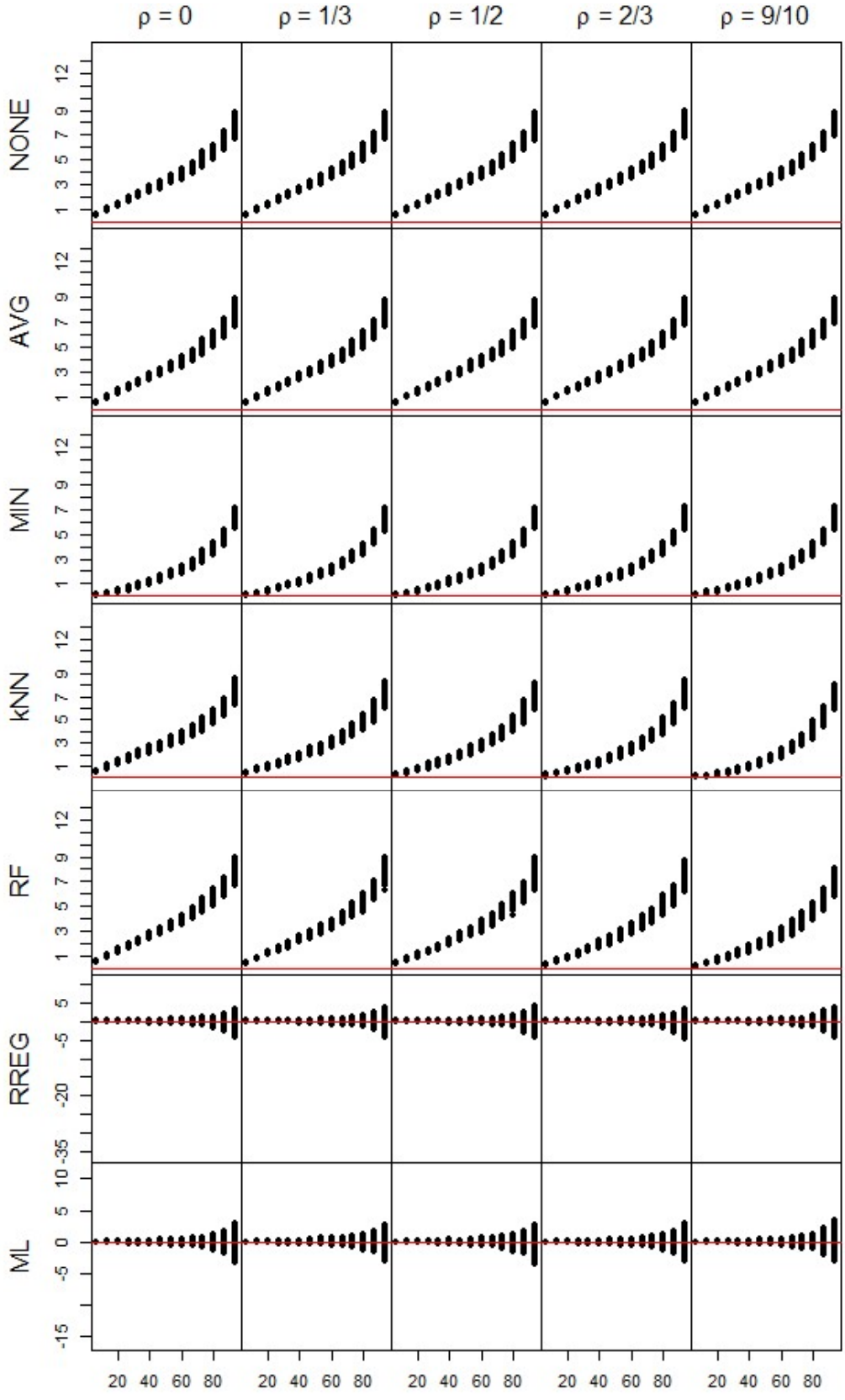


Figure 3.16: Relative error in mean under large sample Normal-LOD simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y=0.

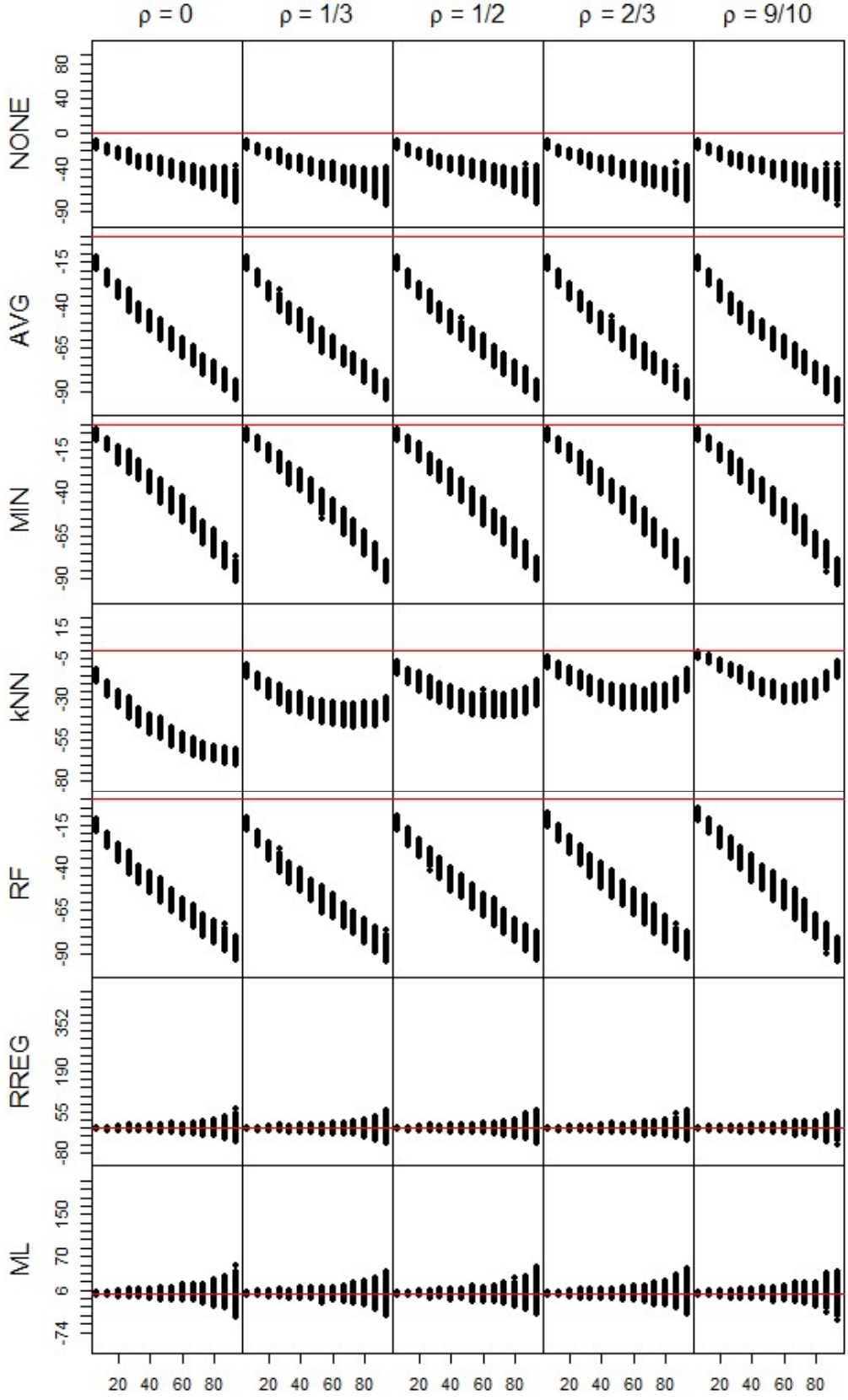


Figure 3.17: Bias in mean parameter under large sample Normal-LOD simulations. X-axis represents percentage of missing values. Y-axis is average relative error.

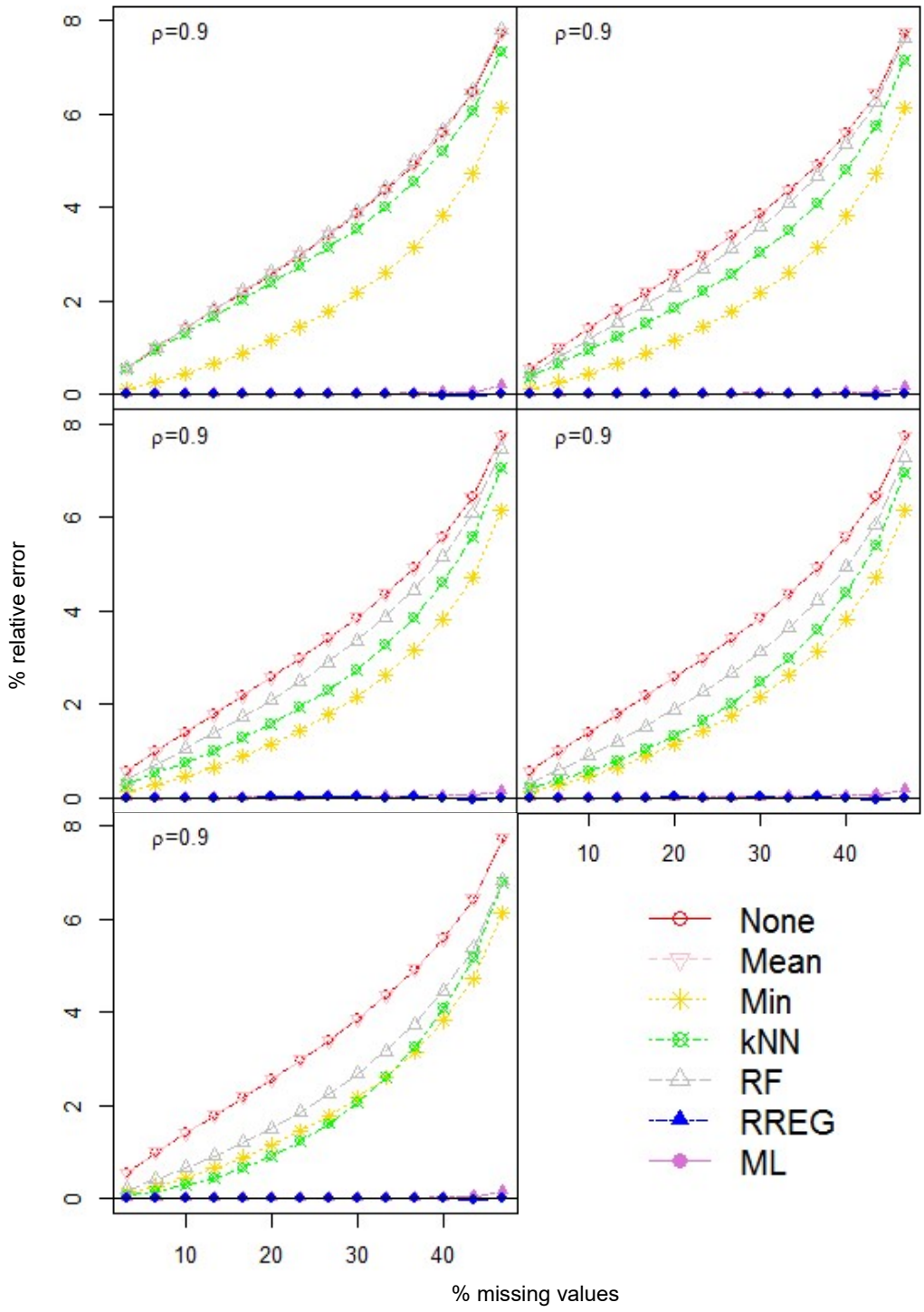


Figure 3.18: Bias in standard deviation parameter under Normal and LOD simulations. X-axis represents percentage of missing values. Y-axis is average relative error.

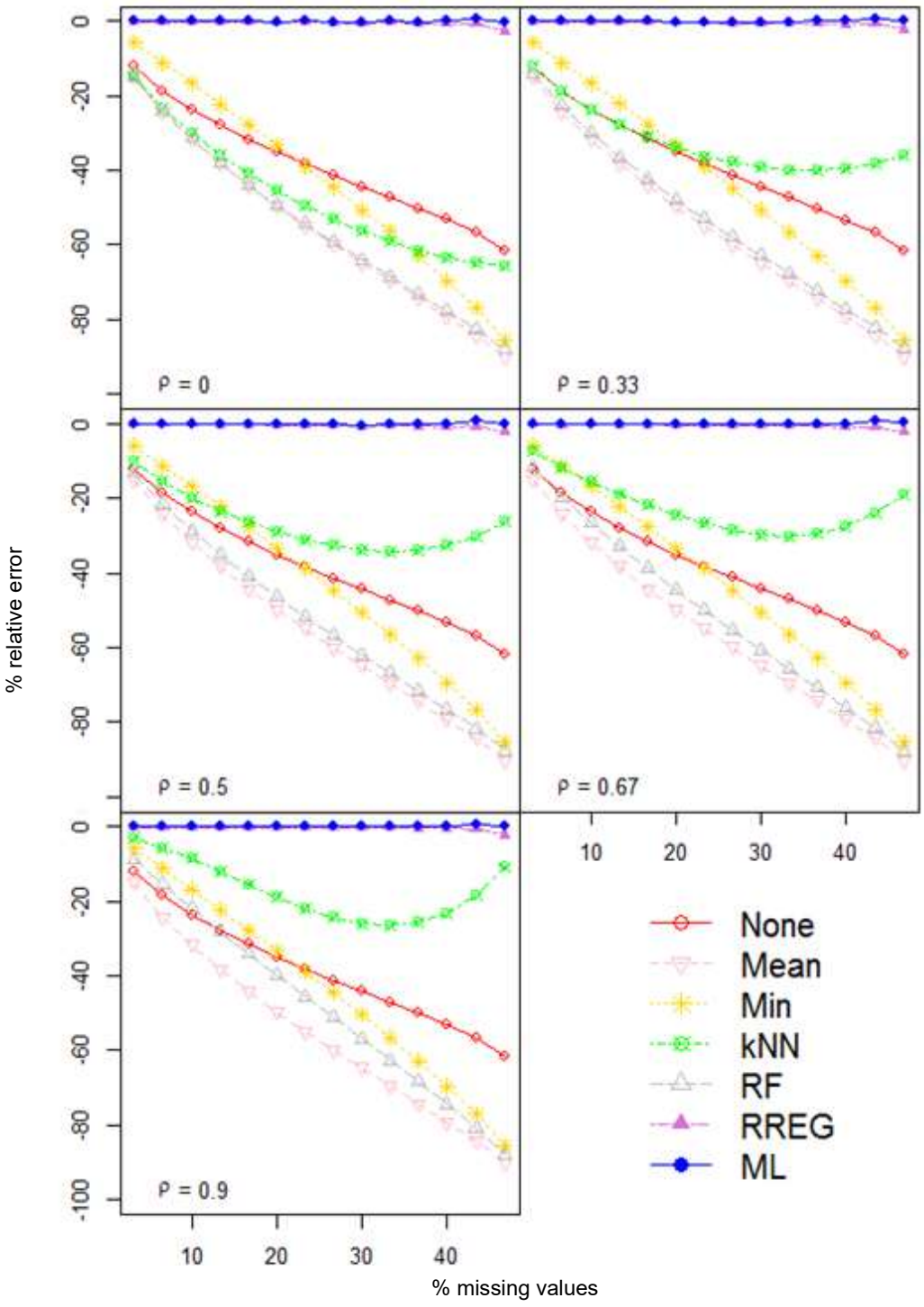


Figure 3.19: Bias in mean parameter under small sample Normal-LOD simulations. X-axis represents percentage of missing values. Y-axis is average relative error.

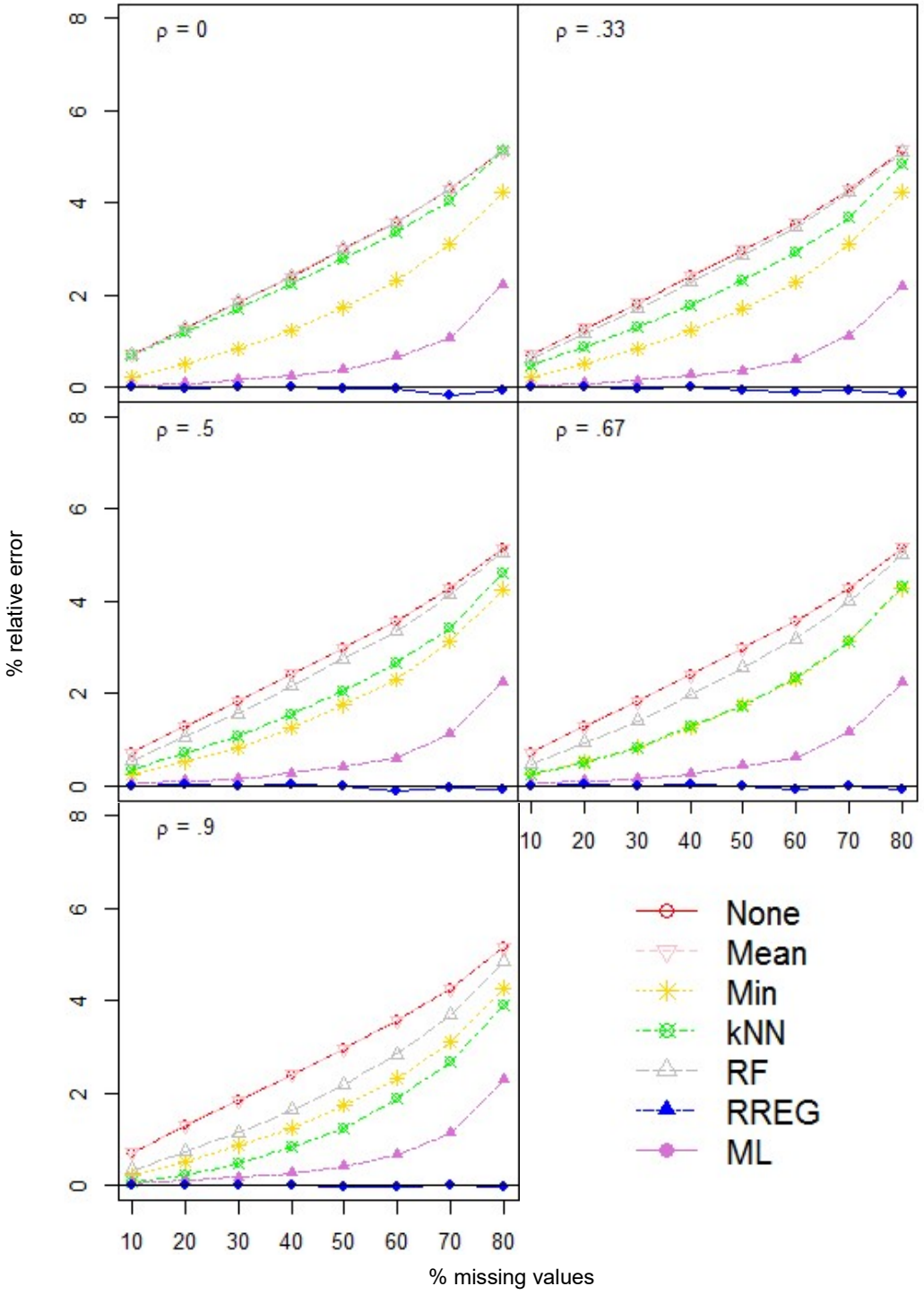
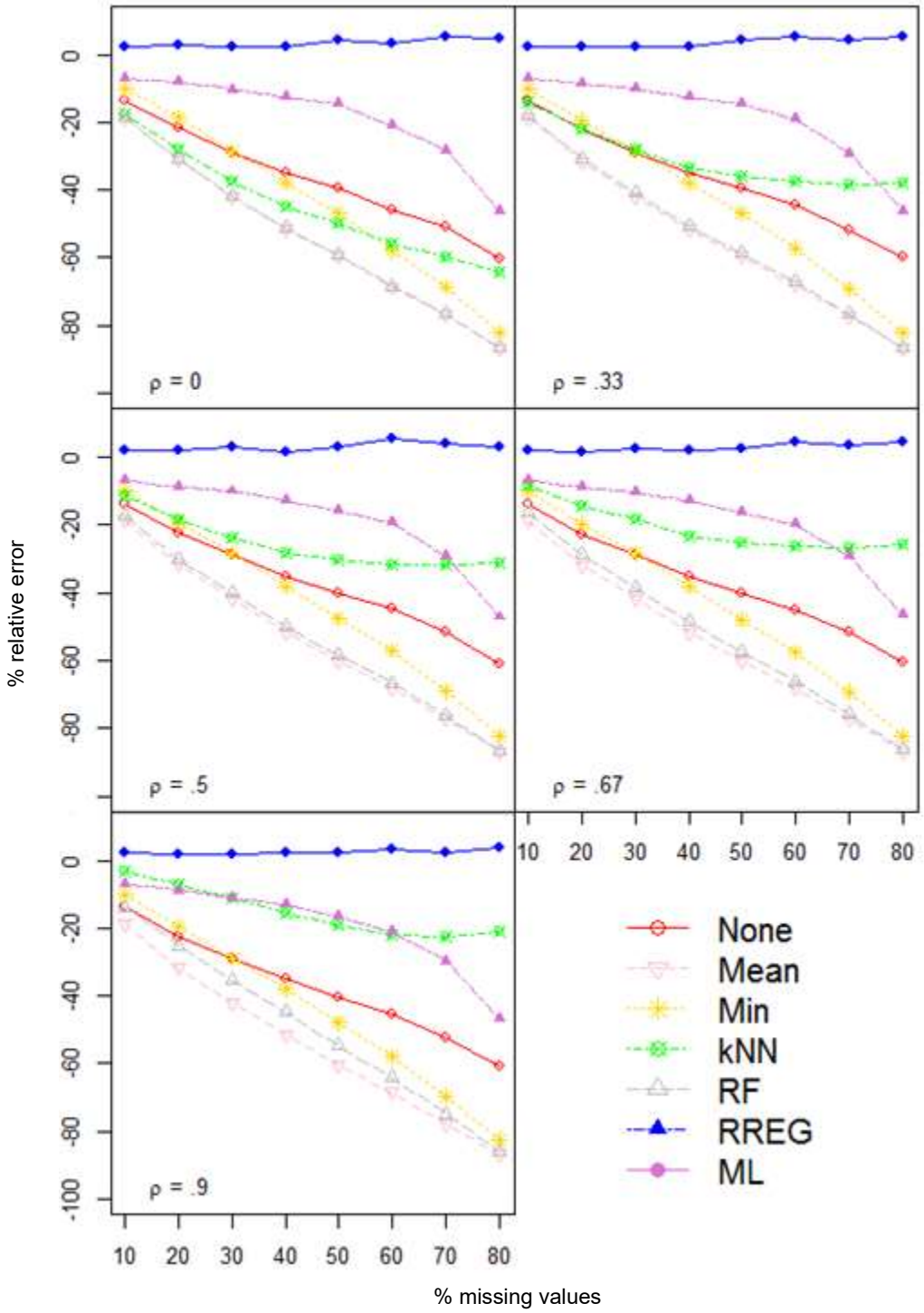


Figure 3.20: Bias in standard deviation parameter under small sample Normal-LOD simulations. X-axis represents percentage of missing values. Y-axis is average relative error.



continue to show much higher variability than ML or any other method.

2.1.1.1.Normal Simulations and MAR

The following simulations explore sensitivity to the maximum likelihood and rankit regression assumptions. Here we begin by removing values at random instead of in a left censored fashion. Multivariate normal datasets are simulated and processed as before, with the only exception being that the 2*j values removed from $\boldsymbol{Y_j}$ are chosen randomly rather than by being the 2*j lowest values. Results are displayed using the same plots and tables for error, relative error and average relative error as before. *Figures 3.21* and *3.22* plot the error by the true parameter for normally distributed data with missing values removed at random. In evaluating of methods at estimating the mean, none of the methods display much of a pattern with regard to true parameter value. NONE, AVG, kNN and RF give estimates that cluster around the true level (red line). MIN overestimates the mean in the rare occasion that the values removed at random happen to favor the lowest values in the set. In general, though, underestimates are more common because MIN is a biased estimate of a randomly remove value and is best suited to the left-censored scenario. Over estimates occur in the rare instances that the values removed happened to be among the lowest in the variable, in which case minimum imputation is appropriate. But in most cases, replacing values removed at random with the observed minimum will result in a lower value for the imputed observation, and thus an under estimate.

When applied to simulated data sets with observations MAR, kNN shows little association under low correlation, but as correlation increases, true mean values tend to be predicted with a mean above 1 while true mean values below 1 tend to be predicted being below 1. Under low correlation RF again behaves very similarly to NONE and AVG, but as correlation increase its profile becomes more like kNN. RREG and ML have similar profiles to MIN, generally underestimating the mean, though in a more extreme manner. The smallest estimate produced by

Figure 3.21: Error in mean under Normal and MAR simulations. Columns represent pairwise correlation between variables in data set. Red line represents y=x.

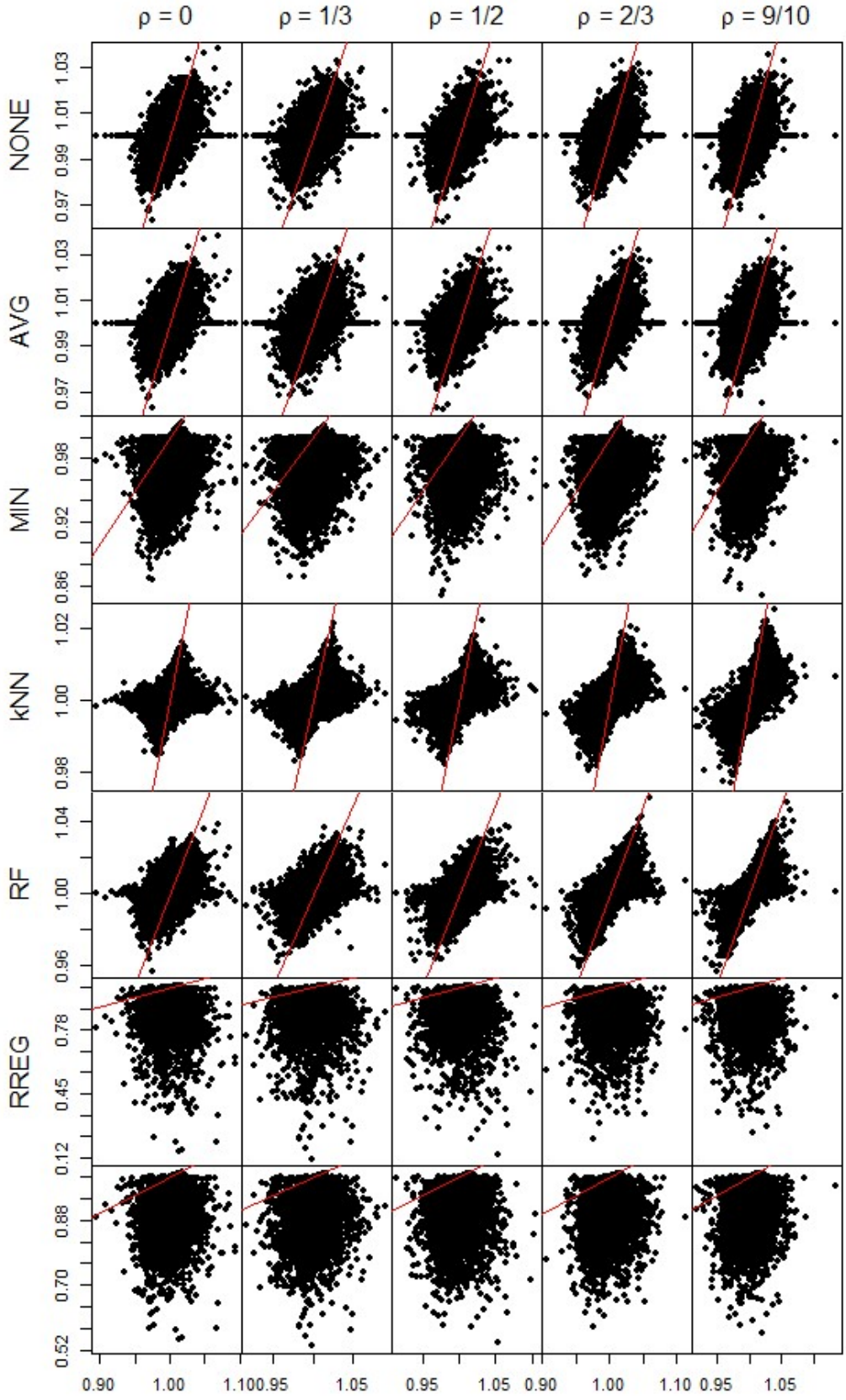
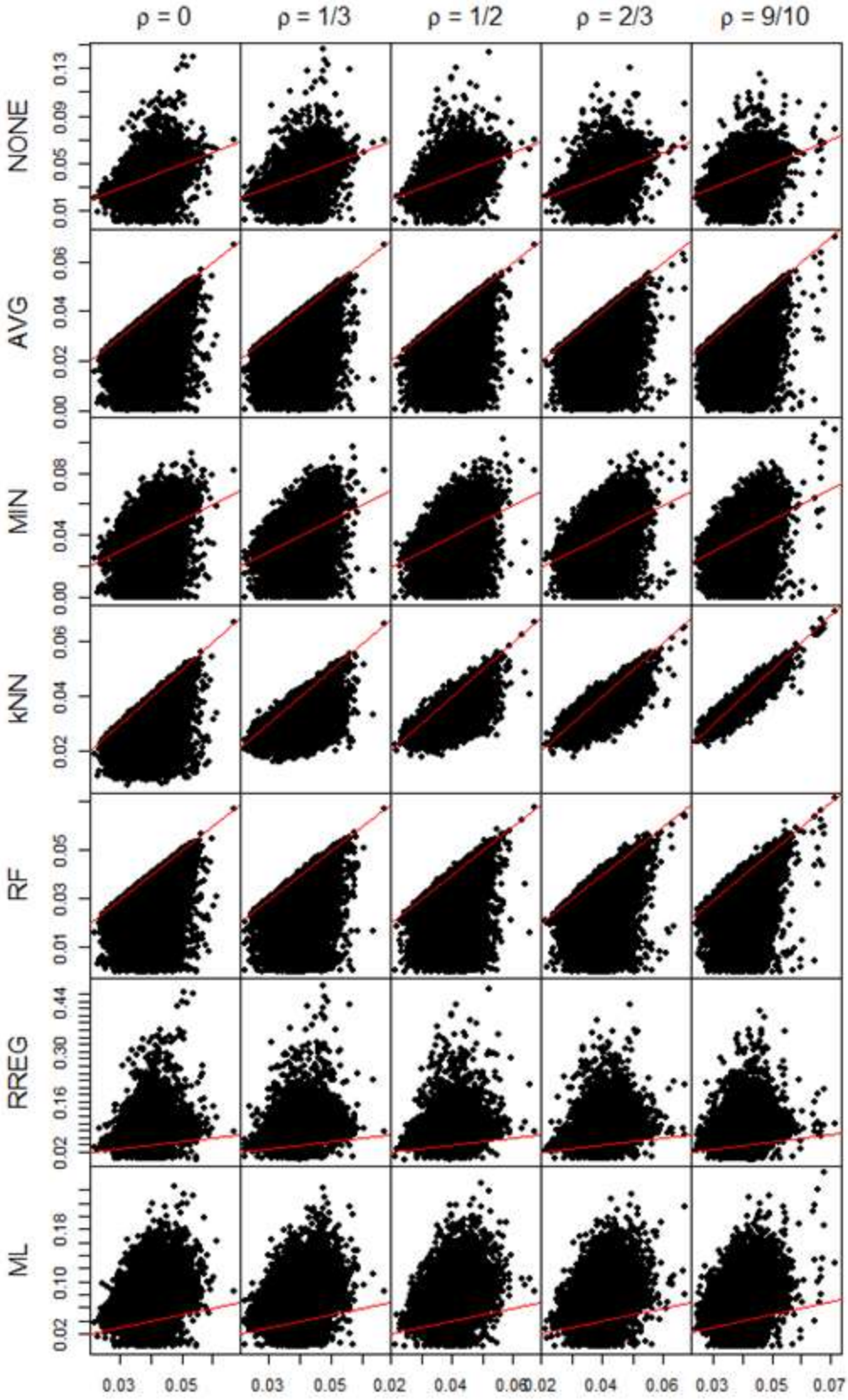


Figure 3.22: Error in SD under Normal and LOD simulations. Columns represent pairwise correlation between variables in data set. Red line represents y=x.



MIN was above 0.85, while ML produces estimates below 0.4 and RREG gives estimates below 0.2.

In the standard deviation, NONE over and underestimates that appear randomly distributed around the true value. AVG always produces lower estimates of the standard deviation, again the further result of imputing with a static value that is equal to the sample average and therefore only serves to depress the observable variation in the dataset. KNN and RF again match very well with AVG under low correlation. MIN, unlike AVG, produces over and under estimates of the standard deviation routinely. While this too imputes with a static value, using the observed minimum can add variation to the imputed variable as a value that is on average equal to the mean is being removed and replaced with a relatively extreme (and low) value.

As the correlation increases RF continues to perform about the same whereas kNN demonstrates a clear linear fit between its estimated standard deviation and the true standard deviation. Meanwhile both RREG and ML continue to display a more extreme form of minimum imputation. Both methods will over and under estimate the standard deviation, but frequently produce extreme over estimates with values of .2 for ML and 0.3 for RREG.

Next, *Figures 3.23-26* show the relative error and average relative error against the proportion of missing values. NONE, AVG, kNN and RF indicate they are unbiased estimators of the sample mean while MIN, RREG and ML tend to underestimate. In the standard deviation, NONE appears unbiased while AVG increasingly underestimates the variation in the data as the proportion of missing values increases. MIN tends to overestimate, but as the proportion of missing values increases and resulting variable becomes more and more imputed the standard deviation begins to lower. At higher amounts of missing values MIN underestimates as a result. When there is no correlation between the variables, kNN and RF behave much like AVG, but

Figure 3.23: Relative error in mean under Normal and MAR simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y=0.

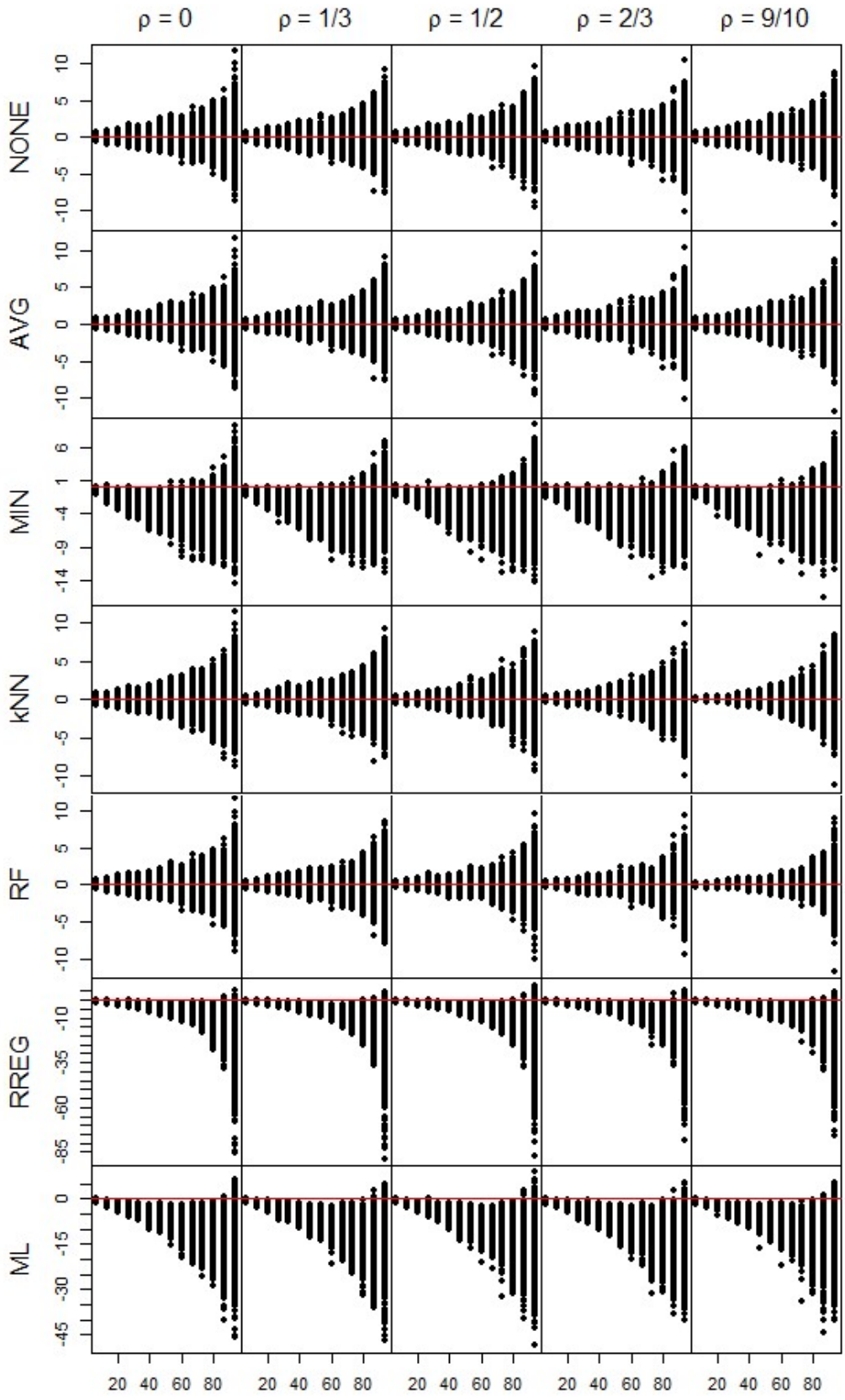


Figure 3.24: Relative error in SD under Normal and MAR simulations. *X*-axis in the proportion of missing values. *Y-a*xis represents percent change from uncensored sample value. Red line represents y=0.

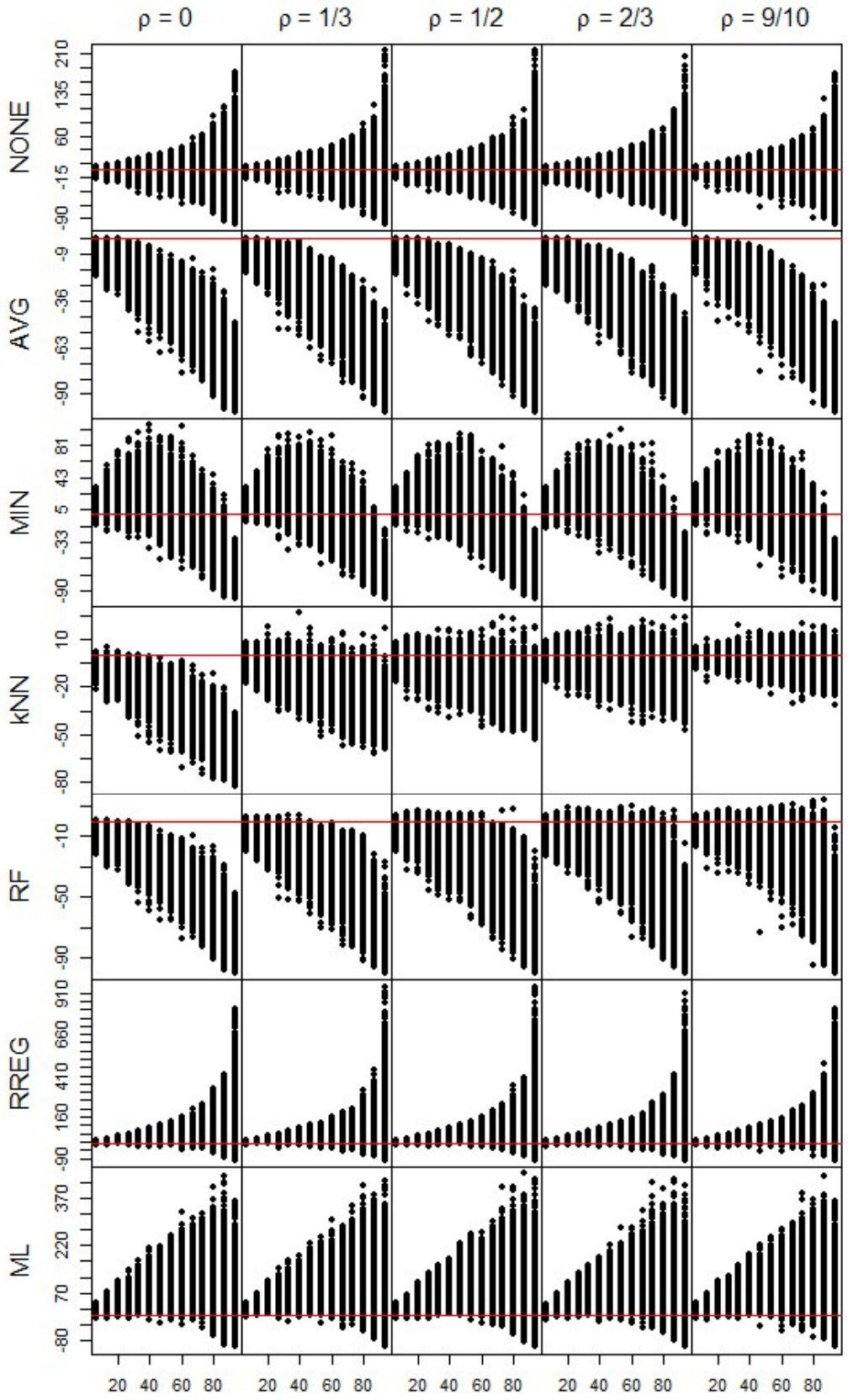


Figure 3.25: Bias in mean parameter under Normal and MAR simulations. X-axis represents percentage of missing values. Y-axis is average relative error.

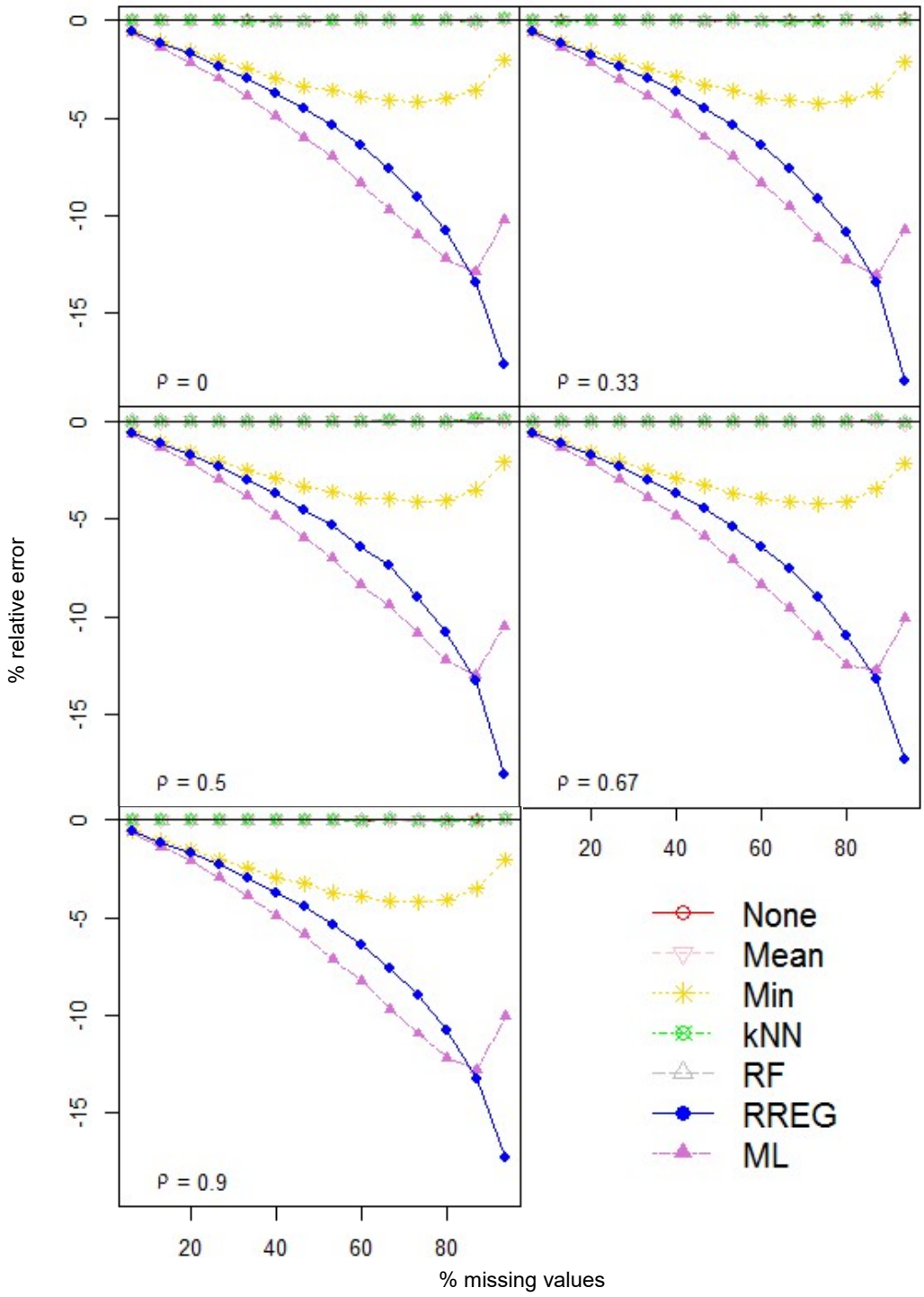
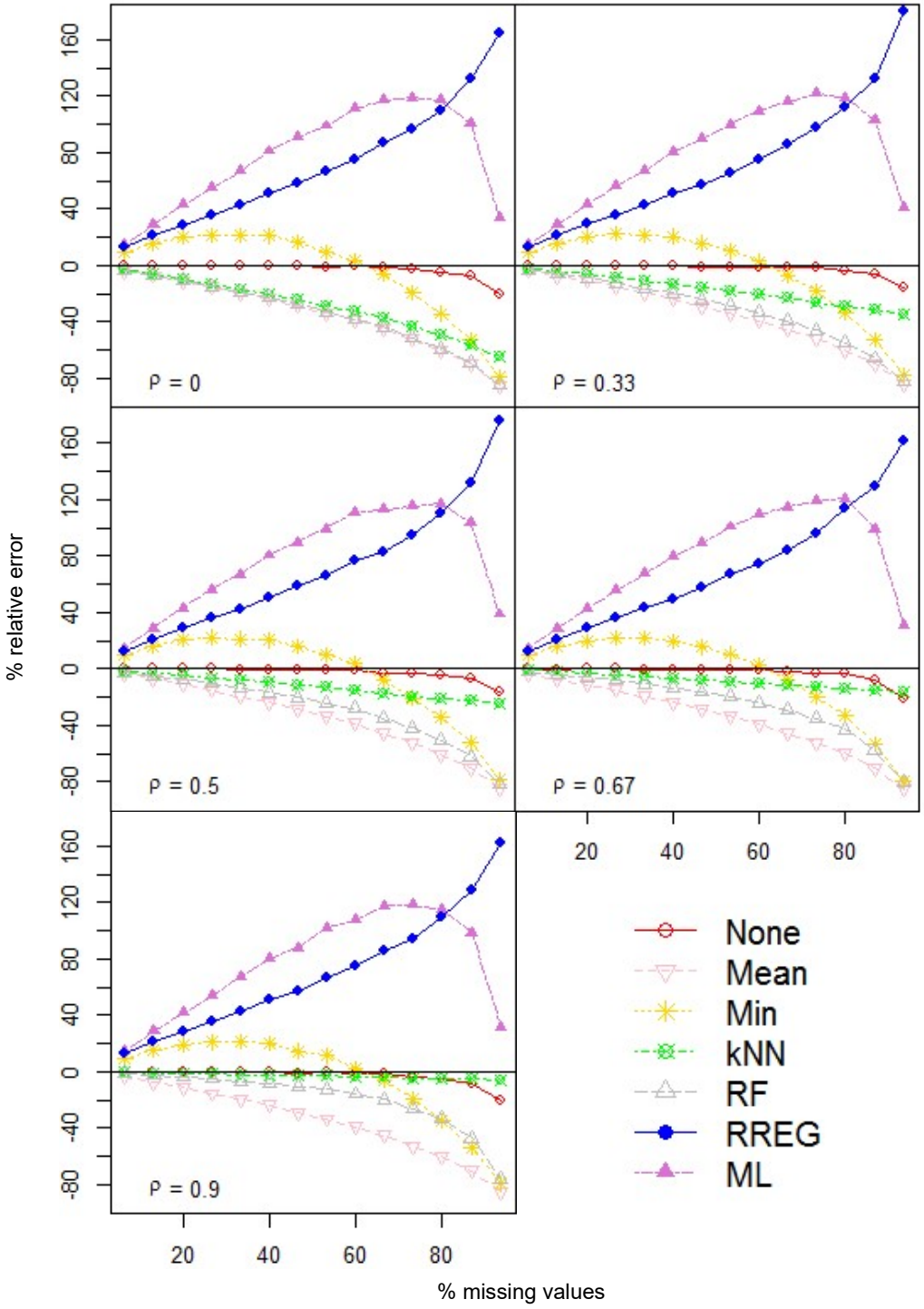


Figure 3.26: Bias in SD parameter under Normal and MAR simulations. X-axis represents percentage of missing values. Y-axis is average relative error.



become less biased as ρ increases. KNN appears roughly unbiased at $\rho = 0.9$ and the estimates are generally contained within 20% of the true sample value. Though through RF estimates improves over MEAN imputation by becoming less biased at correlation increases, the method never becomes completely unbiased even at $\rho = 0.9$. Contrasting with the others methods, RREG and ML are biased upwards in the standard deviation and do so in an extreme fashion consistent with MAR results thus far. Overestimates of 100-200% are not uncommon when the proportion of missing values is around 50%. The other methods are generally contained to 50-60% for the same amount of missing values. The final items in the assessment of normal data with values missing at random are Tables 3.8 and 3.9 showing the average relative error at each proportion of missing for all seven methods when the correlation is one third. These tables reveal a picture that is reversed from left-censored missingness in the mean parameter with the five common imputation approaches being nearly unbiased while the two proposed methods are increasingly

Table 3.8: Relative error averages and variances in Mean parameter from normal simulations. Missing values are MAR and $\rho = 1/3$.

Proportion Missing	NONE		AVG		MIN		KNN		RF		RREG		ML	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
6.7%	0.01	0.04	0.01	0.04	-0.53	0.06	0.00	0.03	0.00	0.03	-0.59	0.04	-0.66	0.06
13.3%	-0.01	0.08	-0.01	0.08	-1.06	0.14	-0.02	0.07	-0.01	0.06	-1.16	0.11	-1.39	0.16
20.0%	0.01	0.13	0.01	0.13	-1.56	0.27	0.01	0.11	0.02	0.10	-1.73	0.20	-2.16	0.34
26.7%	0.00	0.20	0.00	0.20	-2.06	0.46	0.00	0.18	-0.01	0.16	-2.34	0.33	-3.02	0.64
33.3%	0.04	0.28	0.04	0.28	-2.45	0.69	0.04	0.25	0.05	0.21	-2.94	0.54	-3.84	1.08
40.0%	0.02	0.38	0.02	0.38	-2.91	0.94	0.03	0.36	0.02	0.28	-3.66	0.82	-4.86	1.69
46.7%	-0.03	0.51	-0.03	0.51	-3.35	1.36	-0.02	0.45	-0.02	0.37	-4.52	1.22	-5.98	2.68
53.3%	0.02	0.57	0.02	0.57	-3.59	1.64	0.01	0.59	0.02	0.43	-5.33	1.83	-7.00	3.86
60.0%	-0.04	0.76	-0.04	0.76	-3.96	2.02	-0.03	0.85	-0.03	0.61	-6.41	2.64	-8.34	5.69
66.7%	0.01	1.00	0.01	1.00	-4.09	2.59	-0.03	1.12	0.00	0.80	-7.59	4.39	-9.57	8.57
73.3%	-0.01	1.45	-0.01	1.45	-4.25	3.74	-0.04	1.62	-0.01	1.21	-9.11	7.59	-11.13	14.71
80.0%	0.06	2.21	0.06	2.21	-4.04	4.60	0.05	2.45	0.05	1.91	-10.86	14.49	-12.31	21.95
86.7%	-0.05	3.61	-0.05	3.61	-3.63	6.85	-0.03	3.86	-0.03	3.23	-13.45	38.20	-13.06	40.92
93.3%	0.08	7.29	0.08	7.29	-2.14	9.85	0.06	7.08	0.07	7.11	-18.47	201.17	-10.71	72.39

Table 3.9: Relative error averages and variances in SD parameter from normal simulations. Missing values are MAR and $\rho = 1/3$.

Proportion Missing	NONE		AVG		MIN		KNN		RF		RREG		ML	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
6.7%	0.0	11.9	-3.5	11.0	9.8	37.2	-2.0	11.0	-2.9	11.1	10.8	17.1	14.6	42.4
13.3%	-0.3	26.5	-7.4	22.8	15.8	95.3	-4.4	21.2	-6.2	22.9	18.9	41.8	28.7	124.8
20.0%	0.2	41.3	-10.8	32.7	20.6	172.0	-6.0	32.0	-8.8	34.3	26.7	74.3	43.3	260.6
26.7%	-0.2	64.3	-15.0	46.5	22.5	249.7	-8.4	44.1	-12.5	51.0	33.0	128.1	56.3	439.6
33.3%	-0.8	87.7	-19.7	57.5	21.2	315.6	-11.0	52.1	-16.4	64.6	39.1	186.7	67.0	650.3
40.0%	-0.2	110.3	-23.6	64.7	20.6	342.7	-12.9	58.0	-19.7	74.7	46.6	260.0	80.6	840.3
46.7%	-1.1	154.5	-28.9	79.9	15.8	407.3	-15.8	71.5	-24.3	98.4	52.6	408.2	89.9	1200.5
53.3%	-1.0	210.7	-33.7	94.4	10.1	460.2	-18.1	79.7	-28.7	117.6	60.4	592.4	99.8	1663.8
60.0%	-1.1	285.5	-39.1	108.3	2.8	467.0	-20.8	87.6	-33.5	139.1	68.3	907.3	109.2	2119.3
66.7%	-1.4	356.1	-45.1	110.5	-7.0	475.2	-23.2	95.9	-39.2	157.1	77.6	1252.3	115.8	2832.6
73.3%	-1.9	511.6	-51.8	123.5	-18.5	486.1	-26.2	105.9	-46.1	179.3	87.3	2002.4	121.6	3965.9
80.0%	-3.6	751.6	-60.0	129.6	-33.8	433.4	-29.0	104.5	-54.8	191.9	98.3	3415.9	118.8	5164.9
86.7%	-6.8	1436.9	-70.0	148.6	-52.8	404.5	-31.7	111.4	-66.1	210.5	113.6	7807.4	102.6	7907.8
93.3%	-15.6	3701.5	-84.3	127.6	-78.2	246.8	-35.2	105.9	-82.6	166.4	147.1	31708	40.5	10255

biased towards underestimates. However, as with left-censored missingness, this bias, as a proportion of the sample mean, does remain relatively small at less than 10% for moderate to intermediate amounts of missing values.

With regard to the standard deviation, none of the methods are revealed to be unbiased as the five common methods underestimate the standard deviation on average while both proposed methods overestimate. This is not entirely unexpected for AVG and MIN, due to the nature of single imputation, nor for KNN and RF which under correlation of 0.33 have thus far functioned similarly to AVG. It is somewhat surprising that NONE underestimates with increasing magnitude as the amount of missing values increases since the remaining values reflect the true population. As such, one might expect this method to be an unbiased estimator of both the mean and the standard deviation when values are removed at random. The reason for this is the distribution of the sample standard deviation follows a scaled chi-square distribution.

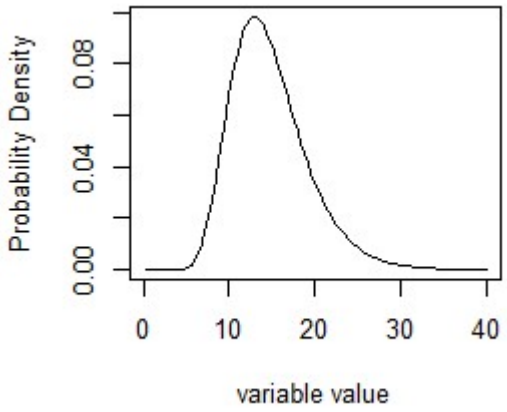
Asymptotically this distribution approaches a normal, by which the symmetry of the distribution allows the relative error to equally balance out. However, as the number of missing values increases, the chi-square creates a more and more right skewed data. The asymmetry causes the relative error to be below zero on average. Aside from this issue, NONE arguably shows the best performance when missing values are MAR, having the least average relative error, although the variance becomes quite large compared to the other methods for moderate to large amounts of missing values. RREG and ML show the worst performance, overestimating the standard deviation by 10% when just two values are removed. Bias in these two methods are routinely 2 to 10 times higher than the five common methods and variation of these estimators is enormous in comparison. Neither of the two proposed methods, which assume left-censored missing values, handle the MAR data even when the distributional assumption is satisfied. Mistakenly applying these left-censored techniques to data which violate this missing assumption can lead to poor estimates. The damage is most apparent in the standard deviation where over-estimates of 50% or more become common for even moderate amounts of missing data.

Conclusions from the MAR simulation is that when values are missing at random NONE appears to be the best option overall, though it will systematically underestimate the standard deviation when the number of observed values gets very small and the variance of this estimate becomes quite large when more than 50% of the data is missing. K nearest neighbors is also effective when the variables in the dataset are moderately correlated, when $\rho \geq 0.33$ or higher, but will still have issues underestimating the standard deviation once the proportion of missing values nears 50%. In these simulations, random forest did not perform any better than kNN and for the most part is very similar to imputation with the observed average except when the correlation is very high, $\rho = 0.9$.

2.1.1.2. Log-Normal

The first examination of non-normality involves the log-normal distribution. The main purpose of using a log-normal distribution is to examine the impact of violation in normality due to right-skew. Since metabolomic datasets often contain extreme values, right skewness is a reasonable possibility. Further, the results of Chapter 2 showed that untransformed ion counts are frequently right skewed. Performance of the missing value methods in log-normal data will elaborate on the potential consequences of not log transforming the data prior to parameter estimation. Datasets were constructed in the same fashion as the normal simulation scheme, but then exponentiated to produce a log-normal dataset:

Figure 3.27: Probability density function of variables used for lognormal simulation.



$$\boldsymbol{Y}_{LN}^{\rho} = exp\left(\boldsymbol{Z}_{n\times m} * Chol\left(\Sigma_{\rho}\right) + \boldsymbol{\mu}\right)$$

The untransformed values $\mu_j = 2.65$ and $\sigma_j = 0.3$ for $j \in \{1, \ldots, 114\}$. are used as this produces a variable with transformed mean:

$$\mu_{LN,\,i} = e^{\left(2.65+\frac{0.09}{2}\right)} = 14.8$$

and is similar to that of the normal simulations. The standard deviation was chosen as this produces a skewness of:

$$\gamma_{LN,i} = (e^{0.09} + 2)\sqrt{e^{0.09} - 1} = 0.94$$

This makes for a moderately right skewed variable as indicated by *Figure 3.27*. The transformed standard deviation however is quite different at a value of:

$$\sigma_{LN,\,i} = \sqrt{(e^{0.36} - 1)e^{2*2.65 + 0.36}} = 5.2.$$

However, it is necessary to accept a larger standard deviation than that used in the normal simulations in order to get the skewness near 1 and maintain a mean around 15.

Results of the log-normal simulations are presented in the same manner as previous simulation results. *Figures 3.28* and *3.29* show the estimated parameter against the uncensored sample values. Patterns for the mean parameter are rather like those of the normal simulation when missing values are left-censored

. Mean estimates from NONE and AVG almost appear random except that neither, naturally, ever underestimates from the true uncensored value. MIN somewhat tracks with the true value, predicting levels above .5 well but performance degrades the lower the true value is and is noticeably worse when the true level drops below .2. Recall that under the LOD mechanism, lower levels of the true parameter are associated with larger amounts of missing values, while higher levels are associated with lower amounts of missing values. RF is very similar to AVG (and NONE) when variable correlation is low, and the pattern moves closer to that of MIN as the correlation increases. Similarly, kNN estimates are also very similar to MIN and AVG under low correlation but, unlike RF, estimation improves as ρ increases. The improvement brought to kNN by higher correlation doesn't produce a patter the same as MIN though, as kNN begins to show unbiasedness around .2 or less.

ML and RREG are shown to underestimate and overestimate at all levels of the true parameter, though of course the accuracy is better at higher levels of the true parameter. While this general pattern is consistent with simulations of normally distributed, left-censored data, the range of the estimates for both parameters is much greater when log normal data is used. In the normal, left censored simulation the mean estimates were always above 0.75, whereas here the

Figure 3.28: Error in mean under Log-normal simulations. X-axis is uncensored sample mean. Y-axis is predicted mean. Columns represent pairwise correlation between variables in data set. Red line represents y=x.

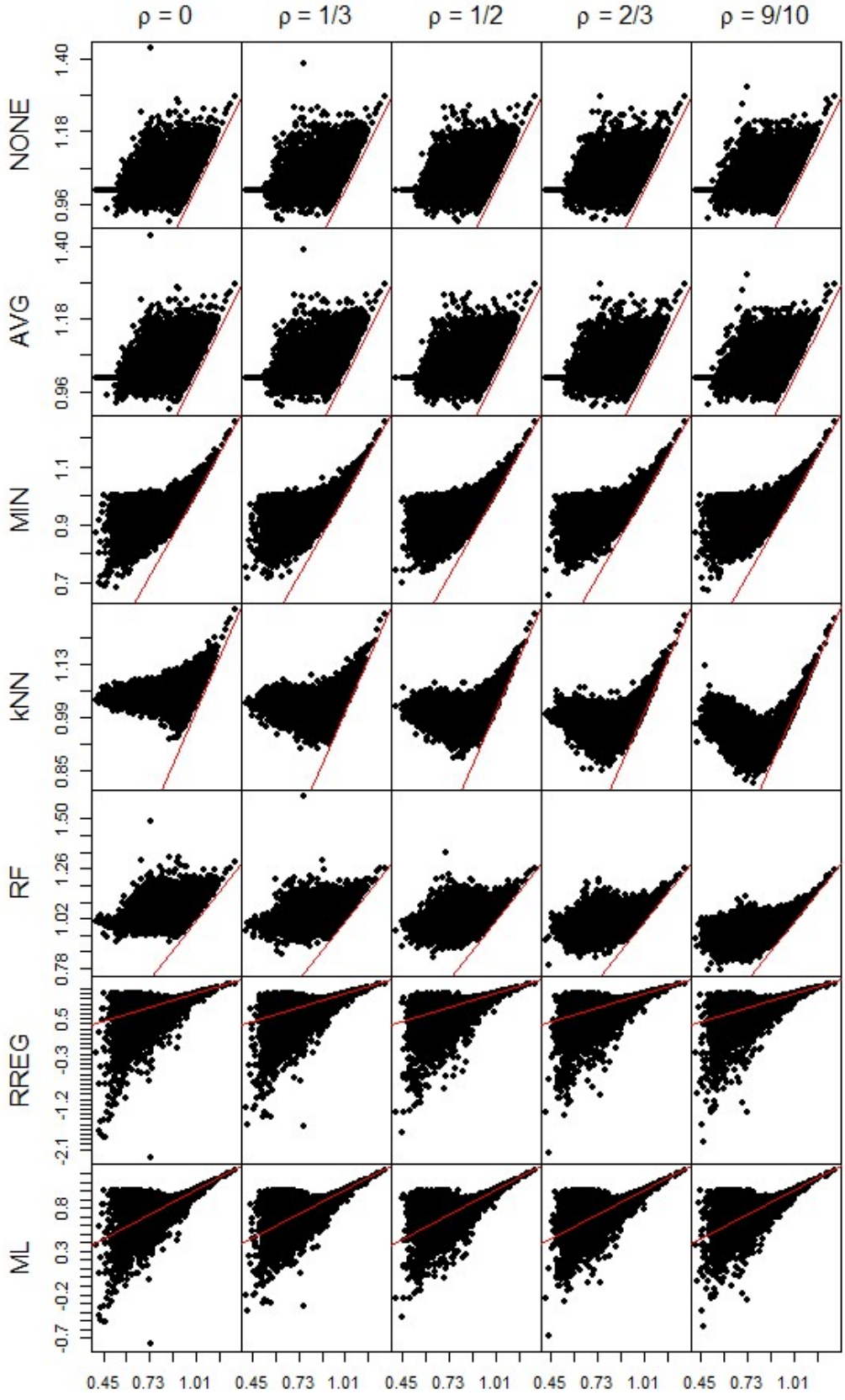
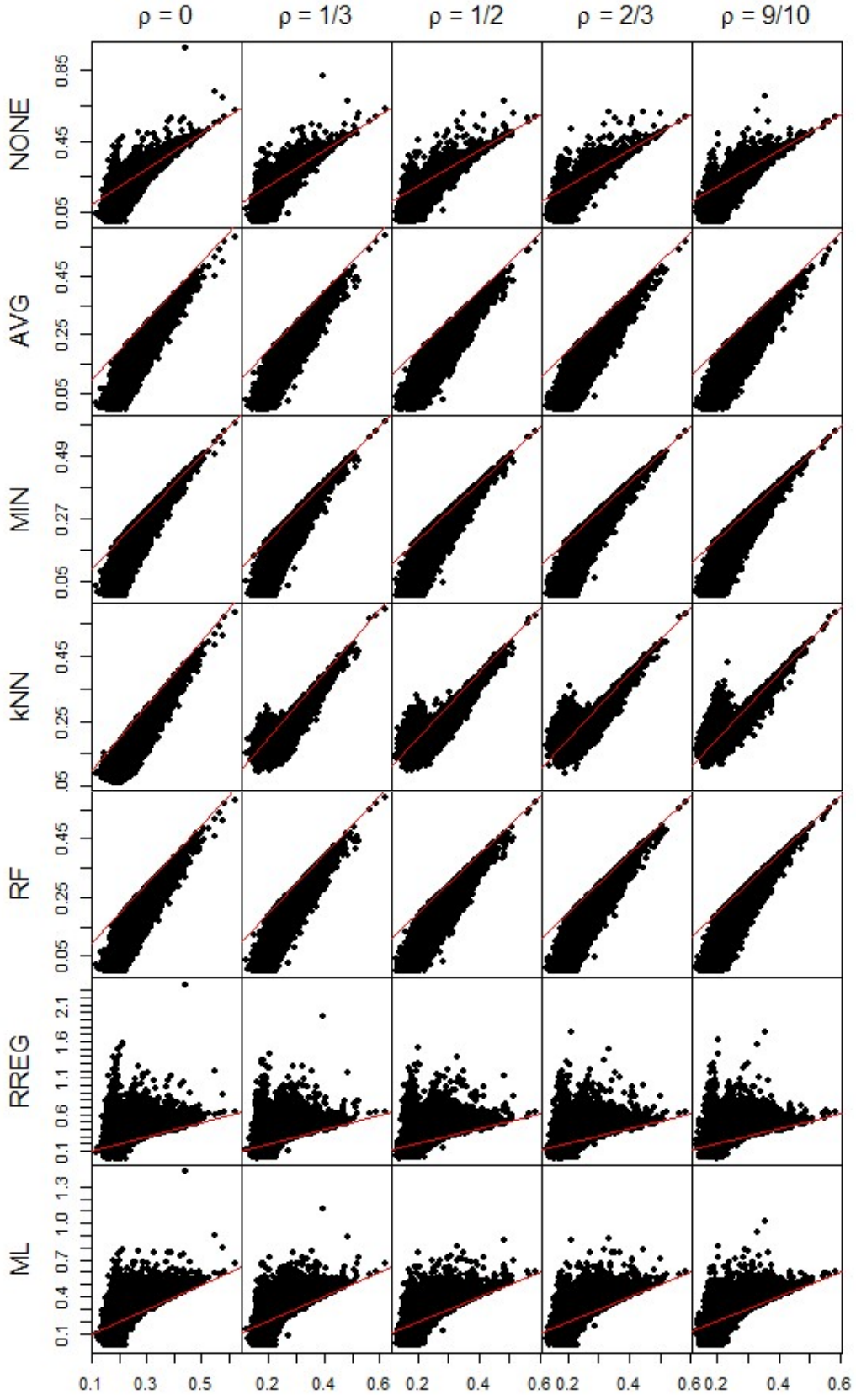


Figure 3.29: Error in SD under Log-normal simulations. X-axis is uncensored sample SD. Y-axis is predicted SD. Columns represent pairwise correlation between variables in data set. Red line represents y = x.



lowest estimates for ML are below 0 and for RREG are below -0.5. The wider range of estimates suggests lower accuracy for log-normal data.

Patterns for the standard deviation are quite different. NONE shows evidence of being unbiased but suffers more variability and the occurrence of large outliers compared to the other traditional methods. RREG and ML show similar patterns to each other though with a greater degree of variability as estimates regularly reach up to .7 while most others are confined to below .45. This suggests, as seen in the mean, a lower level of accuracy for the two parametric methods. Profiles for AVG, MIN, KNN and RF closely resemble each other indicating a similar level of accuracy that appears better than the other three methods. Once again, the value of ρ does not appear to be RF as profiles of standard deviation estimates by this method remain largely unchanged from $\rho = 0$ up to $\rho = 0.9$. The standard deviations calculated using kNN to fill in missing values do change as ρ increases, becoming less biased for lower values of the standard deviation.

Figures 3.30 and *3.31* show the relative error in the estimation of mean and standard deviation respectively as a proportion of missing values. Profiles in the mean parameter for NONE, AVG, MIN, KNN and RF are rather consistent, showing over estimates that increase as the proportion of missing values increases. MIN has less error at the higher levels, peaking at around 2.2 times the uncensored mean level while AVG and NONE peak around 2.5 times. Estimates from kNN show slight improvement as ρ increases, whereas RF does not improve in any meaningful way. ML and RREG produce under and over estimates of both parameters but are prone to extremely low mean estimates. The magnitude of these underestimates of the mean are, once again, most acute in RREG. There is evidence that underestimates of the mean are favored as the amount of missing values increases, and the magnitude of the relative error is

Figure 3.30: Relative error in mean under Log-normal simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y=0.

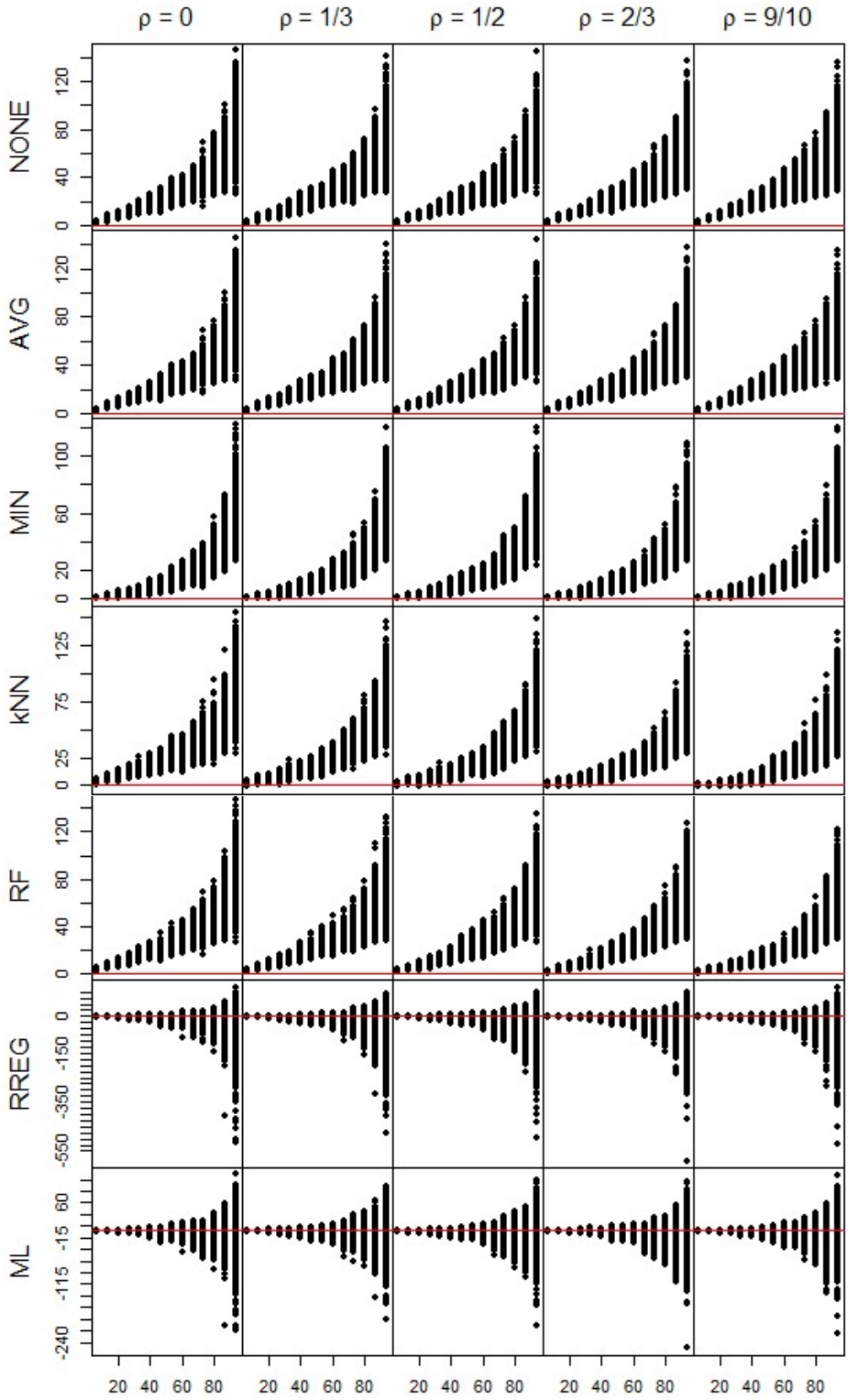
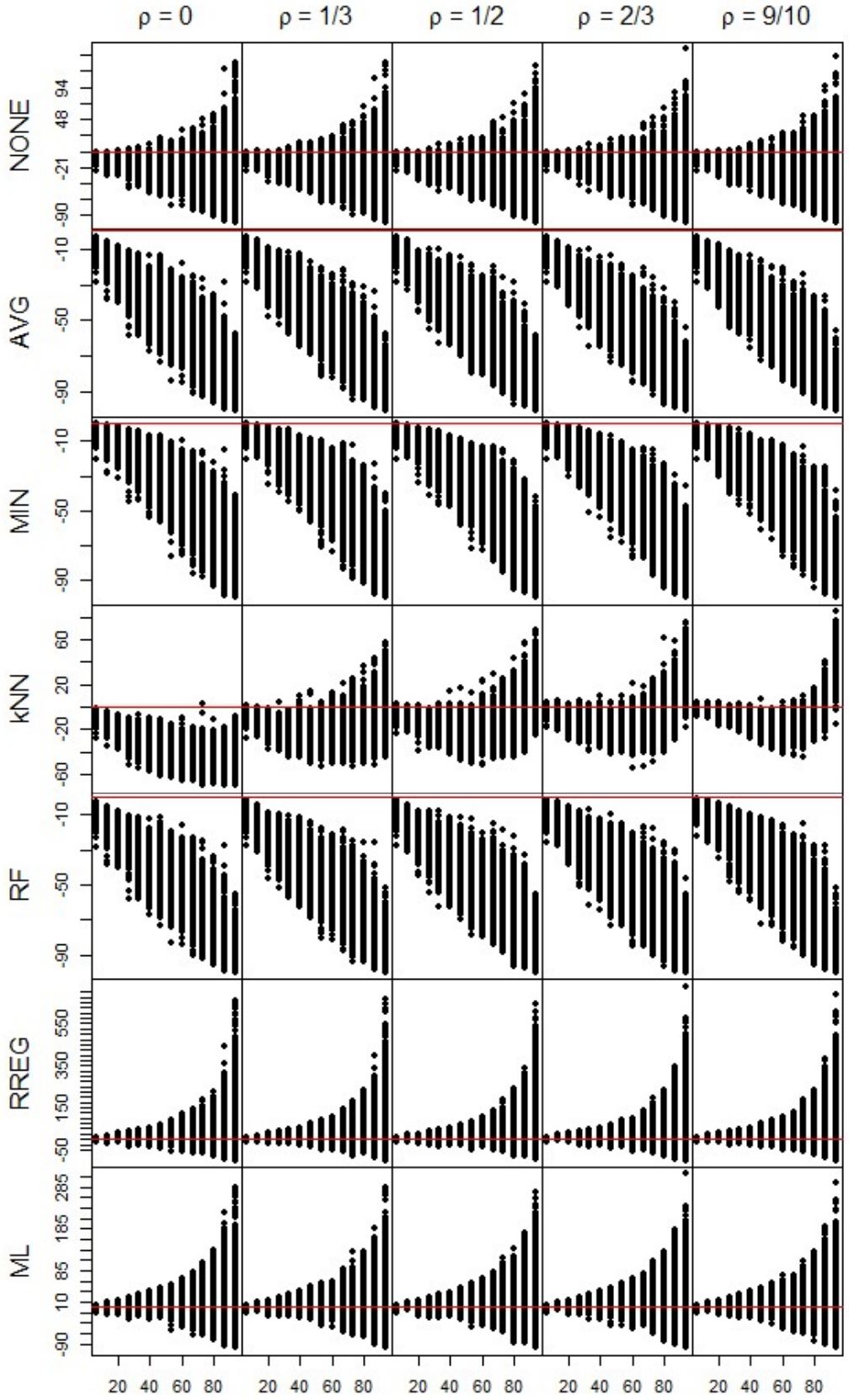


Figure 3.31: Relative error in SD under Log-normal simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y = 0.



extremely large when the proportion of missing values is above 80%.

Some remarkable differences between the five traditional methods are exhibited in the profiles of standard deviation estimates. NONE generally underestimates the standard deviation but extreme overestimates begin to appear around 40% missing and increase in frequency and magnitude as the proportion of missing values increases. AVG and MIN are rather similar, attributed to the right-skew pushing the observed average and minimum closer together. RF continues to track with both AVG and MIN with no meaningful improvement seen with increasing correlation. KNN has the most unusual pattern. When the correlation is low kNN underestimates, though not to the same degree as AVG / MIN / RF. As ρ increases, the bias in kNN standard deviation estimates decrease, showing relative error to shrink substantially from $\rho = 0$ *to* $\rho = 0.9$. However, as the proportion of missing values rises above 50%, kNN tends to overestimate the standard deviation. At $\rho = 0.9$, kNN appears roughly unbiased for low amounts of missing values, biased towards lower values at intermediate to moderate levels of missing values, and then biased toward overestimates for high amounts of missing values. RREG and ML standard deviation estimates remain similar those of the normal, left-censored simulation with the exception of being even more variable and producing even larger overestimates, particularly when the missing proportion is above 80%.

Next, the average relative error plots, given in *Figures 3.32* and *3.33*, show that on average the five common methods (NONE through RF) overestimate the mean parameter with a magnitude that is greater magnitude than the two proposed methods. NONE and AVG have about 5-10% more bias than MIN. In a familiar theme KNN and RF fluctuate, being similar to NONE and AVG when the correlation between variables is zero and moving closer to MIN as correlation increases. ML and RREG are biased towards lower estimates of the mean, though as

Figure 3.32: Bias in mean parameter under Log-Normal simulations. X-axis represents percentage of missing values. Y-axis is average relative error.

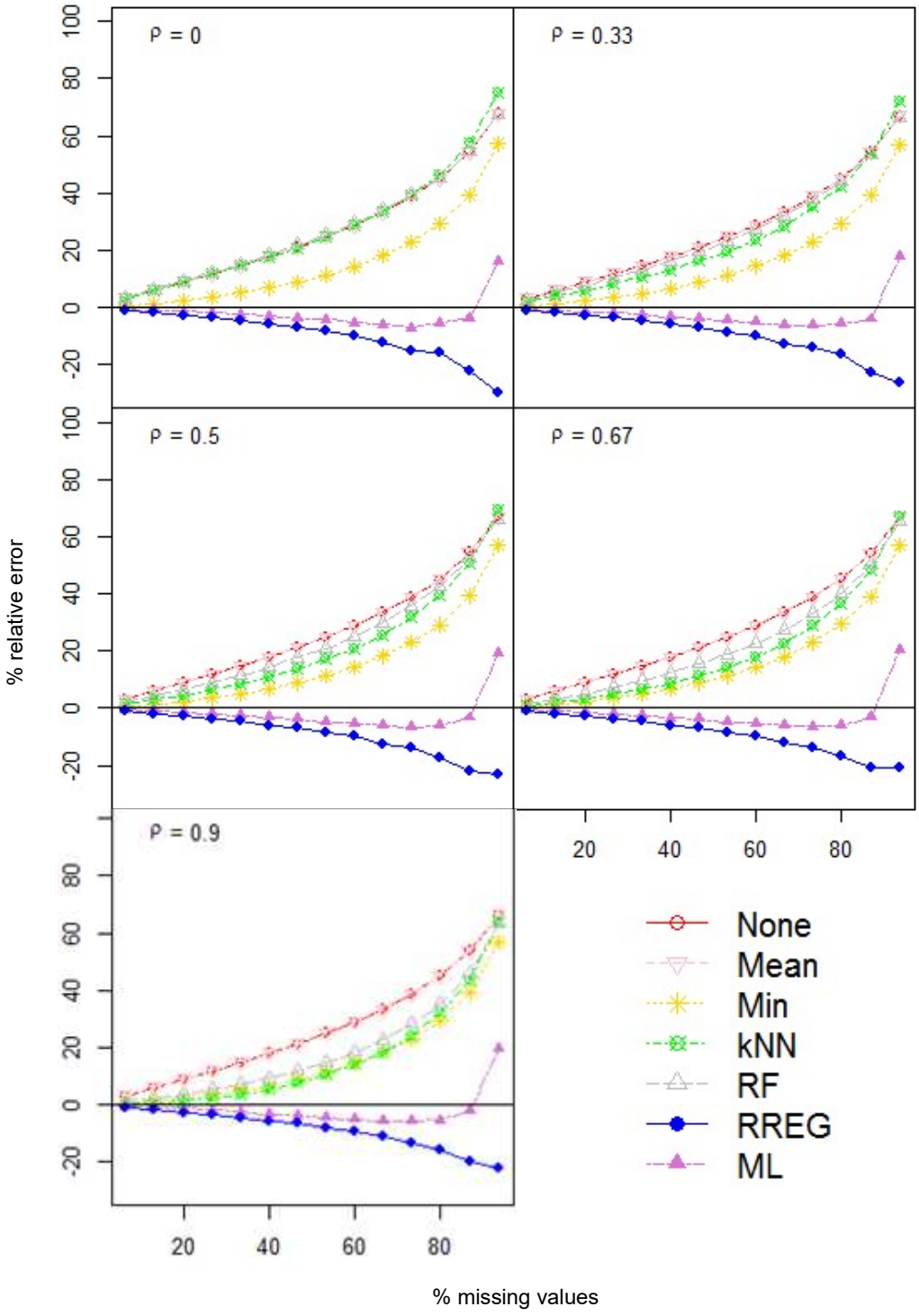
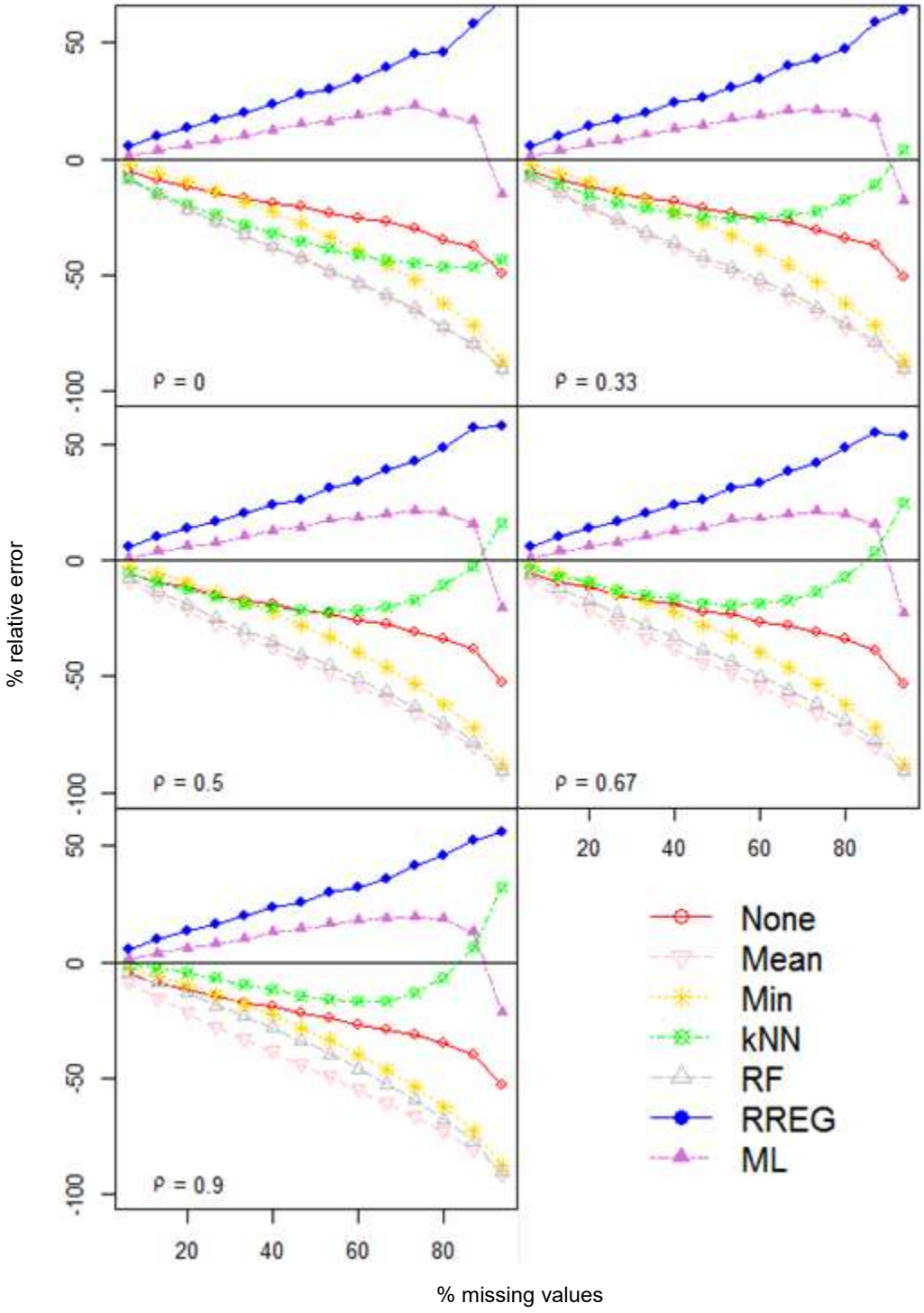


Figure 3.33: Bias in SD parameter under Log-normal simulations. X-axis represents percentage of missing values. Y-axis is average relative error.



a percentage the average error is generally contained to around 20% even as the percentage of missing values exceeds 90%. ML fairs better than RREG at high levels of missing values. The other five methods reach 20% error when the amount of missing values is between 50% to 70% and increase rapidly thereafter.

For the standard deviation, average relative error is similar in magnitude for all seven methods. RREG and ML are biased upwards while the five common methods are biased downward. NONE shows a little less bias compared to AVG and MIN. RF again matches closely to AVG and MIN, as has been typically observed in these simulations. Performance of KNN is similar to AVG and NONE at low levels of missing values but deviates as the missing proportion increases from 50% to 100%. When correlation is low this deviation is minor but as the correlation increases KNN deviates more strongly towards zero as the proportion of missing values increase. In fact, when $\rho > 0$, KNN fully reverses to a bias of overestimates between 80% and 100% missing values. Similar to the mean parameter, ML outperforms RREG in the standard deviation as well.

Finally, Tables 3.10 and 3.11 show the average relative error and variance for the seven methods when $\rho = 0.33$. These show that while RREG and ML are less biased estimators of both sample parameters than the other five methods, but that once again the variability in these methods becomes extremely high as the proportion of missing values increases. Even for 40% missing values these two methods have variance that is 3-4 times higher in the mean and 2-4 times higher in the standard deviation. The simulation results for log-normal data reveal that none of the seven methods to be unbiased in either parameter when variables are left-skewed. As with the previous normal simulations, relative errors in the mean parameter tend to be less than

Table 3.10: Relative error averages and variances in Mean parameter from log-normal simulations. Missing values are left-censored and ρ=1/3.

Proportion Missing	NONE		AVG		MIN		KNN		RF		RREG		ML	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
6.7%	3.16	0.18	3.16	0.18	0.54	0.10	1.97	0.59	2.50	0.24	-1.07	0.39	-0.47	0.16
13.3%	6.07	0.52	6.07	0.52	1.31	0.29	3.94	1.28	5.02	0.60	-1.95	1.49	-0.97	0.58
20.0%	8.89	1.20	8.89	1.20	2.29	0.61	5.84	2.39	7.58	1.20	-2.88	3.50	-1.52	1.28
26.7%	11.81	2.05	11.81	2.05	3.59	1.13	8.10	3.62	10.28	2.03	-3.66	6.82	-1.97	2.87
33.3%	14.71	3.62	14.71	3.62	4.93	1.71	10.45	5.78	13.04	3.28	-4.51	12.2	-2.63	5.01
40.0%	17.82	4.99	17.82	4.99	6.68	2.60	13.01	7.47	16.12	5.08	-5.82	23.4	-3.38	9.04
46.7%	21.15	7.98	21.15	7.98	8.76	3.98	16.17	10.86	19.49	7.97	-6.55	38.5	-3.94	16.16
53.3%	24.85	12.03	24.85	12.03	11.29	6.04	19.60	15.52	23.09	11.75	-8.09	64.6	-4.77	27.3
60.0%	28.76	16.88	28.76	16.88	14.39	9.11	23.48	19.90	27.25	17.09	-9.11	98.2	-5.22	42.2
66.7%	33.38	25.9	33.38	25.9	18.00	12.71	28.37	28.8	31.93	24.7	-11.66	210.5	-6.52	92.9
73.3%	38.95	39.2	38.95	39.2	23.12	22.61	34.86	43.8	37.74	39.7	-12.28	313.7	-6.32	146.9
80.0%	45.03	66.4	45.03	66.4	29.30	33.8	42.0	68.0	44.00	64.1	-14.10	615.1	-5.91	271.0
86.7%	54.6	122.7	54.6	122.7	39.12	67.2	53.6	122.3	53.7	121.5	-18.61	1525	-4.22	655.0
93.3%	67.0	283.1	67.0	283.1	56.8	194.1	71.8	293.6	66.6	274.3	-18.0	5413	17.57	1643

Table 3.11: Relative error averages and variances in SD parameter from log-normal simulations. Missing values are left-censored and ρ=1/3.

Proportion Missing	NONE		AVG		MIN		KNN		RF		RREG		ML	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
6.7%	-5.5	8.2	-8.8	7.7	-2.8	4.4	-6.6	8.4	-8.2	7.7	4.0	9.3	1.1	4.6
13.3%	-9.2	20.4	-15.7	17.6	-6.1	10.3	-11.5	14.9	-14.7	18.0	8.1	23.0	3.6	12.1
20.0%	-11.9	35.4	-21.5	28.1	-9.5	17.4	-15.2	25.8	-20.3	28.6	11.9	46.3	6.1	23.0
26.7%	-14.8	56.8	-27.5	41.1	-14.0	31.8	-18.9	34.9	-26.2	43.2	14.6	85.6	7.7	47.7
33.3%	-17.3	75.9	-33.0	49.7	-18.0	39.9	-21.5	41.8	-31.5	52.6	17.7	135.0	10.4	68.8
40.0%	-18.7	115.9	-37.8	67.9	-22.4	55.0	-23.2	52.0	-36.2	71.9	21.4	225.3	12.8	107.7
46.7%	-21.6	153.2	-43.6	79.2	-27.8	72.9	-25.0	61.2	-41.8	85.1	23.2	340.4	14.5	171.3
53.3%	-23.2	205.2	-48.6	92.0	-33.1	94.0	-25.8	64.4	-46.7	100.0	27.1	523.2	16.9	267.9
60.0%	-25.7	241.7	-54.2	91.7	-39.3	100.0	-25.5	85.7	-52.3	100.4	29.7	676.7	18.4	355.3
66.7%	-27.1	391.5	-59.4	121.5	-45.4	146.3	-24.2	95.9	-57.4	133.0	34.4	1255	20.7	665.8
73.3%	-30.8	450.8	-66.0	108.8	-53.3	146.0	-22.5	129.1	-64.1	122.6	36.1	1702	20.9	921.9
80.0%	-34.6	634.0	-72.8	109.3	-62.0	158.7	-18.0	147.0	-71.0	123.9	38.7	2859	19.6	1492
86.7%	-37.5	991.5	-79.9	102.6	-71.6	167.8	-11.3	176.0	-78.4	118.6	46.8	5568	17.2	2746
93.3%	-50.6	1873	-90.8	64.6	-87.3	124.9	3.7	253.6	-90.3	73.6	44.5	16044	-17.8	5189

those in the standard deviation, but the overall magnitudes are much more similar. Both of the proposed methods are less biased on average than the five common methods; however, once again the variability in these methods is much higher, leading to a degree of unreliability.

2.1.1.3.Uniform Simulations

The second attempt to examine non-normality involves the Uniform(14,16). As with the simulation of log-normal distributions, missing values are removed in a left-censored fashion. Results are shown in *Figures 3.34-39*. Error plots show the five common methods to have a similar pattern in both parameters to those of the normal simulation when missing values are also left-censored. One minor difference is that here NONE never overestimates the standard deviation. Both proposed methods show notable differences compared the simulation results using the normal distribution with left-censored missingness. Both proposed methods almost exclusively overestimate the mean parameter and neither produces the severely overestimated standard deviations seen in the normal, left-censored simulations.

As with data simulated using other distributions, the range of both parameter estimates from RREG and ML compared to the others suggests that these methods are more variable. In the relative error plots, the mean is consistently over-estimated by the five common methods and at a similar percentage, which ranges from 2% at low to intermediate proportions of missing values up to 8% at high levels of missing values. While RREG and ML both produce estimates with similar errors, these methods also produce estimates that are close to the original, uncensored sample mean. One unusual observation, compared to the previous simulations, is the variation in RREG estimates are comparable that of ML. All previous simulations experiments have shown the opposite. In the standard deviation, NONE, AVG and MIN show similar patterns of underestimation in which the bias consistently increases as the percentage of missing values

Figure 3.34: Error in mean under Uniform simulations. X-axis is uncensored sample mean. Y-axis is predicted mean. Columns represent pairwise correlation between variables in data set. Red line represents y=x.

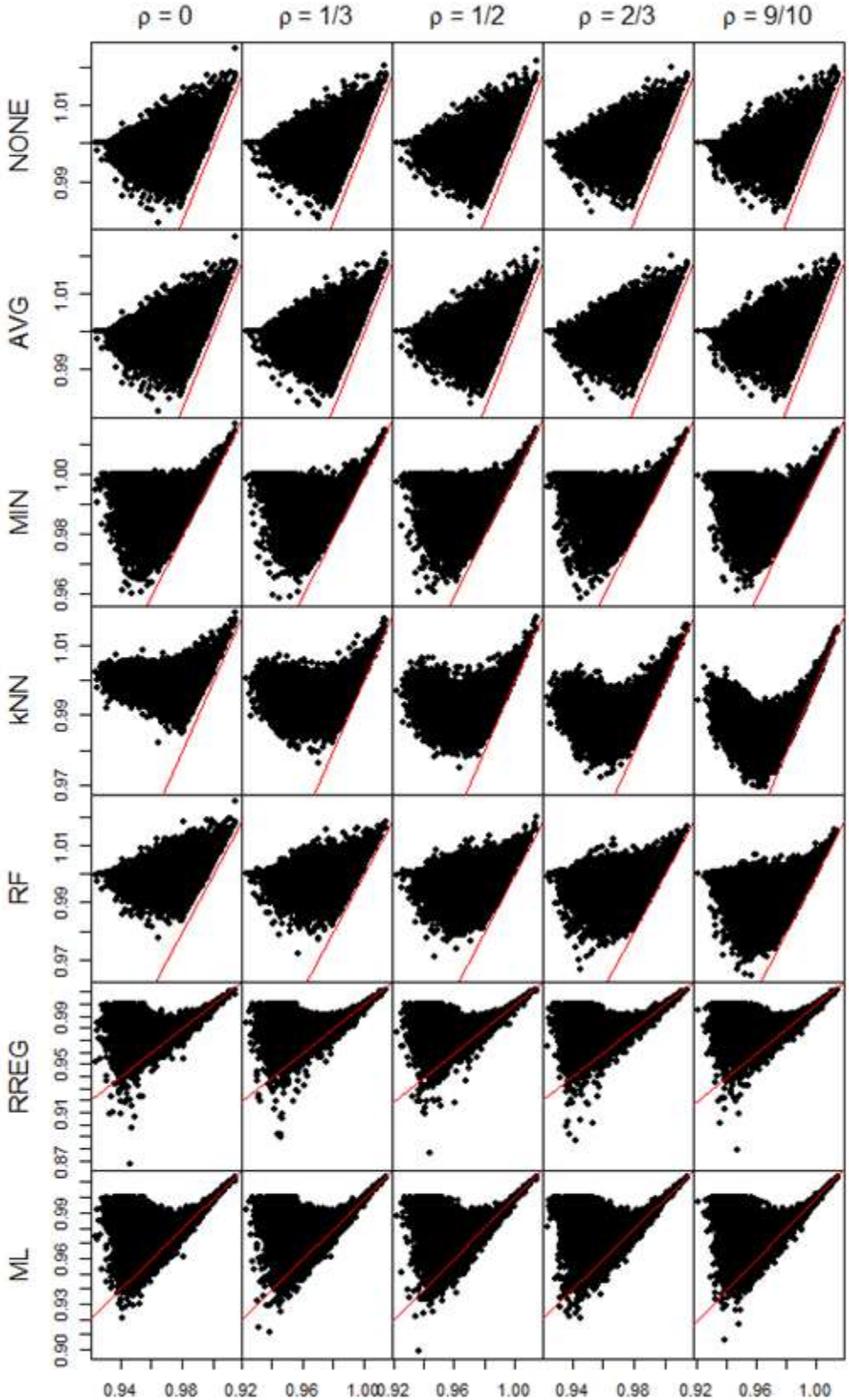


Figure 3.35: Error in SD under Normal and MAR simulations. X-axis is uncensored sample SD. Y-axis is predicted SD. Columns represent pairwise correlation between variables in data set. Red line represents y=x.

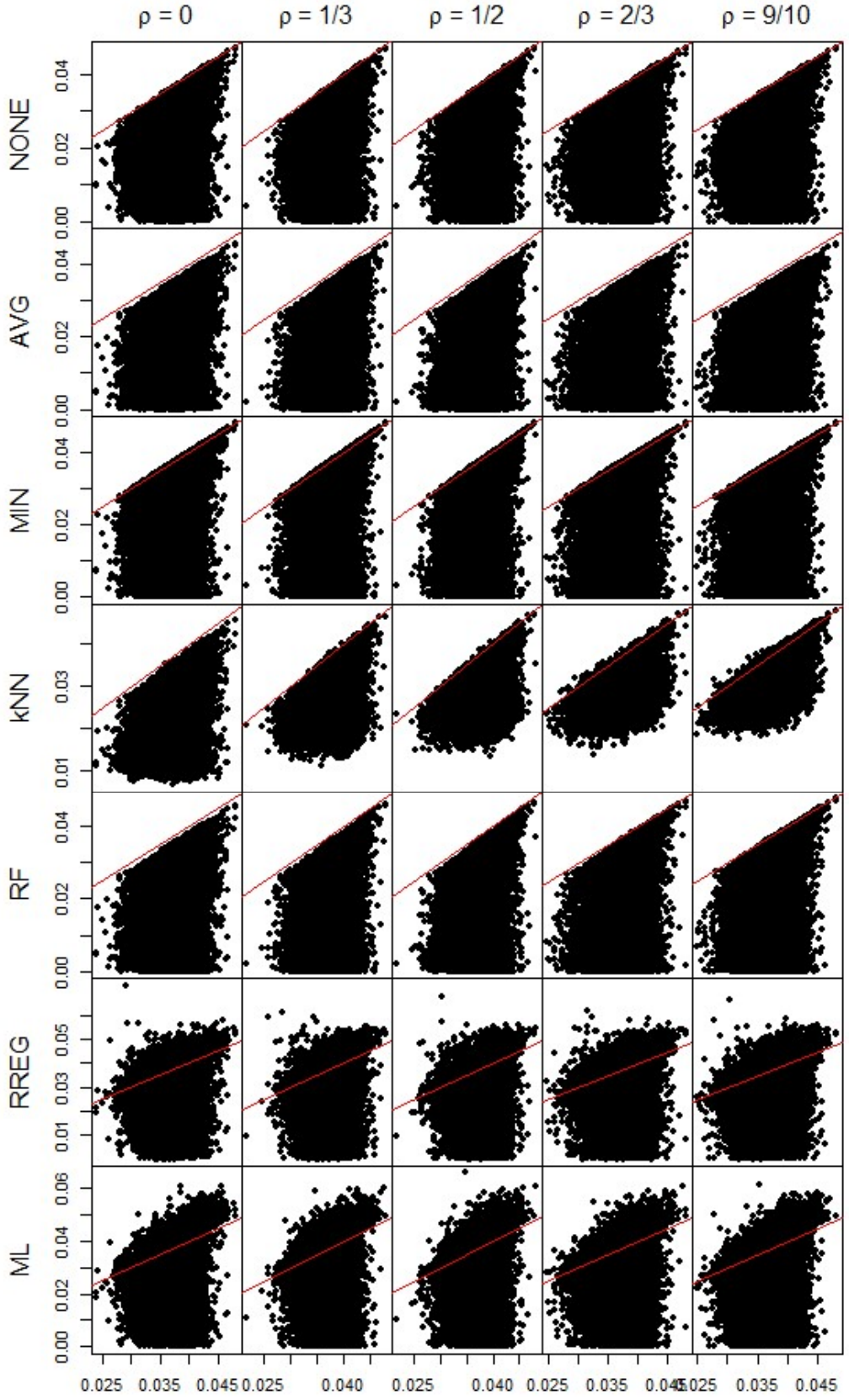


Figure 3.36: Relative error in mean under Uniform simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y=0.

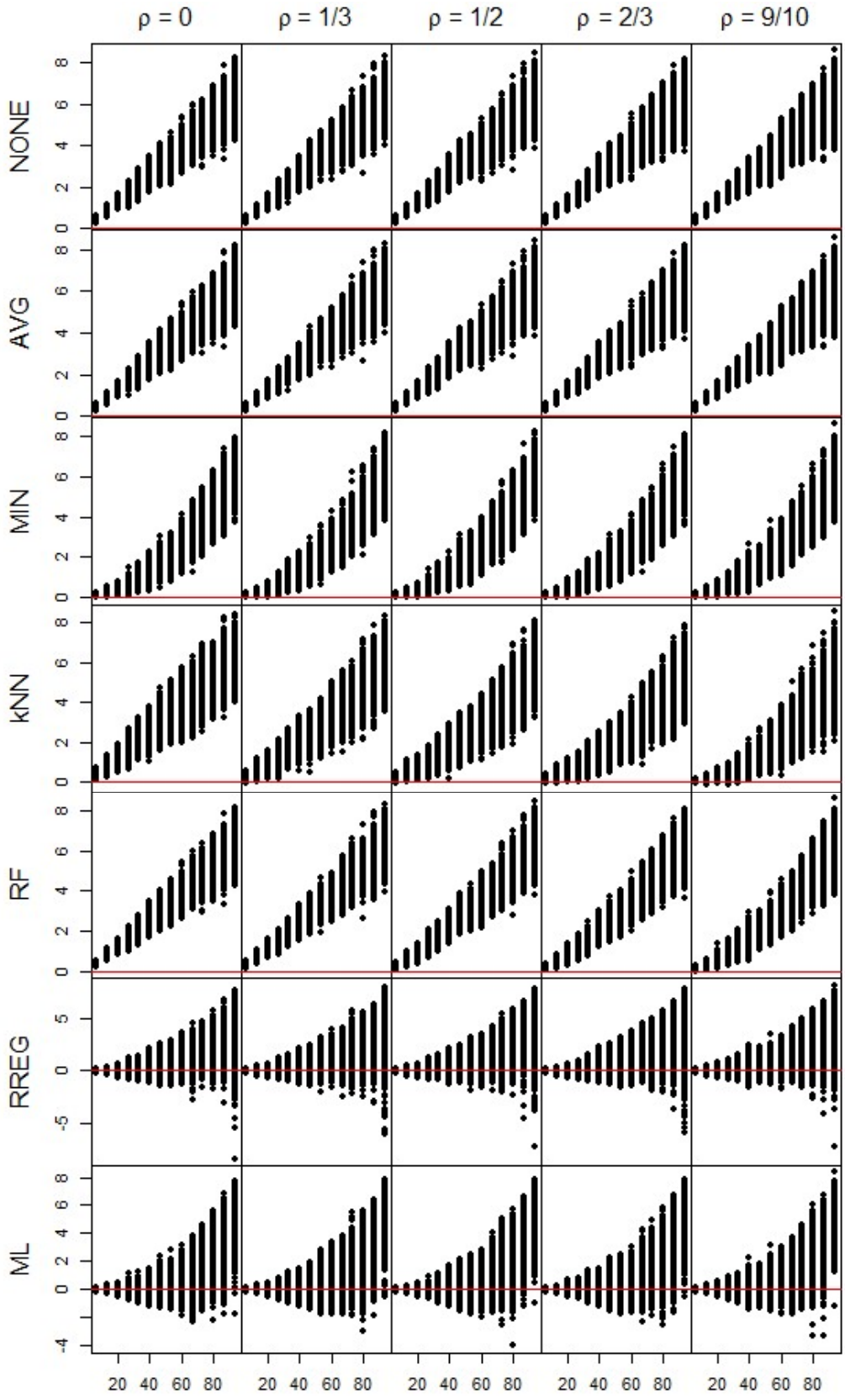


Figure 3.37: Relative error in SD under Uniform simulations. X axis in the proportion of missing values. Y axis represents percent change from uncensored sample value. Red line represents y=0.

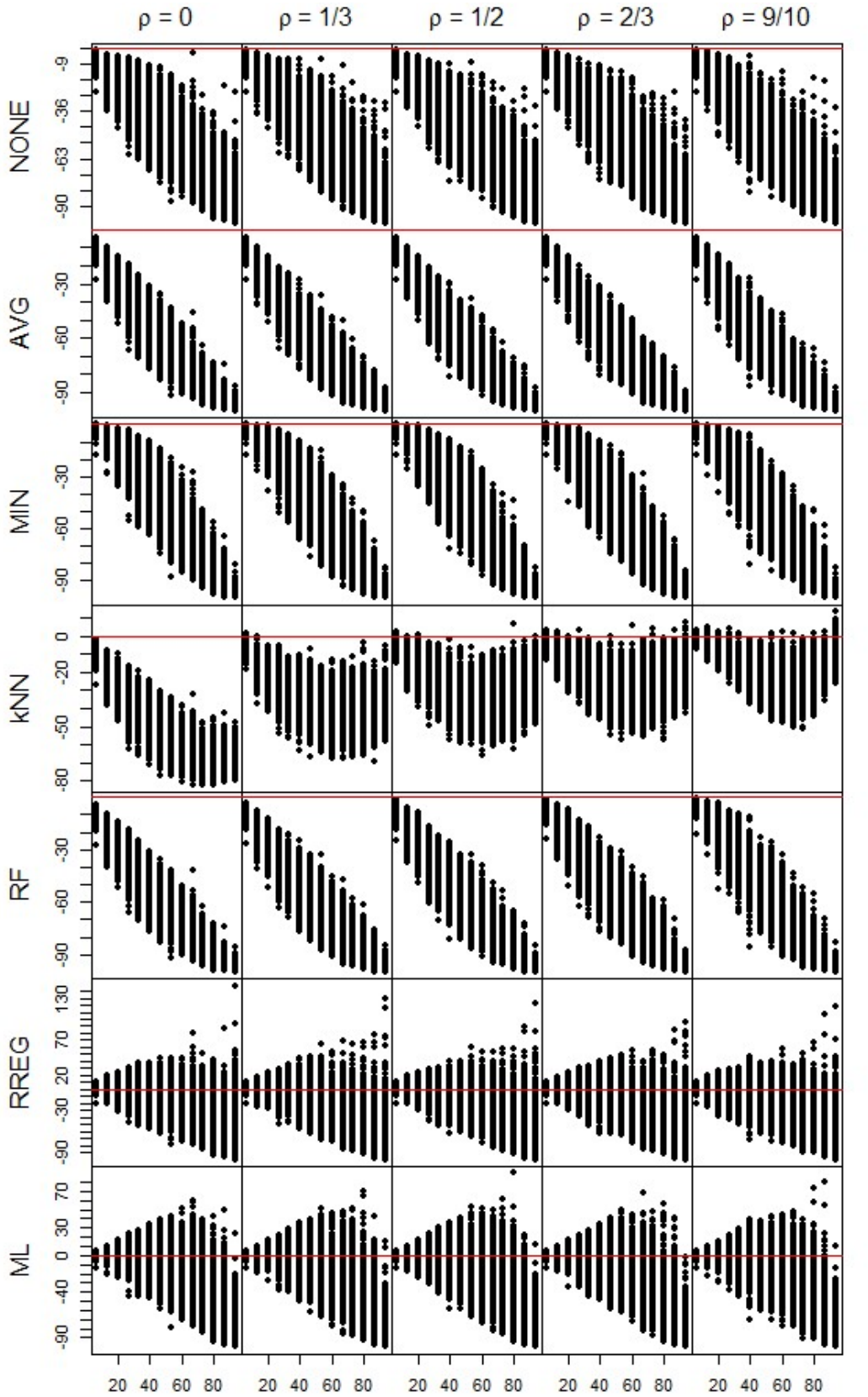


Figure 3.38: Bias in mean parameter under Uniform simulations. X-axis represents percentage of missing values. Y-axis is average relative error.

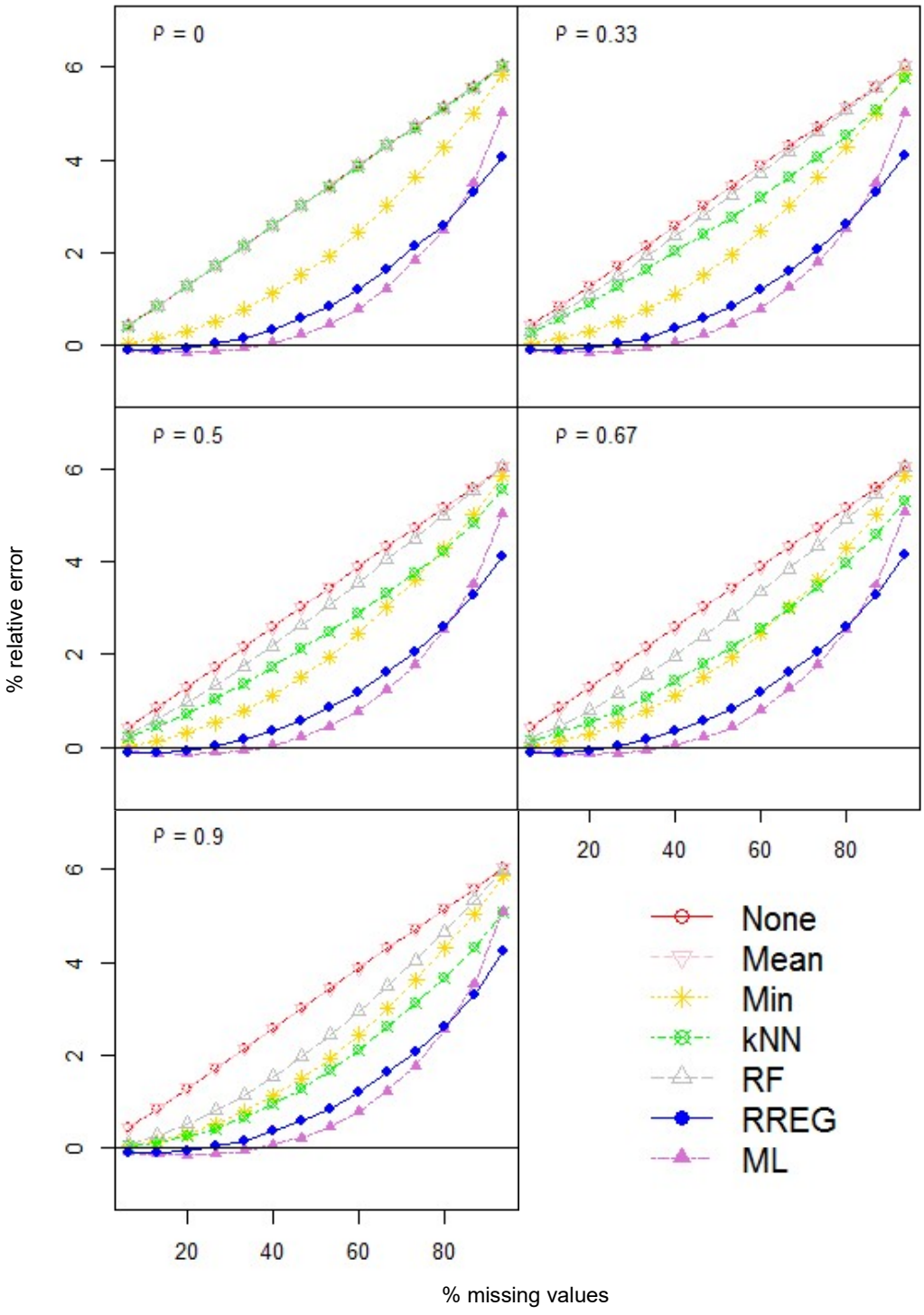
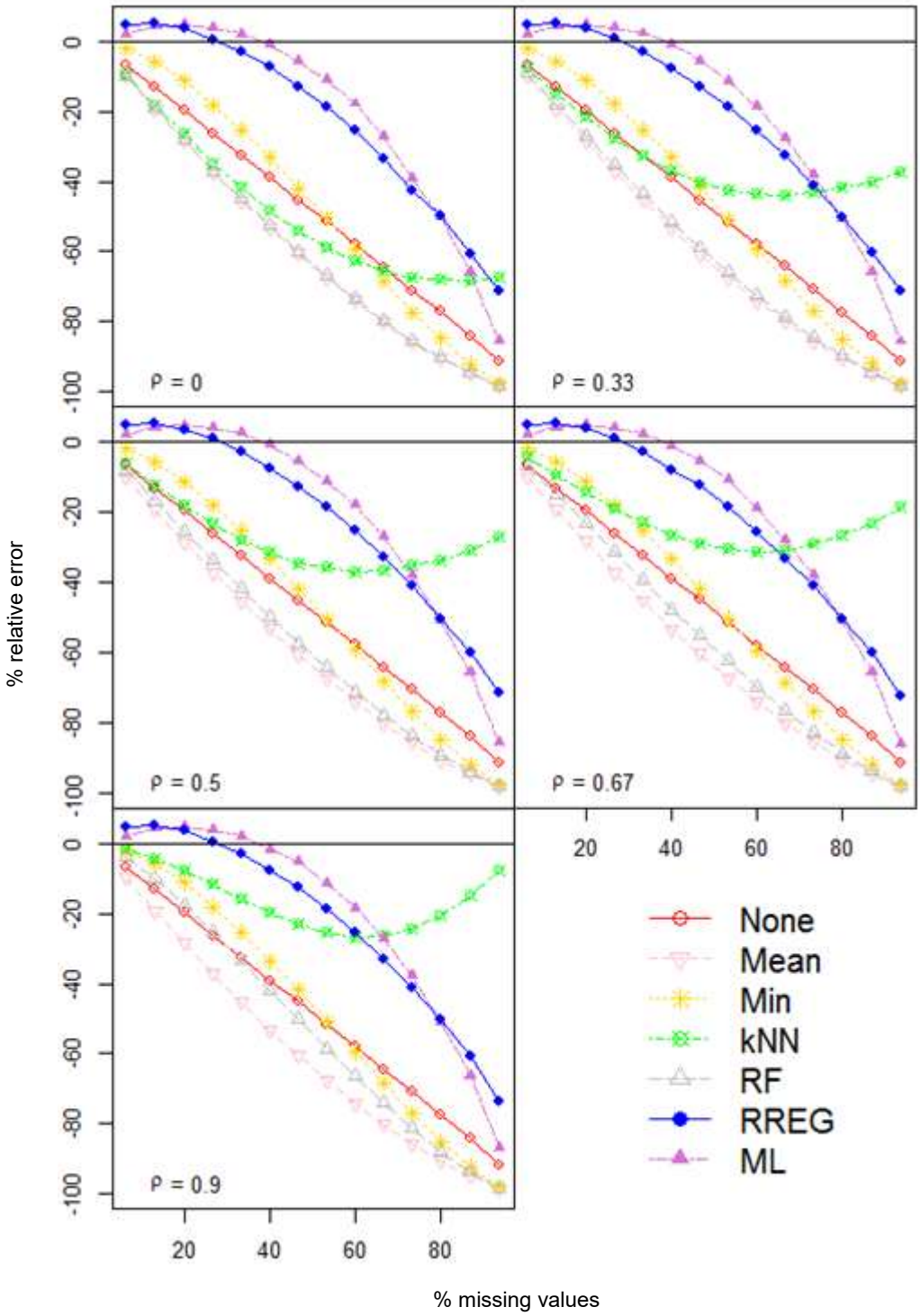


Figure 3.39: Bias in SD parameter under Log-normal simulations. X-axis represents percentage of missing values. Y-axis is average relative error.



increases. RF is virtually indistinguishable from NONE, AVG and MIN even as the correlation between variables in the dataset increases. KNN on the other hand once again improves substantially in terms of bias as this correlation increases. RREG and ML continue to over and under estimates in both parameters, but, like the other five methods, the greatest magnitudes of the error are overestimates of the mean and in underestimates of the standard deviation. The average relative error plots are interesting as roughly all seven methods are shown to be quite similar in both parameters. This is particularly true with the mean parameter as all methods steadily increase from roughly unbiased when the proportion of missing values is close to 0 to overestimates of about 6% as the proportion of missing values approaches 1. MIN tends to be a little bit better than AVG / NONE while RF and KNN are closer to AVG/NONE when the correlation is low and migrate towards MIN as the correlation improves. RREG and ML do a little bit better than the other methods, most notably when the proportion of missing values is between 50-70%, but the performance of all methods in the mean parameter is fairly close throughout.

Average profiles in the standard deviation are also similar across the methods. All seven methods are biased towards underestimation and, with the exception of KNN, the profiles move from near unbiased to standard deviations that are basically zero as the proportion of missing values approaches 1. Variance estimates approaching zero is expected for the single imputation methods, but it is notable that the phenomenon also occurs in both parametric methods. AVG has the worst bias followed closely by RF, which is generally different by only a few percentage even when $\rho = .9$. NONE and MIN demonstrate very similar performance and do 10-20 percentage points better than AVG and RF when the missing proportion is between 20% and 80%. RREG and ML have less bias than the other five methods up to missing proportions of 60

to 70%. Between 60-70% missing values, kNN becomes the least biased and bias actually decreases for this method as the missing proportion increases further.

The final item for the uniform simulations are Tables 3.12 and 3.13 giving the average relative error and associated variance of the seven methods in both parameters when the between correlation is 0.33. Estimates of the mean with RREG and ML are about 1-2% better than the other methods on average while the variance of their estimates is anywhere from two to ten times higher than other methods. Both the bias and variance increase steadily in all methods as the proportion of missing values increases. Table 3.13 indicates that RREG and ML have anywhere from a one-quarter to one-third the bias that NONE, AVG, MIN and RF do for the most part.

Table 3.12: Relative error averages and variances in Mean parameter from uniform simulations. Missing values are left-censored and $\rho=1/3$.

Proportion Missing	NONE		AVG		MIN		KNN		RF		RREG		ML	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
6.7%	0.43	0.00	0.04	0.00	0.04	0.00	0.27	0.01	0.33	0.00	-0.10	0.01	-0.08	0.00
13.3%	0.86	0.01	0.14	0.01	0.14	0.01	0.59	0.03	0.71	0.01	-0.11	0.02	-0.13	0.01
20.0%	1.29	0.02	0.31	0.01	0.31	0.01	0.91	0.05	1.11	0.02	-0.05	0.05	-0.14	0.03
26.7%	1.72	0.03	0.52	0.03	0.52	0.03	1.26	0.08	1.51	0.04	0.05	0.09	-0.12	0.06
33.3%	2.14	0.06	0.79	0.06	0.79	0.06	1.62	0.12	1.92	0.05	0.19	0.2	-0.06	0.12
40.0%	2.59	0.08	1.11	0.09	1.11	0.09	2.00	0.17	2.37	0.08	0.39	0.2	0.05	0.21
46.7%	3.03	0.10	1.51	0.14	1.51	0.14	2.38	0.21	2.81	0.10	0.64	0.4	0.23	0.35
53.3%	3.44	0.13	1.95	0.17	1.95	0.17	2.75	0.26	3.25	0.12	0.91	0.5	0.47	0.5
60.0%	3.89	0.19	2.46	0.21	2.46	0.21	3.17	0.33	3.71	0.17	1.28	0.6	0.80	0.6
66.7%	4.31	0.2	3.02	0.3	3.02	0.30	3.59	0.4	4.17	0.2	1.67	0.8	1.24	0.9
73.3%	4.71	0.3	3.61	0.3	3.61	0.35	4.02	0.5	4.61	0.3	2.16	1.1	1.79	1.2
80.0%	5.13	0.3	4.27	0.4	4.27	0.4	4.5	0.6	5.07	0.3	2.72	1.4	2.52	1.5
86.7%	5.6	0.4	5.0	0.5	5.01	0.5	5.1	0.6	5.5	0.4	3.43	2	3.50	1.6
93.3%	6.0	0.5	5.8	0.5	5.8	0.5	5.8	0.7	6.0	0.5	4.3	3	5.02	1

Table 3.13: Relative error averages and variances in SD parameter from uniform simulations. Missing values are left-censored and ρ=1/3.

Proportion Missing	NONE		AVG		MIN		KNN		RF		RREG		ML	
	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var	Avg	Var
6.7%	-6.6	7.1	-1.9	2.5	-1.9	2.5	-7.4	7.8	-9.1	6.8	2.2	2.7	3.0	11.9
13.3%	-13.1	24.0	-5.8	13.1	-5.8	13.1	-14.6	21.2	-18.0	21.2	4.3	15.9	3.4	40.2
20.0%	-19.7	41.1	-11.3	28.2	-11.3	28.2	-21.4	33.0	-27.0	33.9	4.9	38.9	1.5	75.9
26.7%	-26.2	66.3	-17.9	54.1	-17.9	54.1	-27.3	46.1	-35.5	52.3	4.0	86.7	-1.6	131.2
33.3%	-32.4	90.3	-25.1	80.2	-25.1	80.2	-32.2	56.1	-43.5	63.8	2.3	149.8	-5.2	193.5
40.0%	-38.9	106.1	-33.1	99.7	-33.1	99.7	-36.4	59.4	-51.5	67.9	-0.7	220.9	-10.3	249.1
46.7%	-45.3	128.9	-41.8	122.0	-41.8	122.0	-39.6	66.6	-58.9	75.2	-5.5	323.8	-15.8	324.5
53.3%	-51.5	132.2	-50.6	117.5	-50.6	117.5	-41.7	62.8	-65.9	66.9	-11.3	381.8	-21.5	367.5
60.0%	-57.9	126.2	-59.5	104.4	-59.5	104.4	-42.9	69.3	-72.5	56.2	-18.5	430.7	-28.3	386.3
66.7%	-64.2	130.3	-68.5	97.1	-68.5	97.1	-43.3	68.8	-78.8	47.4	-27.5	523	-35.6	432.0
73.3%	-70.8	124.7	-77.0	75.7	-77.0	75.7	-42.5	67.6	-84.5	36.8	-38.2	561	-44.2	470.0
80.0%	-77.3	111.6	-85.0	49.9	-85.0	49.9	-41.2	72.4	-89.8	24.0	-50.8	553	-53.3	480
86.7%	-84.0	95.7	-92.0	23.6	-92.0	23.6	-39.5	66.7	-94.4	12.1	-65.7	441	-63.6	504
93.3%	-91.4	75	-97.8	5.0	-97.8	5.0	-36.9	69.4	-98.3	3.0	-85.7	207	-74.8	640

However, bias for these methods still reaches about 20% when around half of the values are missing. ML also degrades more quickly than RREG as the missing proportion increases. kNN becomes biased very quickly, underestimating the standard deviation parameter by more than 20% with only 20% of the values are missing. But, the bias in the standard deviation parameter levels off quickly at just under -40 and becomes less biases above 70% missing values while RREG and ML continued to degrade.

2.1.2. Conclusions

These simulations show the five common methods consistently underestimate the standard deviation and with magnitude that, relative to the original parameter, is much greater than the

error in the mean parameter. This result is not surprising for the single imputation methods AVG and MIN, which underestimate in every condition considered here, since the replacement of missing values with a constant value naturally decreases variability in the imputed feature. No imputation of the data at all is arguably preferable to these methods for this reason, producing less biased estimates of standard deviation and having a similar amount of relative bias in the mean.

In every simulation experiment, RF estimates in both parameters proved to be little better than average imputation except when inter-variable correlation was the very high ($\rho = 0.9$) and missing values were left-censored, in which case average relative errors in this method were closer those of minimuvm imputation. The strong resemblance to average imputation implies that the algorithm employed by random forest imputation has little effect except when the correlation between the missing variable and the neighbors is very high. In the left-censored case, close correspondence to MIN further signals that RF correctly associates missing values with low values; however, the imputed values estimated by RF remain at a near constant value. The performance of RF in these simulations is surprising given the success of the method in other studies [22, 65]. The main difference between these simulations and those studies is that here the dataset consists of a single, homogenous population. Datasets examined by others involved multiple groups which were found to be statistically significant by hypothesis testing or separate by the first two components in a PCA analysis. Such strong group differences are likely to enhance the correlation of the data set as within group samples are more similar than between group samples. The implication of this is that random forest is effective at picking out averages between different populations but does not do well at discriminating between individual values within a single population.

The best performer of the five common methods is kNN. Similar to random forest, kNN always overestimates the mean, being close to average imputation when $\rho = 0$ and close to minimum imputation when $\rho = 0.9$. However, kNN preserved the standard deviation much better even under "weak" inter-variable correlation of $\rho = 0.33$. This performance in weakly correlated data suggests kNN does better than RF at producing an imputed variable whose distribution is closer to that of the censored variables.

Both proposed methods do well when values are left-censored. When the distribution is normal, maximum likelihood and rankit regression are nearly unbiased for both the mean and standard deviations up to a missing proportion of 80%. In the leptokurtic setting, represented here by the uniform distribution, the methods exhibited bias in both parameters, but the magnitude of this bias was less than the other methods. Finally, under a log-normal distribution, which here was both left skewed and platykurtic, both proposed methods had less bias, in terms of magnitude, in the mean while similar bias in the standard deviation.

It is notable that in the log-normal case maximum likelihood overestimates the standard deviation while the other methods underestimated this parameter. The one condition in which ML and RREG were shown to struggle was when values were removed randomly rather than from the lower tail. Overestimates of the standard deviation by ML and RREG were more severe than the underestimates seen in the other methods. This suggests that ML and RREG are more sensitive to LOD assumption than to the normality assumption. Another notable observation for both proposed methods is that they consistently show higher variance in their estimates than the other methods which could lead to a loss of confidence in the methods. Though error estimates may be close to zero on average, if the errors vary widely from data set to data set clinicians may not feel comfortable relying on the patient z-score results. Confidence is needed for each

individual instrument run. As the variance of both parametric methods degrades quickly at high proportions of missingness, it seems wise to restrict estimation to only those features with a maximum of 70% or 80% missing.

2.2. Metabolomic Datasets

2.2.1. Data Summary

Metabolomic datasets utilized in this paper consist of the three same sets used in Chapter 2: plasma (n = 31), cerebral spinal fluid (CSF; n = 31) and urine (n = 40). Relevant characteristics to these experiments are summarized in Table 3.14. The proportion of metabolites containing MVs varies between 46% and 60%, fitting with the authors' previous experience that, as a general rule of thumb, roughly half the metabolites have some level of missing values while the other half are completely observed.

2.2.2. Evaluation

To assess performance of the considered methods in the three metabolomic sets, all metabolites with missing values are first discarded. Then, known values from the remaining metabolites are removed in a left-censored fashion. Each method is then applied to the artificially censored dataset, from which the error between the original sample parameter and the estimated value is found. Continuing with the simulated work, "full" datasets are created in order to provide data to kNN and Random Forest for which to fill in missing values. Such mimic datasets

Table 3.14: Summary of metabolomic data sets indicating the number of samples, metabolites and amount of missing data per set.

		Identified metabolites					Proportion with	Overall proportion of
Matrix	Samples (n)	Neg 1	Neg 2	Pos 1	Pos 2	Total (m)	missing values (%)	missing values (%)
Plasma	31	522	79	218	187	1006	46.10%	16.20%
CSF	31	171	56	203	93	523	51.10%	26.70%
Urine	40	715	125	376	22	1238	60.40%	17.10%

are created by taking those metabolites that are fully observed and randomly selecting half (rounding down) to be censored. For each one of the selected metabolites the lowest k values are removed with $k \sim U(1, n-2)$. This process was repeated 500 times for each matrix.

Following results from Chapter 2, metabolites are log transformed in order to help induce normality. Following censoring, metabolites are also median scaled in order to maintain consistency with the simulations. The complete steps in order are as follows beginning with fully observed data, (1) log transform, (2) randomly sample metabolites for censoring, (3) metabolites scaled based on median of "observed" values, (4) estimated parameter value and compare to uncensored parameter, (5) repeat steps 2 thru 4 five hundred times.

To illustrate the value of variable scaling for kNN, *Figures 3.40-43* show the estimated parameters and their percent bias versus the corresponding sample values in the seven methods. There is a pronounced association between error and average metabolite abundance level under kNN. The effect is subtle in the mean but more pronounced in the standard deviation. Metabolites with high mean values tend to be under predicted in the mean while metabolites with lower mean values tend to be over-predicted in the standard deviation. Association between performance and metabolite abundance is a result of there being fewer metabolites at the extreme margins. For example, in the plasma there are only five metabolites with a mean log abundance level above 22. Thus, when one of these metabolites has missing values some of the neighbors must be of lower average ion abundance, causing these high abundance metabolites to be imputed with lower abundant ones and thus much lower values, dragging down the imputed average. Conversely, the lowest abundant metabolites are forced to rely on metabolites with larger ion counts to replace their missing values, but since kNN generally over imputes missing values the effect is indistinguishable. A similar problem is seen in the standard deviation, where,

Figure 3.40: Estimated mean by true mean after natural log transformation and no further scaling of the metabolites (Plasma).

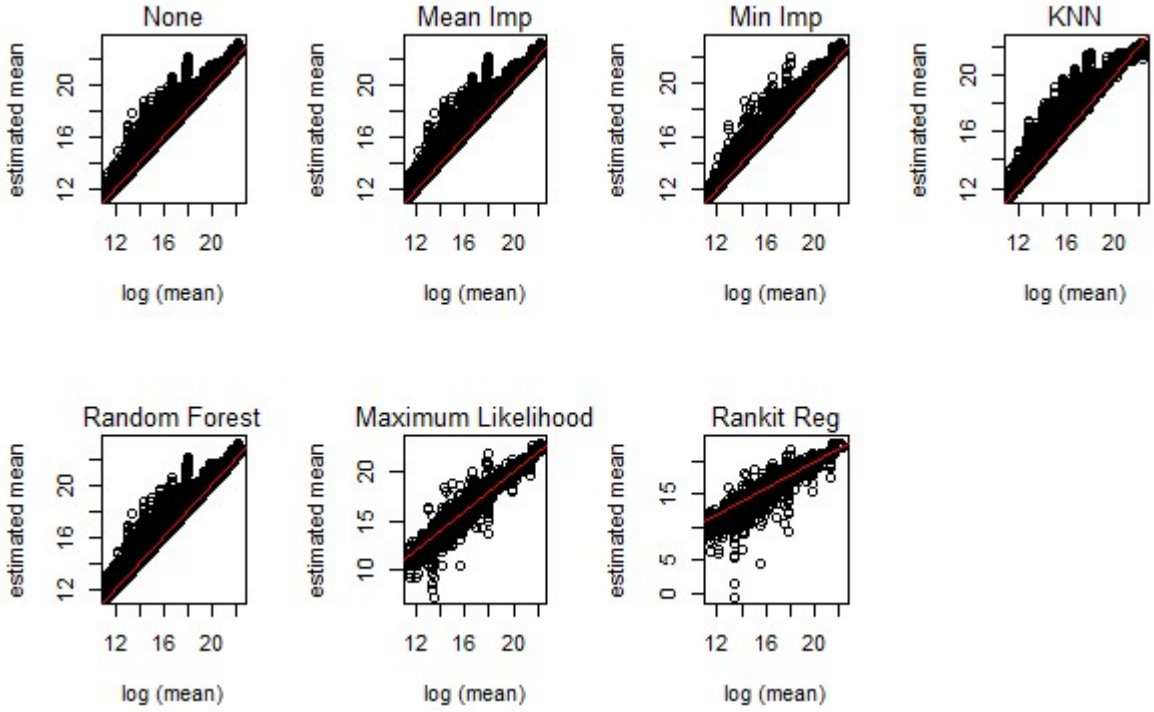


Figure 3.41: Percent bias in mean parameter by average log abundance without median scaling. kNN under predicts highest abundant metabolites due to neighbors generally being of lower abundance. (Plasma)

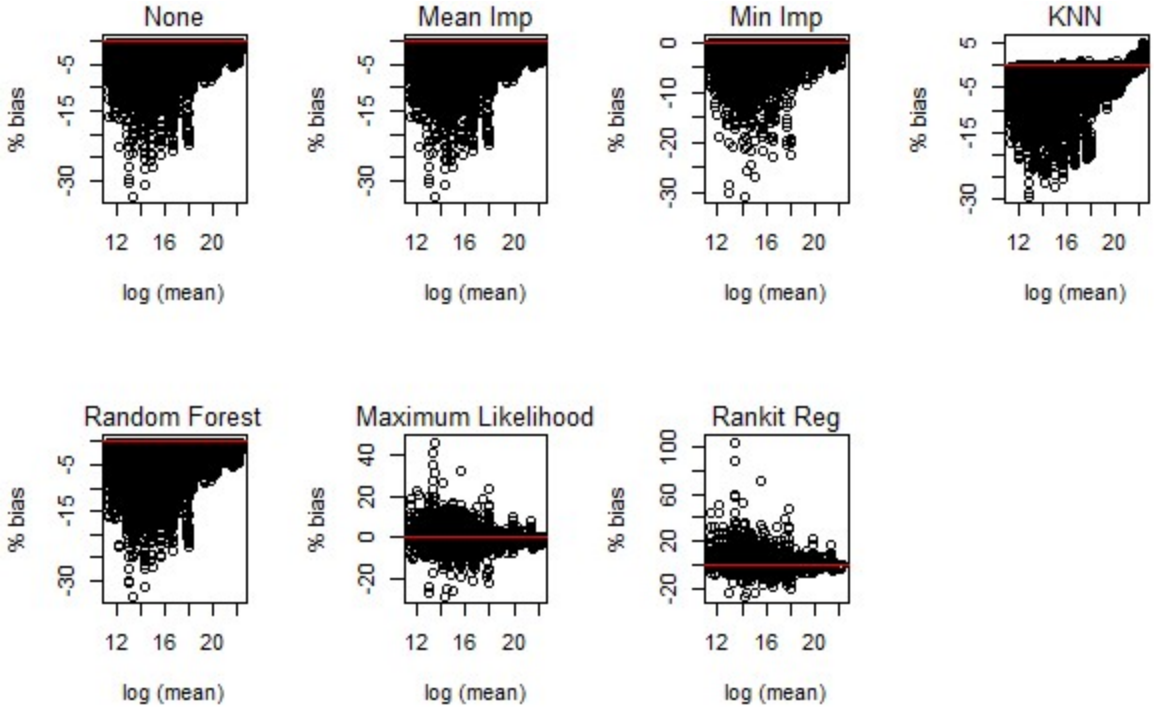


Figure 3.42: Estimated standard deviation by true standard deviation after natural log transformation and no further scaling of the metabolites. Plasma data set.

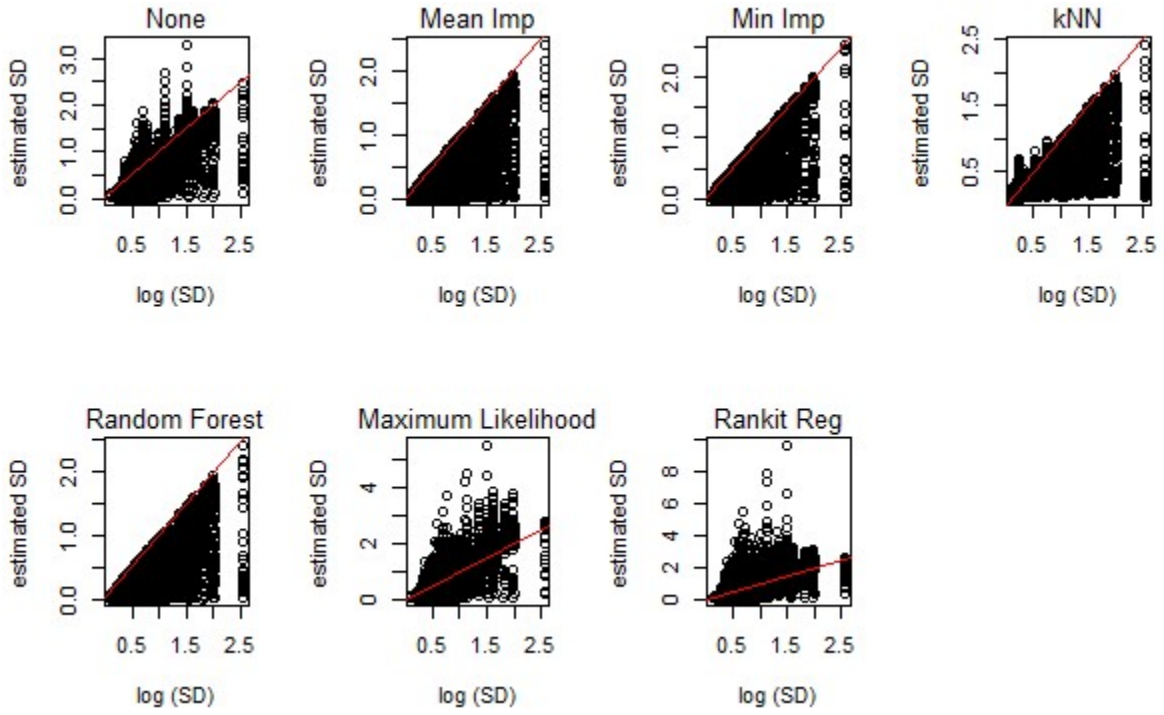
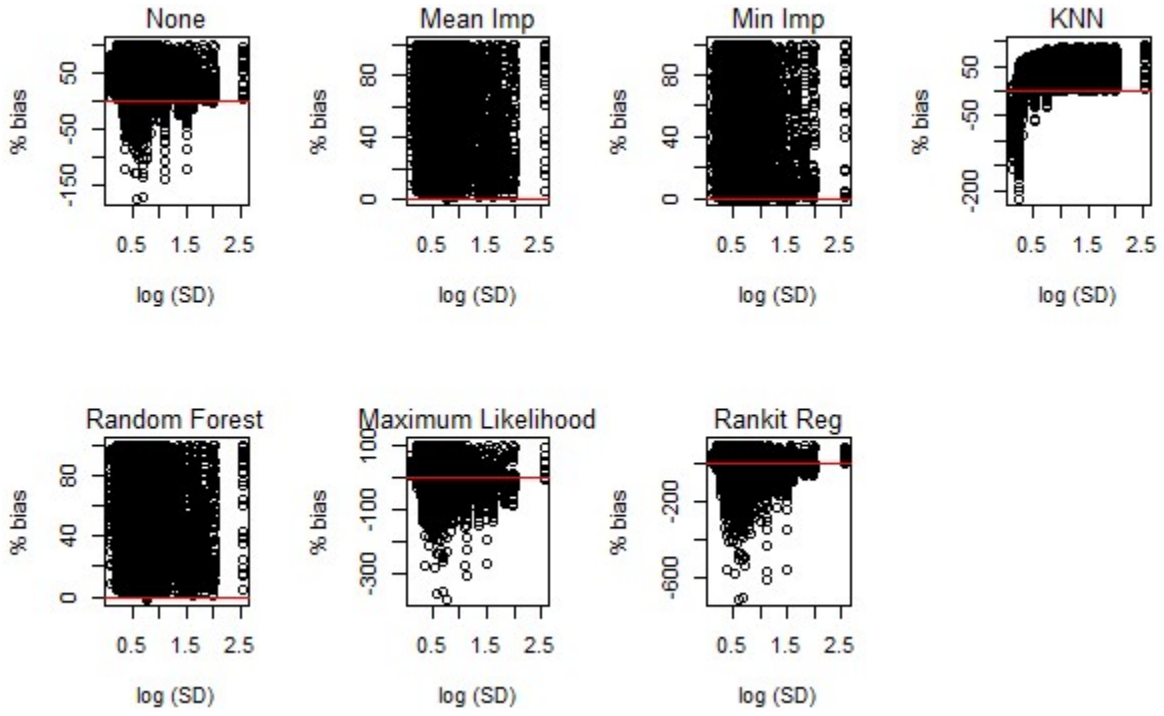


Figure 3.43: Percent bias in standard deviation parameter by log standard deviation without median scaling.



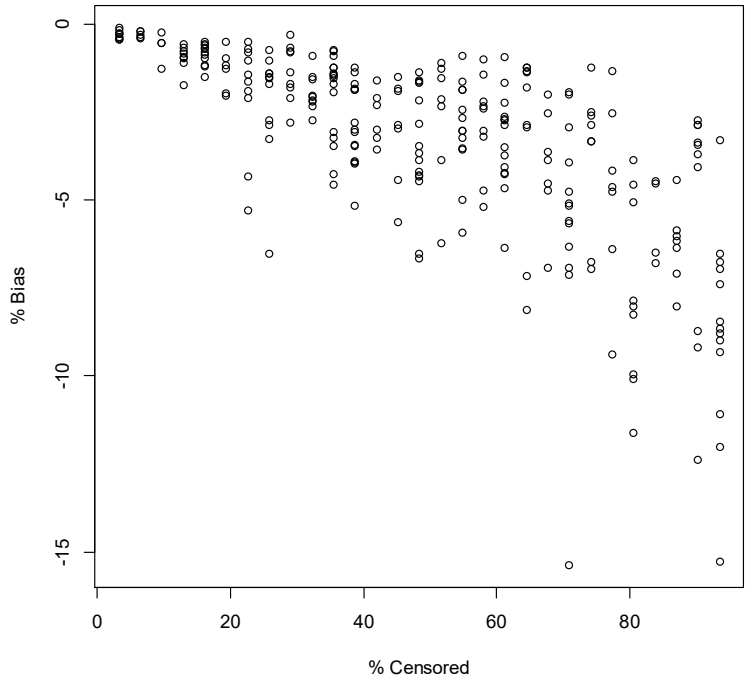
due to the different scaling, it is relatively easy to manufacture a 2- to 3-fold increase. However, the other methods are shown to be consistent across the range of the uncensored parameter values and, as seen in the subsequent results, variable scaling helps to remove this phenomenon in kNN as well.

The methods considered will be evaluated by examining the error, as described in section 3.6, between the estimated mean and standard deviation versus their original, uncensored parameter values. The intention of viewing the results in this manner is to address the concept of bias, which relates the expected value of an estimator against the parameter it is trying to estimate. However, direct comparison to the population level mean and standard deviation for the metabolites is challenging for two reasons. First, knowledge of metabolite behavior at the population level may not be readily available. The Human Metabolome Database [42] does provide references to normal levels for many metabolites, but when performing a global profile of a sample there is no guarantee that out of the many hundreds detected a reference range will be available for all metabolites given the patient demographics and sample type being profiled. Second, the semi-quantitative nature of the instrumentation implies that the ion counts in any batch do not directly infer a concentration from which population knowledge can be translated. Use of the dataset leads to comparisons with the fully observed sample mean and standard deviation, which are themselves unbiased estimators of their respective population values. How the methods compare to these statistics thus provides some inference about bias while also informing how well these methods compared to the fully observed data.

Lastly, all elements of data pre-processing and processing are worthy candidates for influencing the results of data analysis. However, these items work is beyond the scope of this dissertation. Instead, we use the best available practice, which in the case of urine involves

normalization [100, 102-104]. While plasma and CSF are homeostatic, metabolite concentrations in urine are known to be closely associated with creatinine levels and as a result creatinine is frequently used to adjust for dilution effects and was previously performed in this data [3]. A creatinine factor was created by dividing each sample's creatinine level by the overall average creatinine level and then normalizing by this factor prior to log transformation. This way the overall ion counts of the features have roughly the same abundance level after normalization as

Figure 3.44: Percent bias under NONE from one image of plasma. Each point represents a biochemical selected for censoring in the image.



they did before. Maintaining the original abundance scale in the urine prior to log transformation retains consistency with the processing steps in the other matrices.

2.2.3. Results

Results of NONE estimating the Mean parameter in one simulation from the plasma dataset are shown in *Figure 3.44*. Each point represents a metabolite selected for censoring with the x-axis being the proportion of samples removed, k_j/n, an d the y-axis being the relative error. Points below the line y = 0 indicate estimated mean for that metabolite was more than the true

mean while positive values indicate the estimate was lower. For this particular method and parameter, all the points are negative as the mean of the observed values will always be more than uncensored mean when values are removed from below. As expected, error increases as the proportion of values removed increases, though clearly some metabolites are more impacted than others. When comparing all seven methods it is most convenient to examine the relative error as a function of k_j/n, similar to the simulated result average bias plots using in the simulations. Performance was remarkably similar across the different datasets. The two parameters are first

2.2.3.1. Mean Parameter

Figures 3.45-53 displays the error, relative error and average relative error plots, just as the simulations, in the mean parameter for all seven methods across the three datasets. These figures are rather consistent across the three matrices, indicating that the five common methods consistently overestimate the mean parameter while both proposed approaches are less biased. The pattern is similar to those of the normal simulation when missing values are left-censored. One exception is that the two proposed methods sometimes demonstrate a notable bias towards over or under estimates across the three sets. The magnitude of the errors is consistent with the normal, left censored simulations though the relative percentage does appear a bit higher in urine.

Random Forest and kNN track very closely to NONE/AVG indicating that, as expected, the low correlation in chemo-centric sets is not sufficient to strengthen these methods very much. A series of splinal plots are shown in *Figures 3.54-56* representing a smoothed curve for relative error as function of k_j/n. The solid black line gives the smooth spline of the average, while the dotted bands are the 1st and 3rd quartiles, essentially represent Inner Quartile Range as a function of x. Average bias most commonly manifests as over-estimates of the true value. All five of the

Figure 3.45: Estimated by uncensored mean in Plasma. Each point represents an individual biochemical selected for censoring. Red line represents y = x.

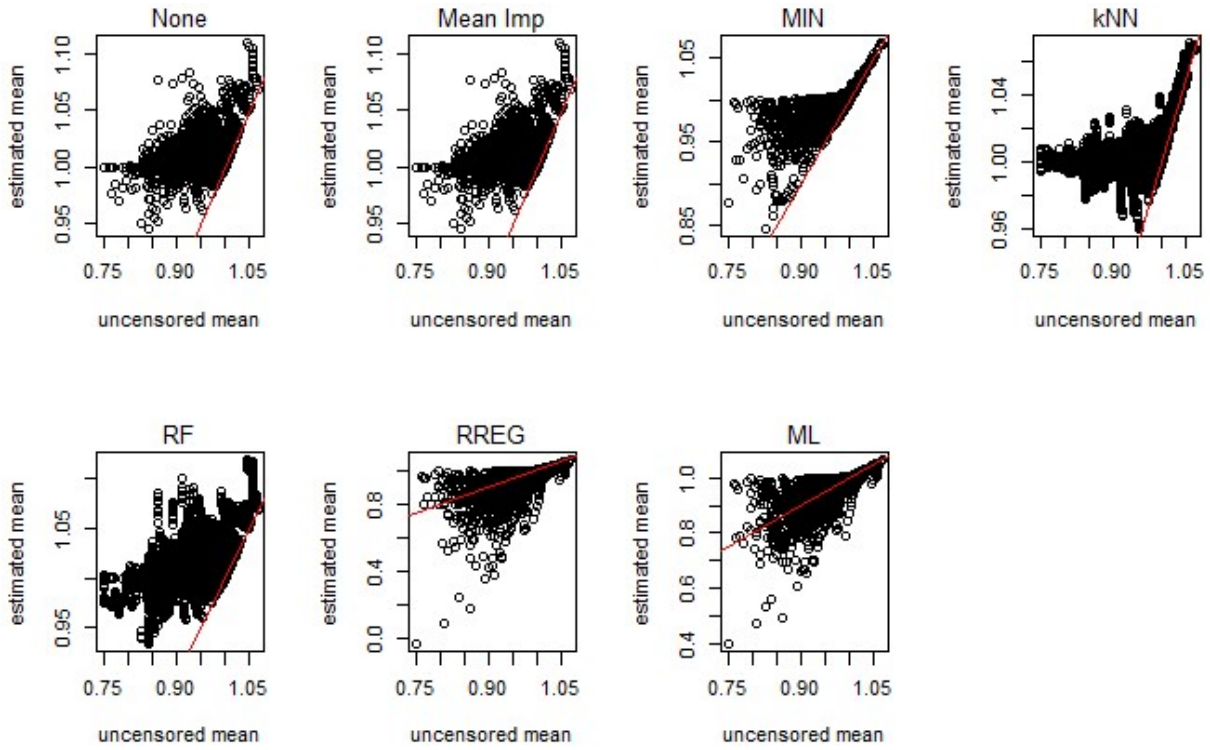


Figure 3.46: Percent error of mean parameter in Plasma. Each point represents an individual biochemical selected for censoring. Red line represents y=0.

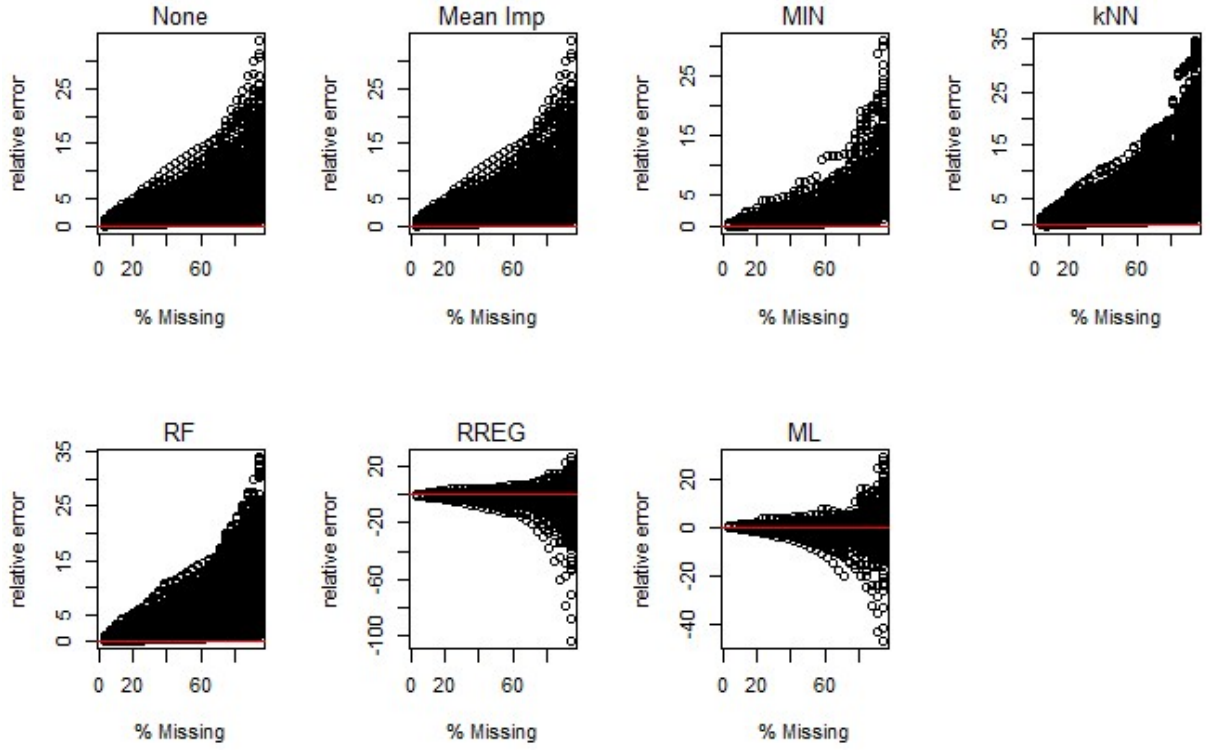


Figure 3.47: Percent error of mean parameter in CSF. Each point represents an individual biochemical selected for censoring. Red line represents y=0.

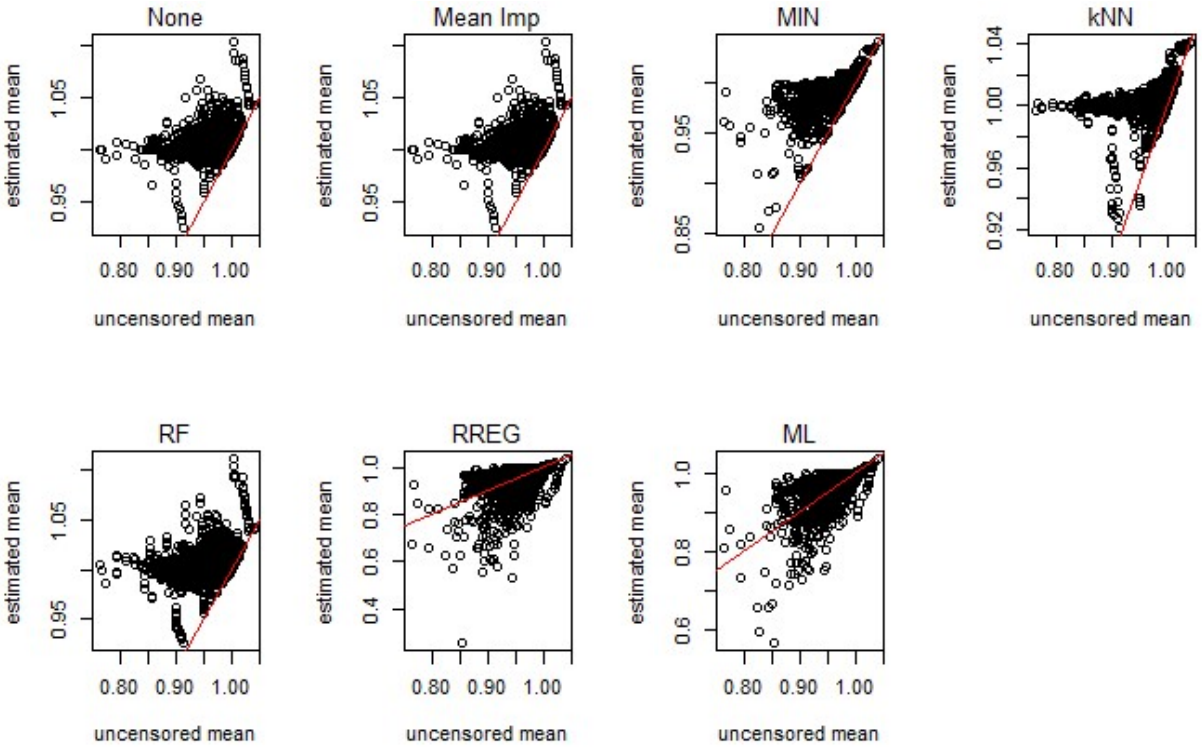


Figure 3.48: Estimated by uncensored mean in CSF. Each point represents an individual biochemical selected for censoring. Red line represents y=x.

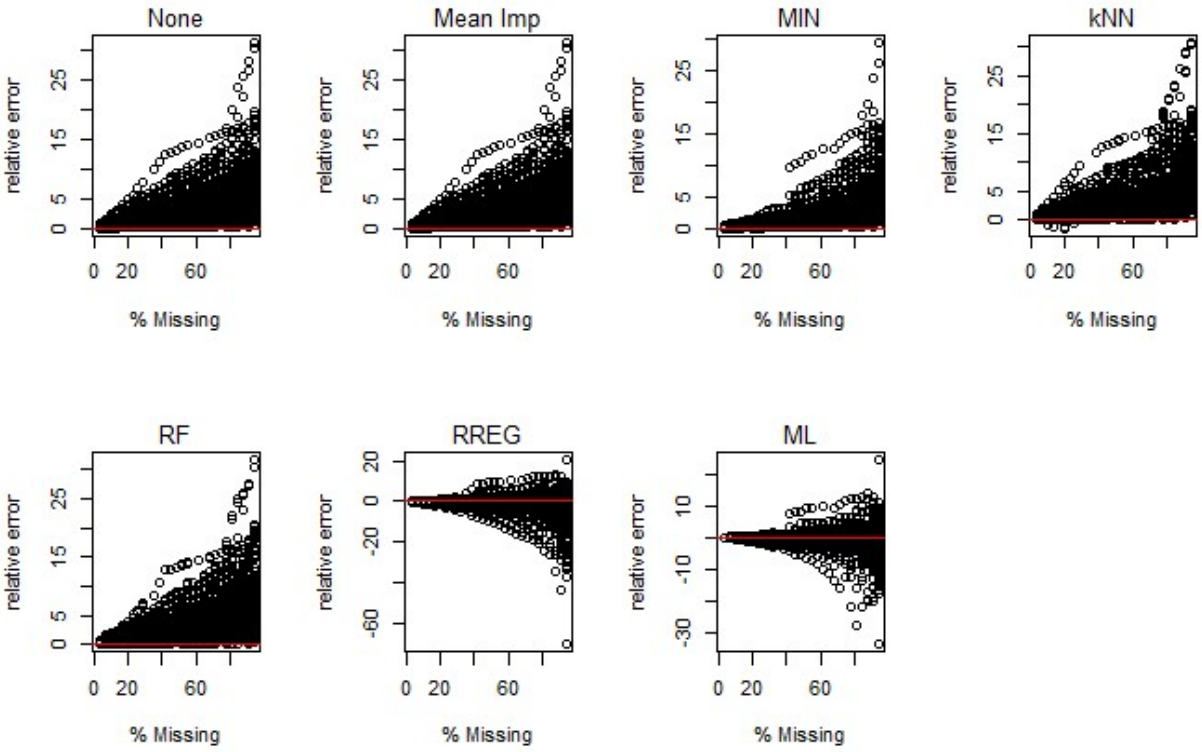


Figure 3.49: Percent error of mean parameter in Urine. Each point represents an individual biochemical selected for censoring. Red line represents y=0.

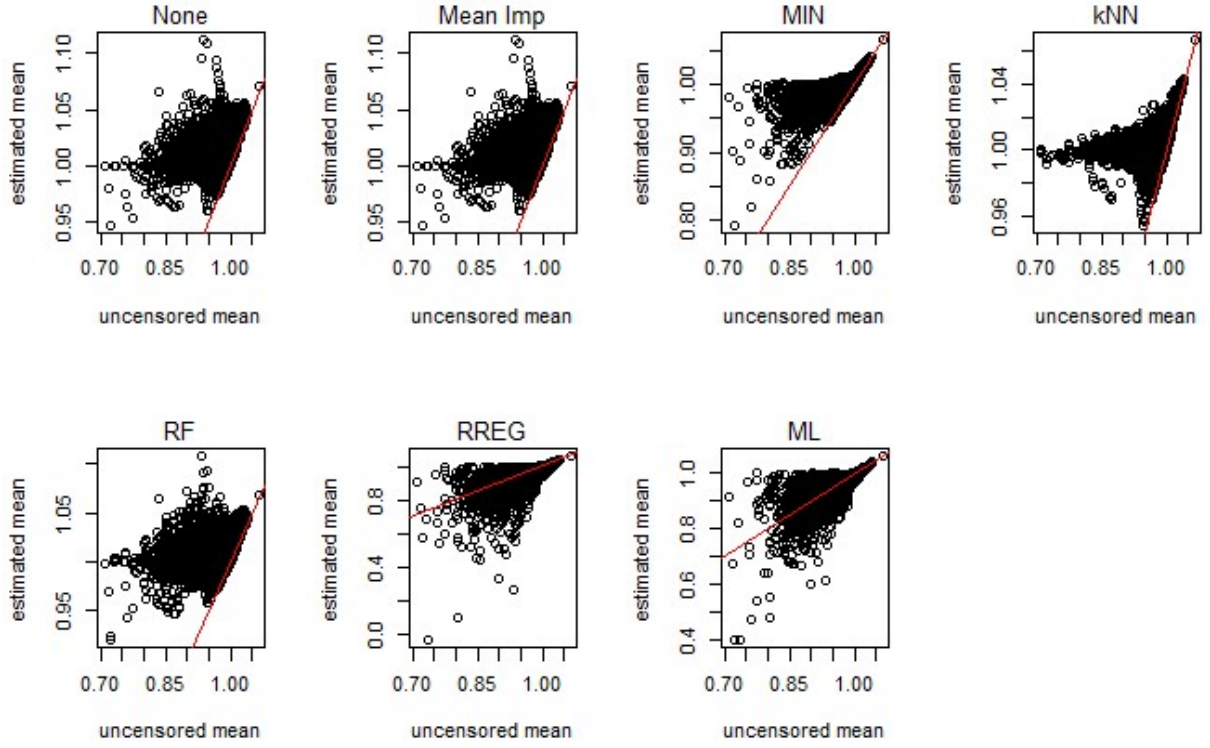


Figure 3.50: Estimated by uncensored mean in Urine. individual biochemical selected for censoring. Red line indicates y=x.

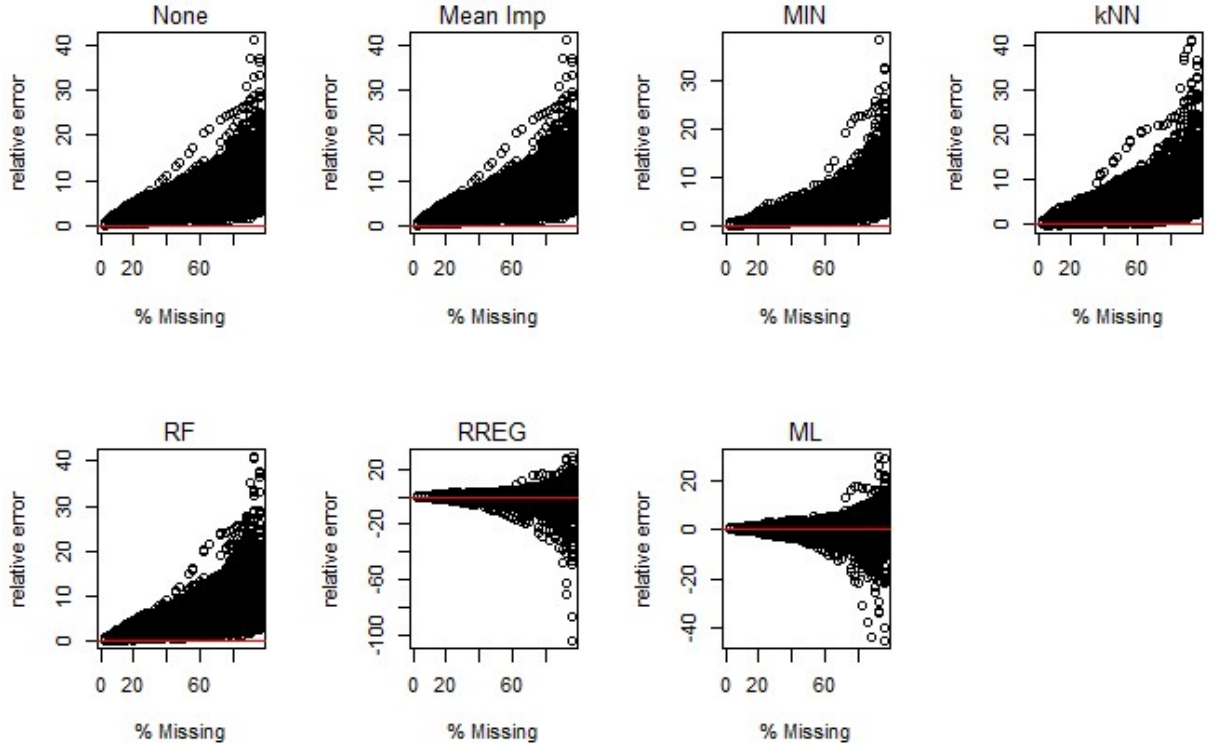


Figure 3.51: Average relative error of missing data methods for mean parameter in Plasma.

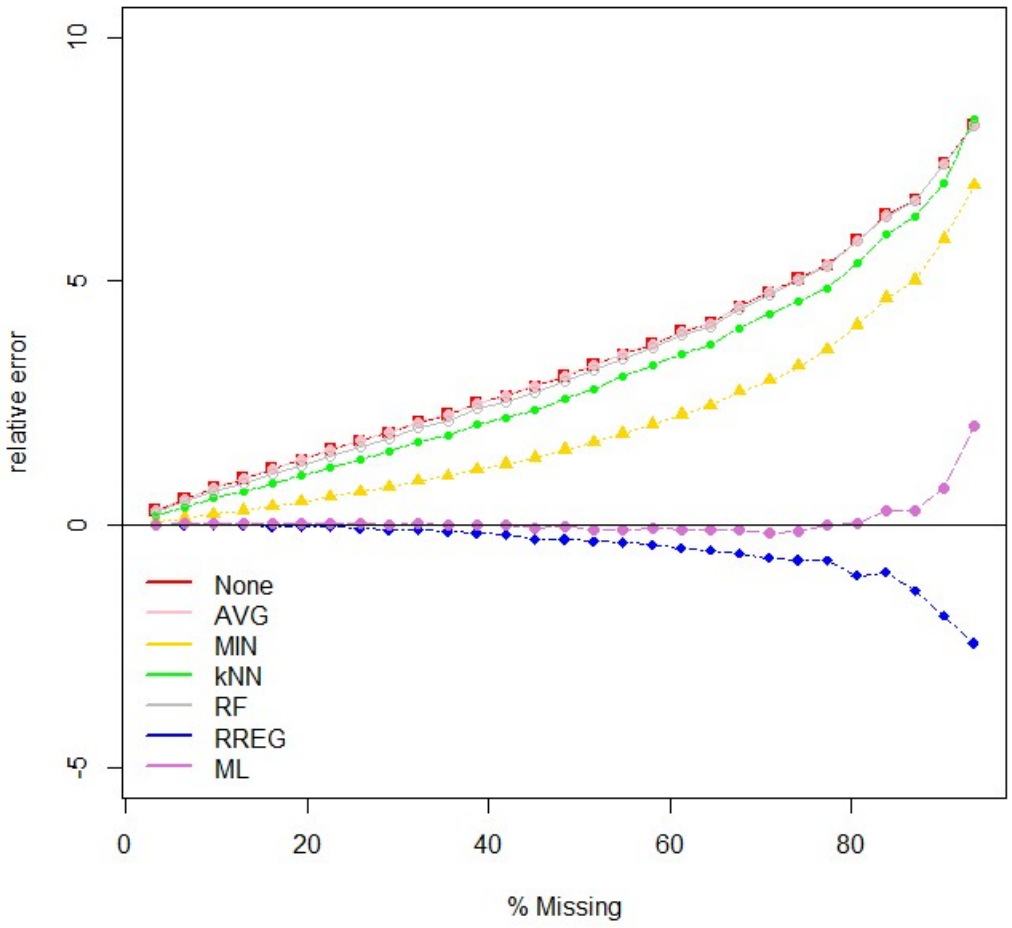


Figure 3.52: Percent bias of missing data methods for mean parameter in CSF.

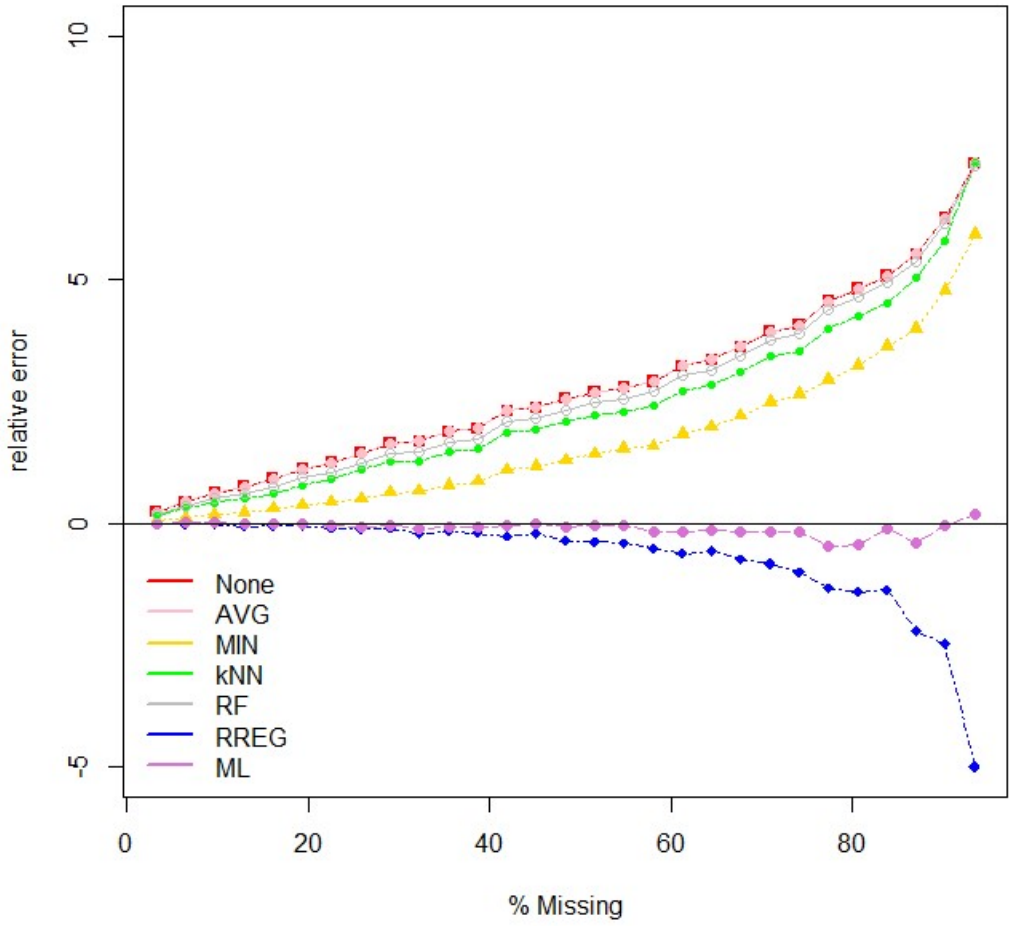


Figure 3.53: Percent bias of missing data methods for mean parameter in Urine.

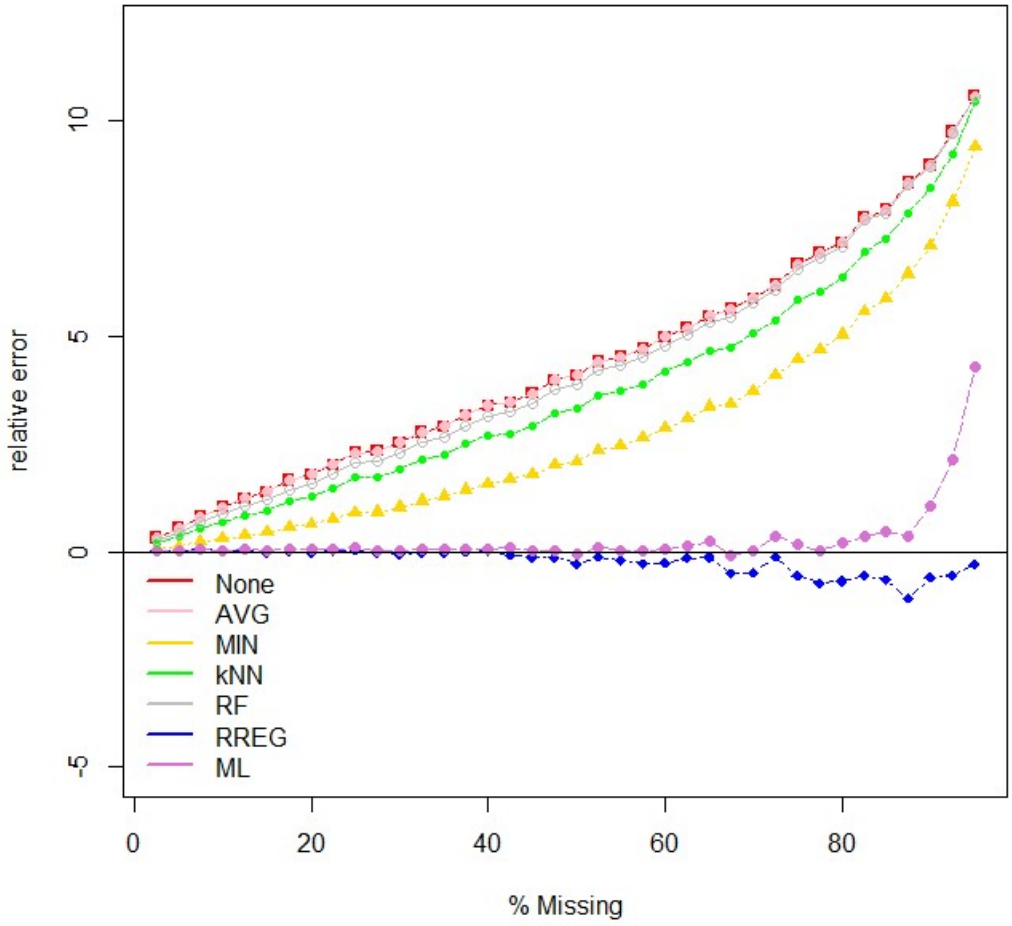


Figure 3.54: Trend plot for percent bias of mean parameter in Plasma. Trend lines are based on splinal fits. Solid line is the average while dotted are first and third quartiles. Red line indicates no error.

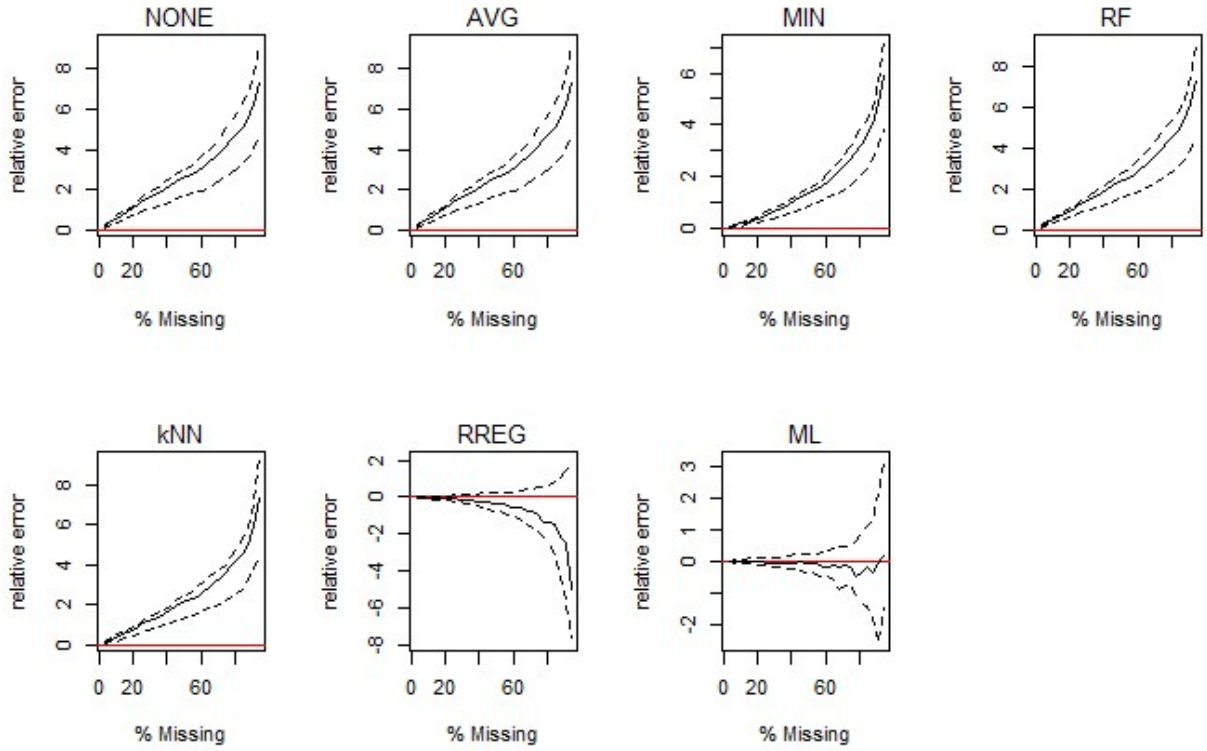
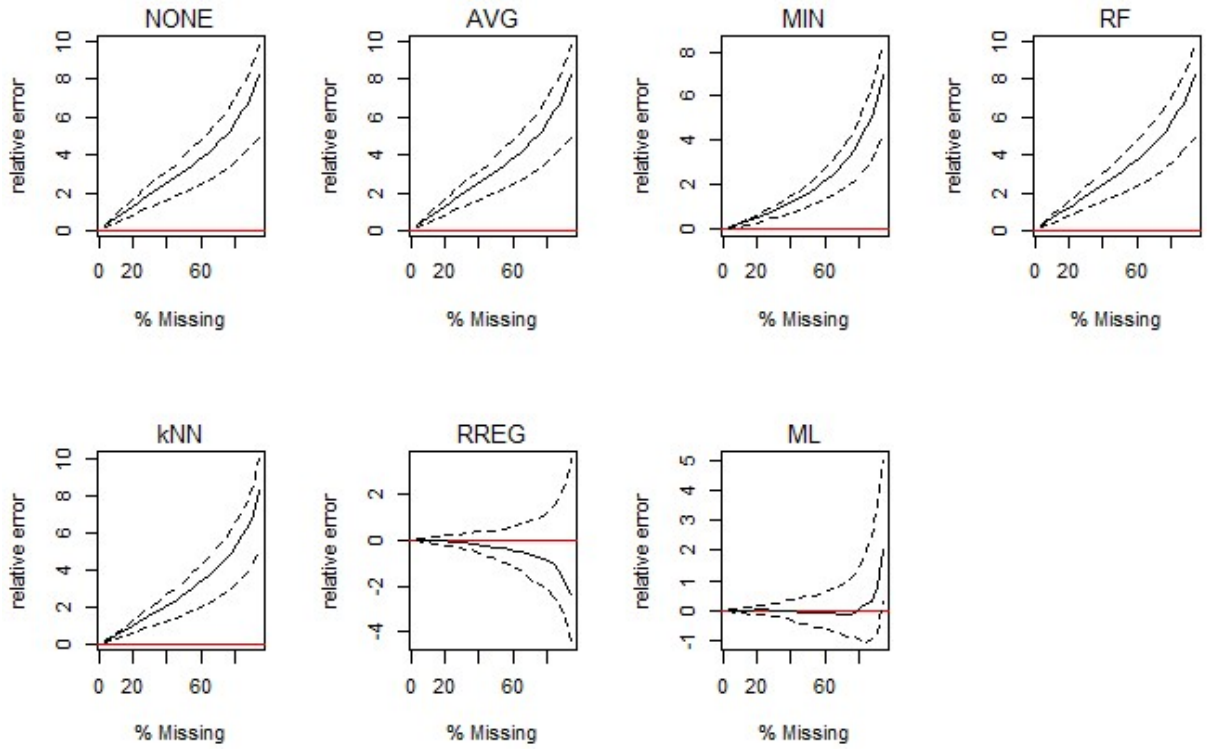


Figure 3.55: Trend plot for percent bias of mean parameter in Plasma. Trend lines are based on splinal fits. Solid line is the average while dotted are first and third quartiles.



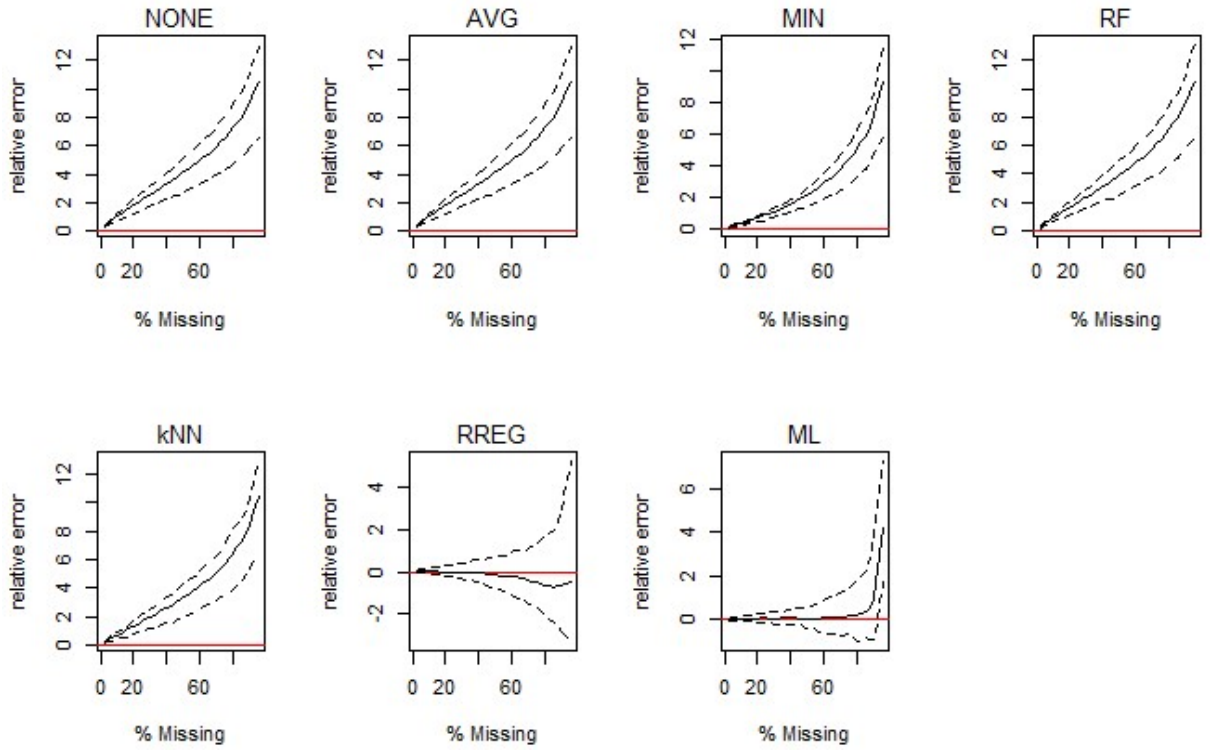


Figure 3.56: Trend plot of percent bias of mean parameter in Urine. Trend lines are based on splinal fits. Solid line is the average while dotted are first and third quartiles.

common metabolomics estimators have similar patterns, with the amount of bias steadily increasing as the amount of censored data increases. Both kNN and RF have strong similarity with the single imputation methods implying the observed data is not very useful at informing on the unobserved portions. In comparison both parametric methods were found to be unbiased up until the proportion of missing values reaches 80%. However, for all methods the bias is generally small, being no more than 2-5% for even moderate amounts of missing data. Variance is also notable. These plots make clear another similarity between the simulated output and the real data, namely the increased variability of the proposed methods as the proportion of missingness increases. The IQR bands dramatically increase as the missing proportion reaches 80%, which in these datasets implies 6 or fewer observations, and this behavior is strongest for RREG.

2.2.3.2. Standard Deviation Parameter

Plots for the standard deviation are given in *Figures 3.57-68*. As with the mean parameter, the results are very similar across the three matrices and are consistent with the normal, left-censored simulations. The five methods consistently underestimate the true variation in the metabolites. AVG, MIN and RF appear as the most biased with magnitudes reaching around 50% for missing proportions between 30-60%, and the error nears 80% when the proportion of missing values exceeds 80%. kNN demonstrates slightly better behavior with higher levels of missing values, indicating the neighbor approach can be useful at extracting structural information in extreme circumstances, but overall the standard deviation has been significantly decreased. Of the common metabolomics approaches, NONE yields as the least bias. In contrast to the common methods, both ML and RREG are practically unbiased throughout. ML is roughly unbiased on average up through 80% missing values, at which point it begins to underestimate the standard deviation. Conversely, RREG begins to show bias towards over estimate beyond 50% missing values in the plasma and CSF. However, in urine this RREG shows less bias than ML throughout the range of missing values.

2.3. Summary

This paper demonstrates the usefulness of parametric approaches to handle missing data in metabolomics, specifically a Gaussian model for the metabolites. Following a natural log-transformation, maximum likelihood and rankit regression are shown to produce estimates of population parameters that are, for all practical purposes, unbiased up to even large proportions of MVs. The ability to produce accurate population estimates in the presence of missing values has direct value for diagnostic screening, including the identification of IEMs. Additionally, these results were observed in three different types of human material following a log

Figure 3.57: Estimated by uncensored SD in Plasma. Each point represents an individual biochemical selected for censoring. Red line represents y=x.

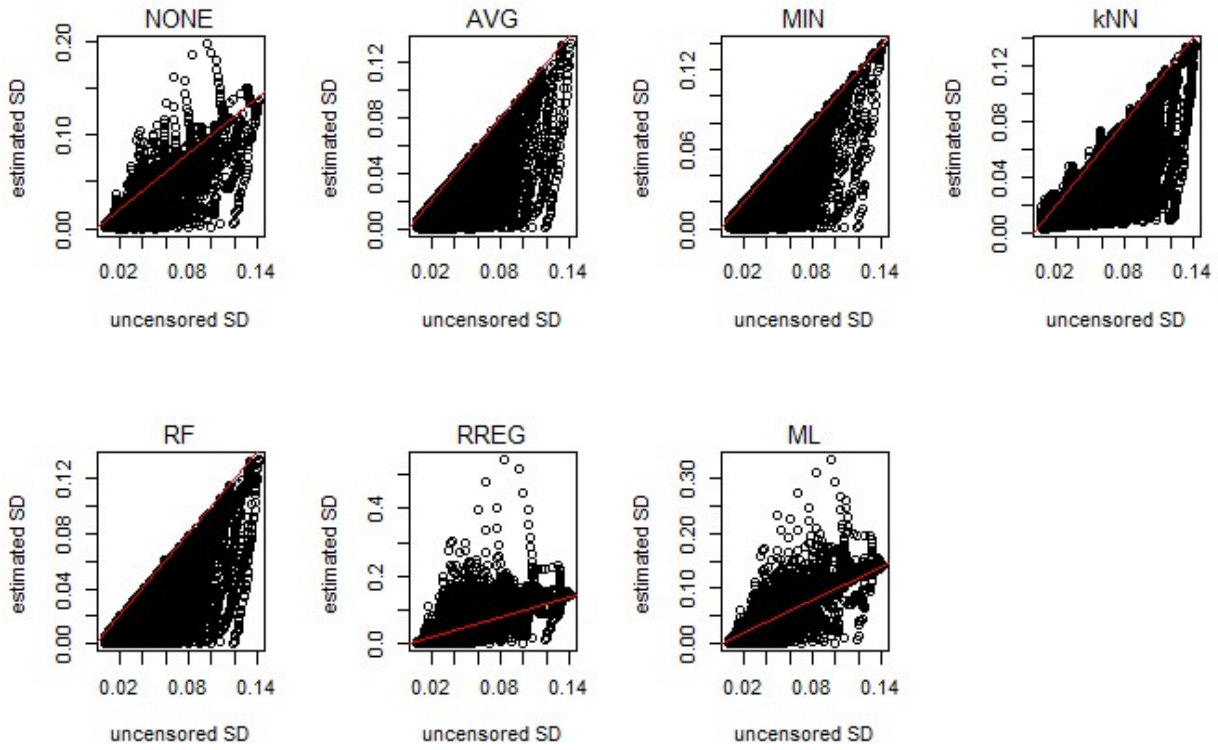


Figure 3.58: Percent error of SD parameter in Plasma. Each point represents an individual biochemical selected for censoring. Red line represents y=0.

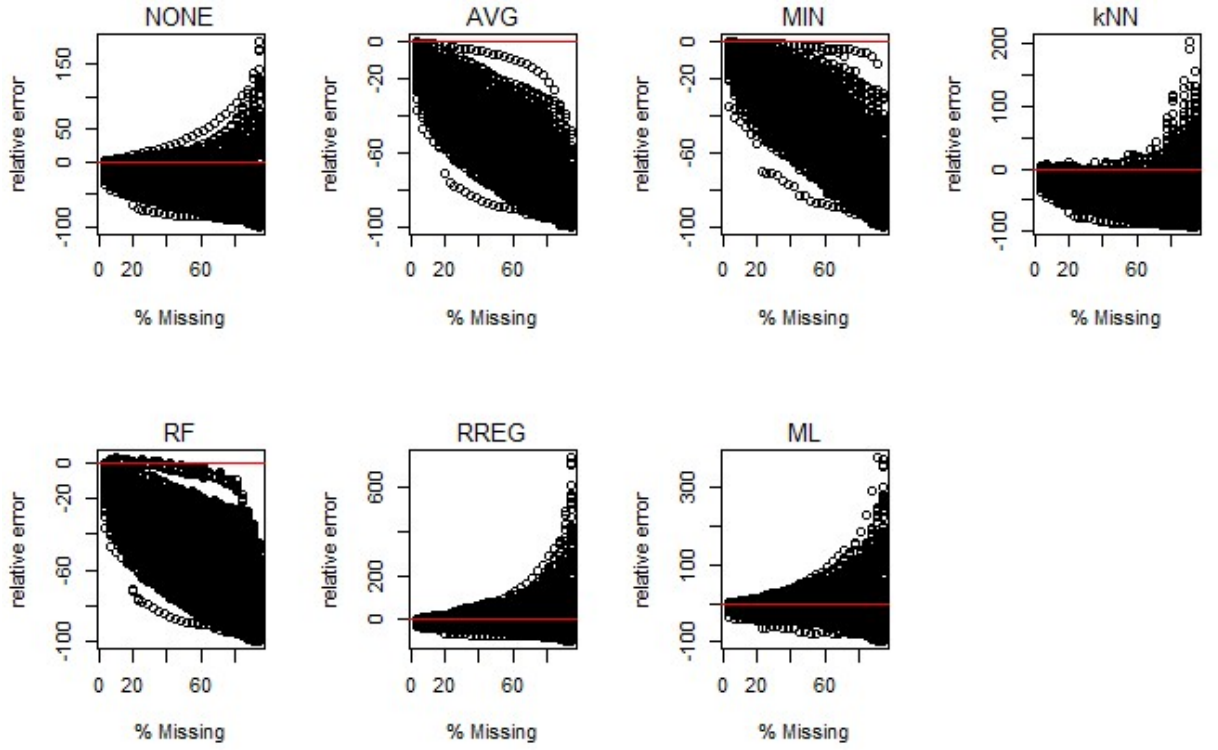


Figure 3.59: Estimated by uncensored SD in CSF. Each point represents an individual biochemical selected for censoring. Red line represents y = x.

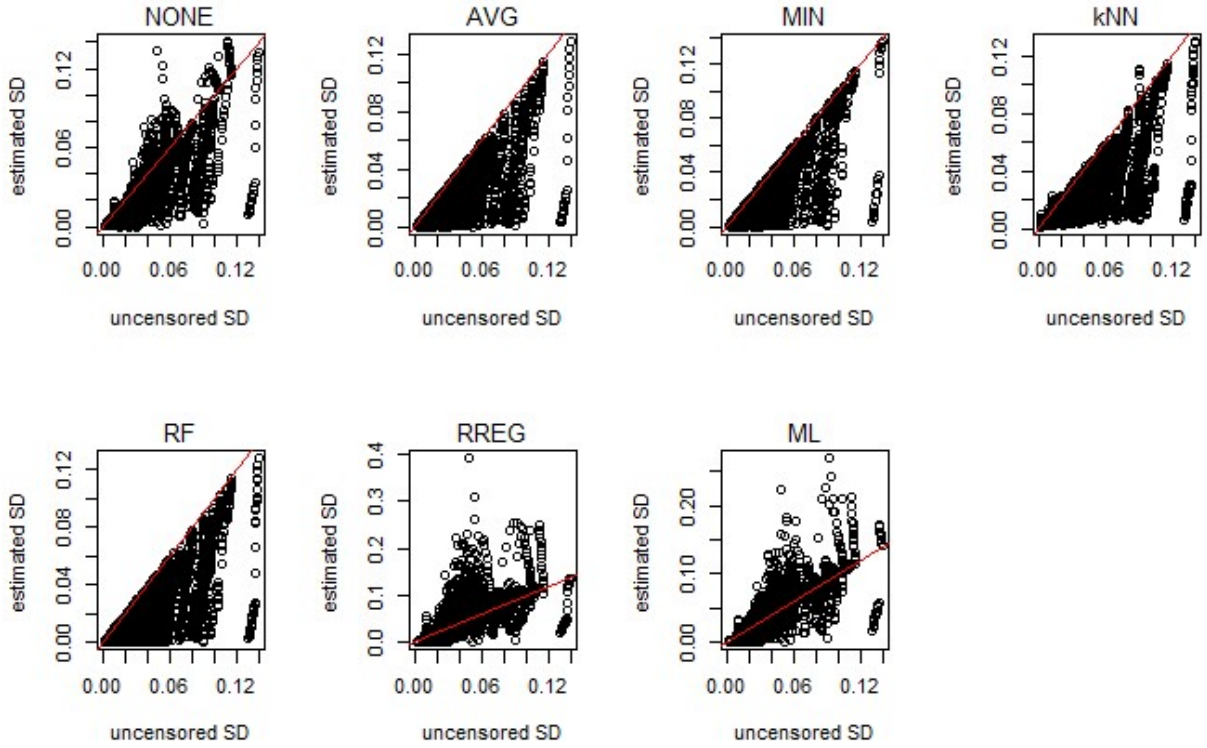


Figure 3.60: Percent error of SD parameter in CSF. Each point represents an individual biochemical selected for censoring. Red line represents y=0.

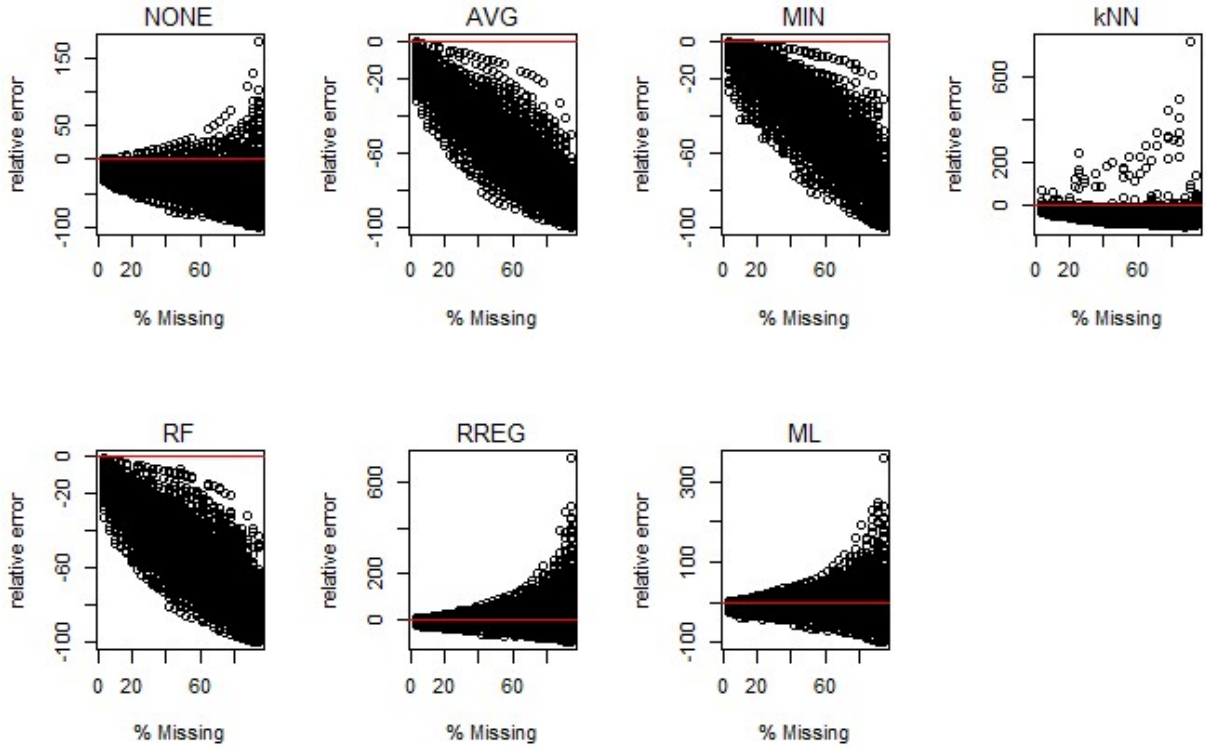


Figure 3.61: Estimated by uncensored SD in Urine. Each point represents an individual biochemical selected for censoring. Red line represents y = x.

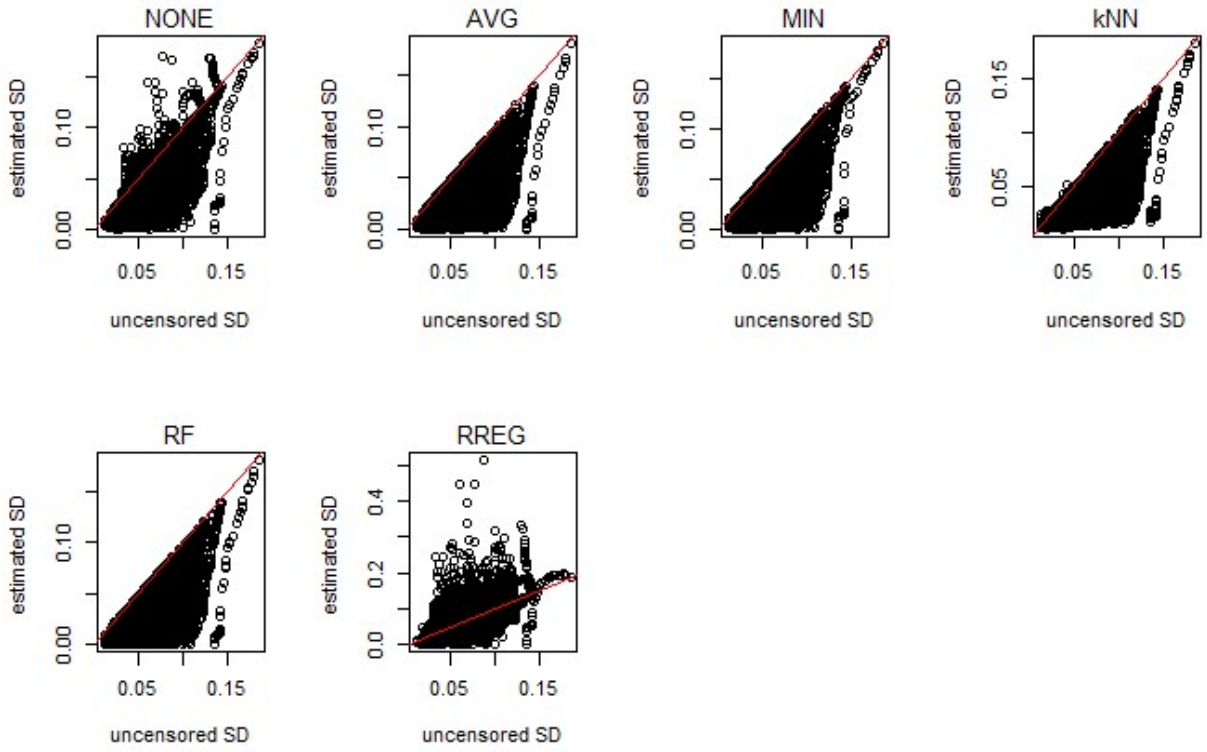


Figure 3.62: Percent error of SD parameter in Urine. Each point represents an individual biochemical selected for censoring. Red line represents y=0.

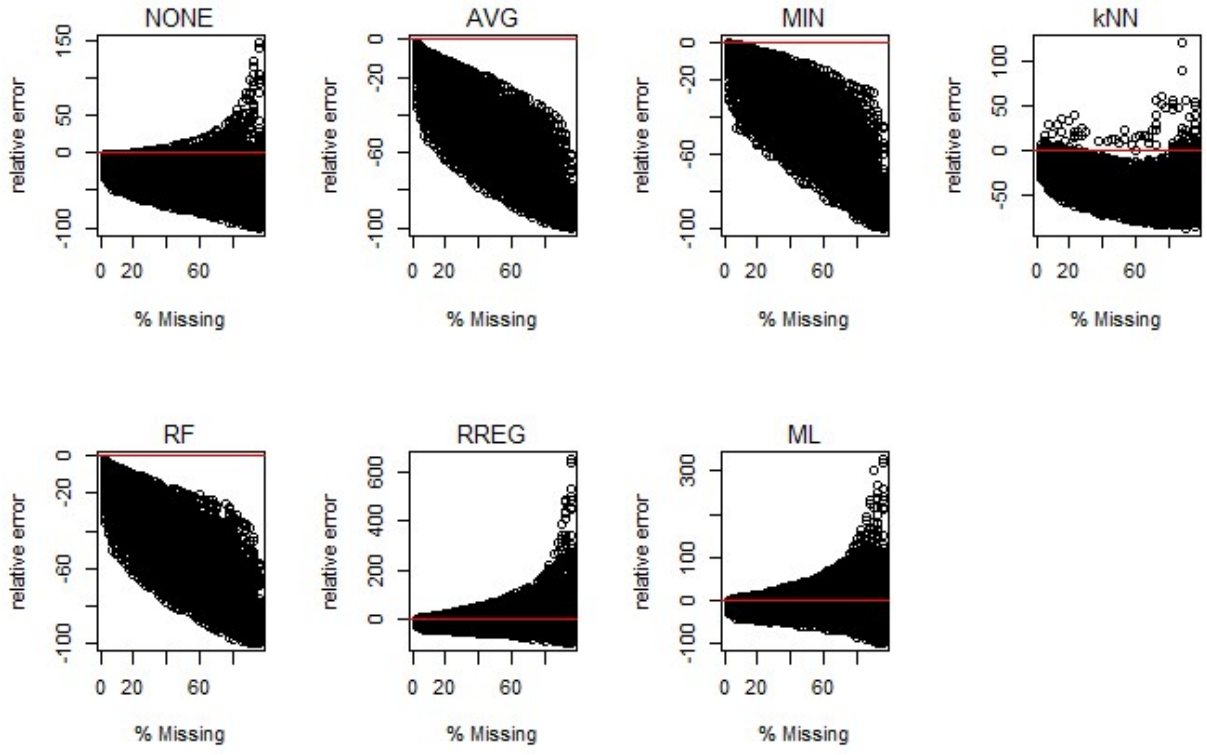


Figure 3.63: Percent bias of missing data methods for SD parameter in Plasma.

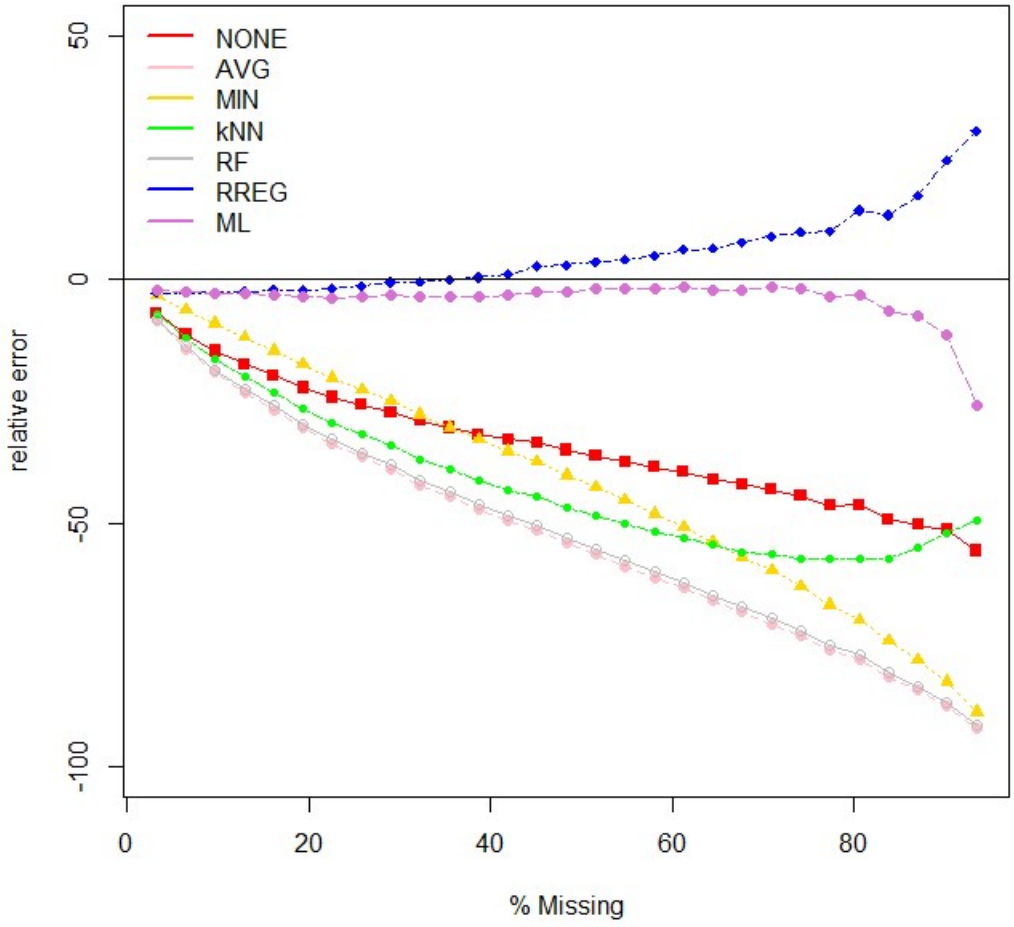


Figure 3.64: Percent bias of missing data methods for SD parameter in CSF.

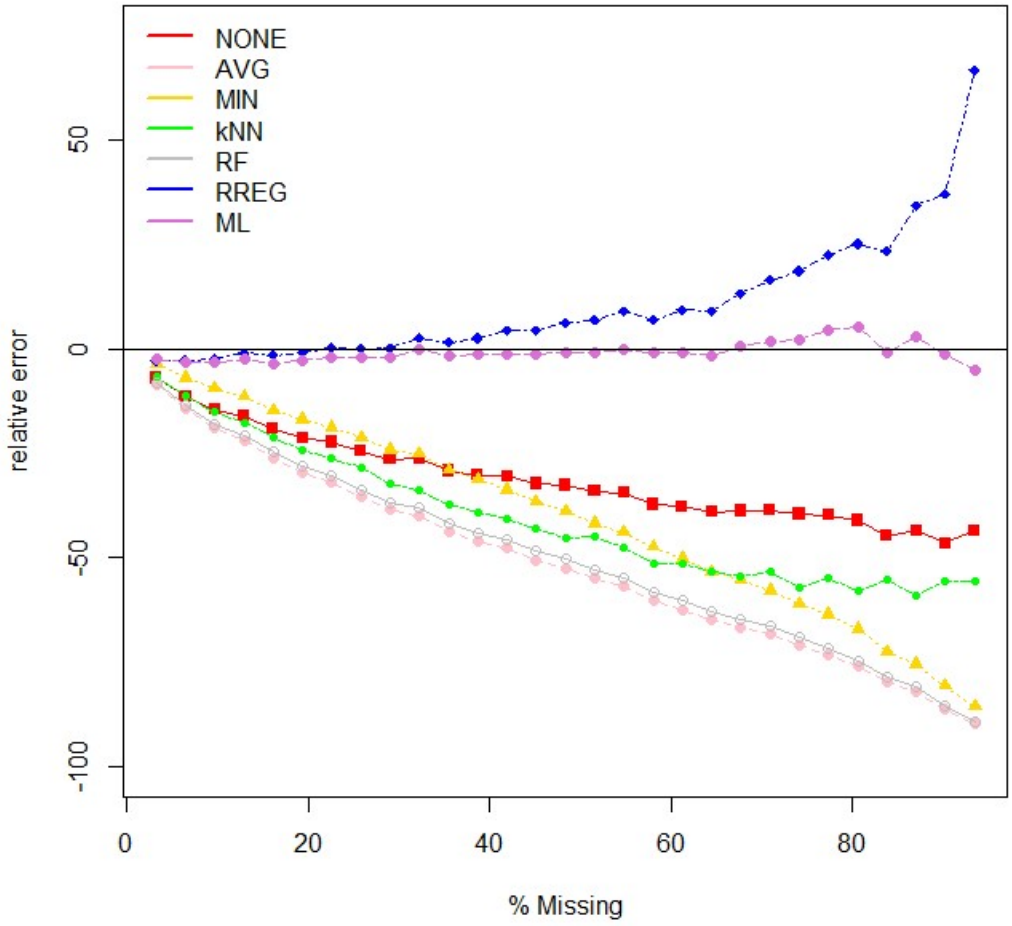


Figure 3.65: Percent bias of missing data methods for SD parameter in Urine.

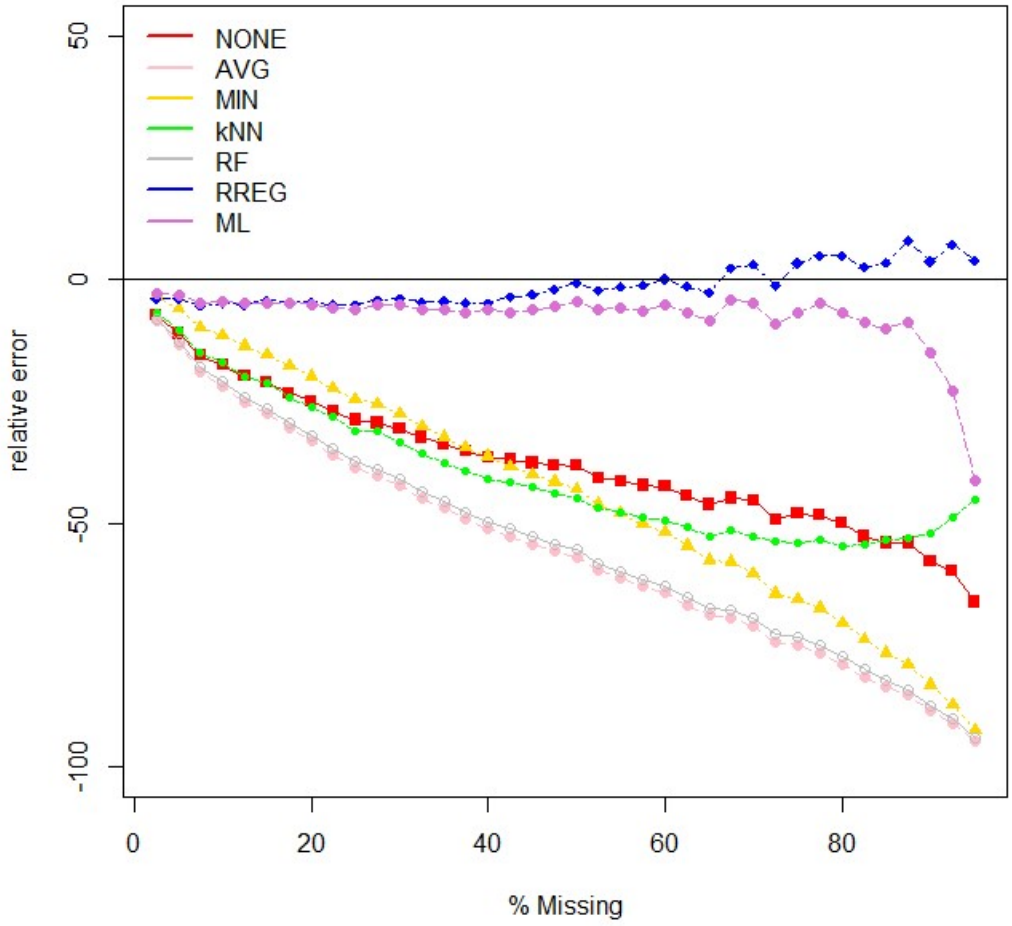


Figure 3.66: Trend plot for percent bias of SD parameter in Plasma. Trend lines are based on splinal fits. Solid line is the average while dotted are first and third quartiles.

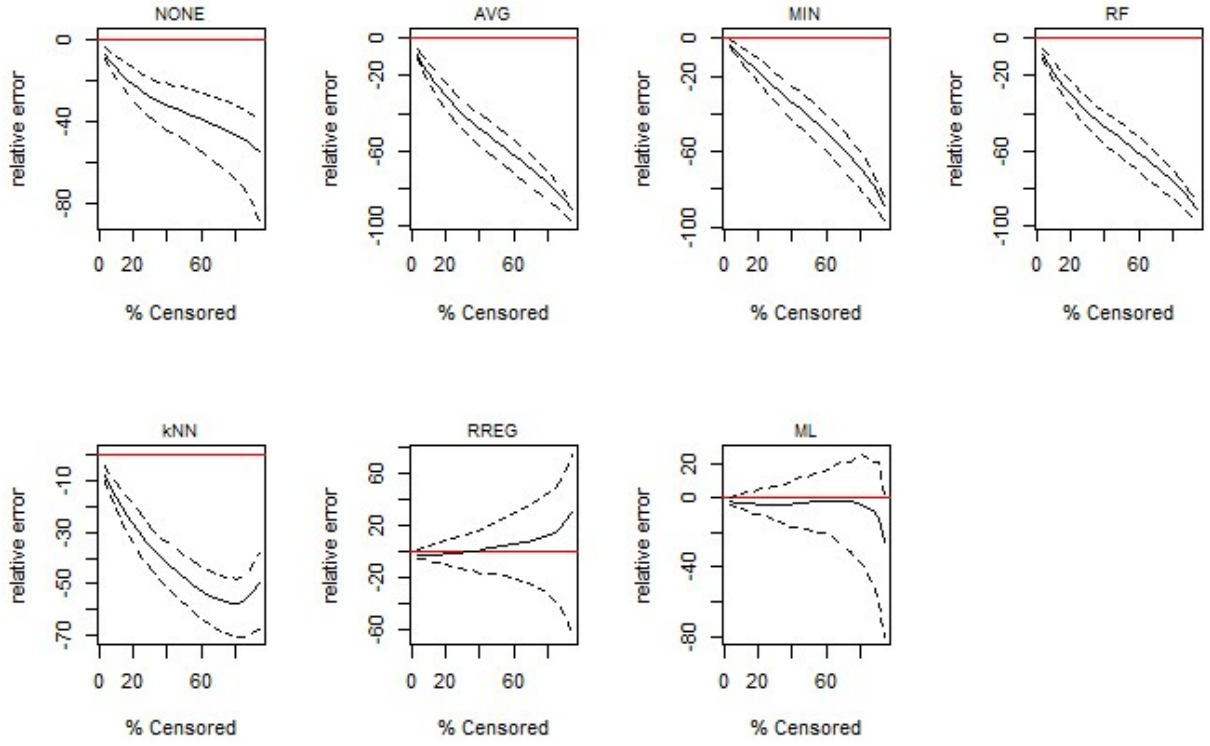
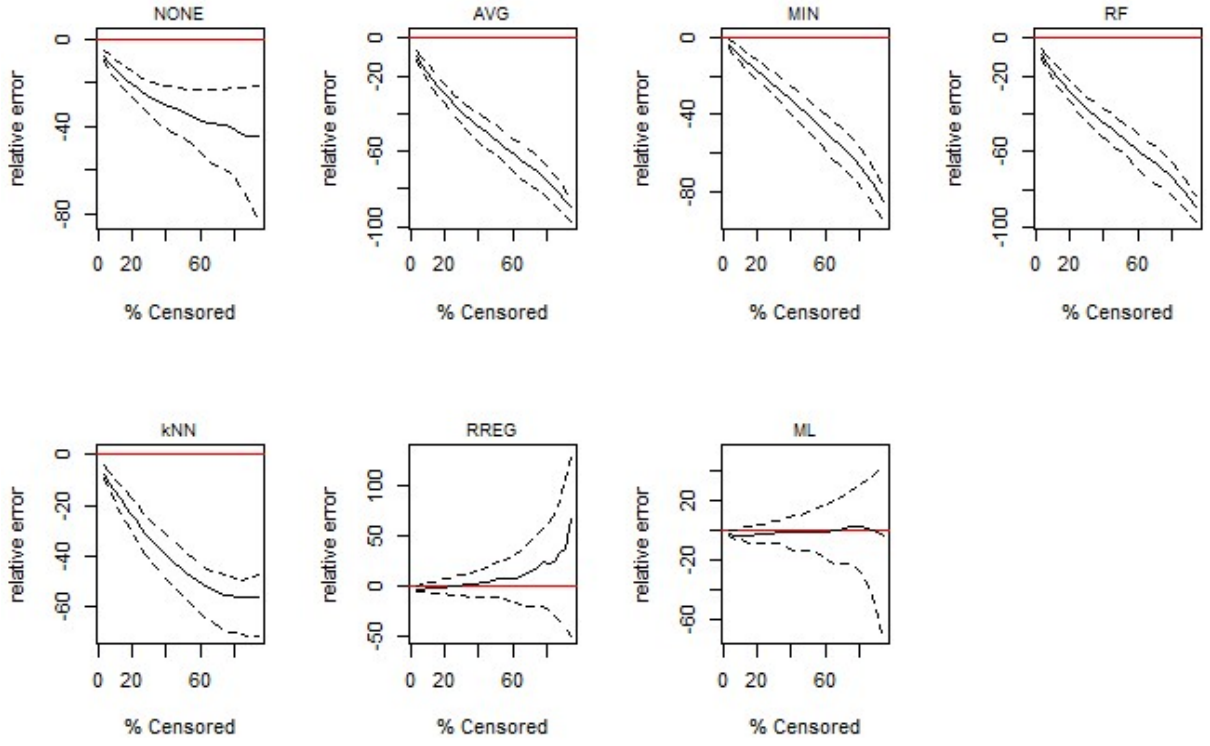


Figure 3.67: Trend plot for percent bias of SD parameter in Plasma. Trend lines are based on splinal fits. Solid line is the average while dotted are first and third quartiles.



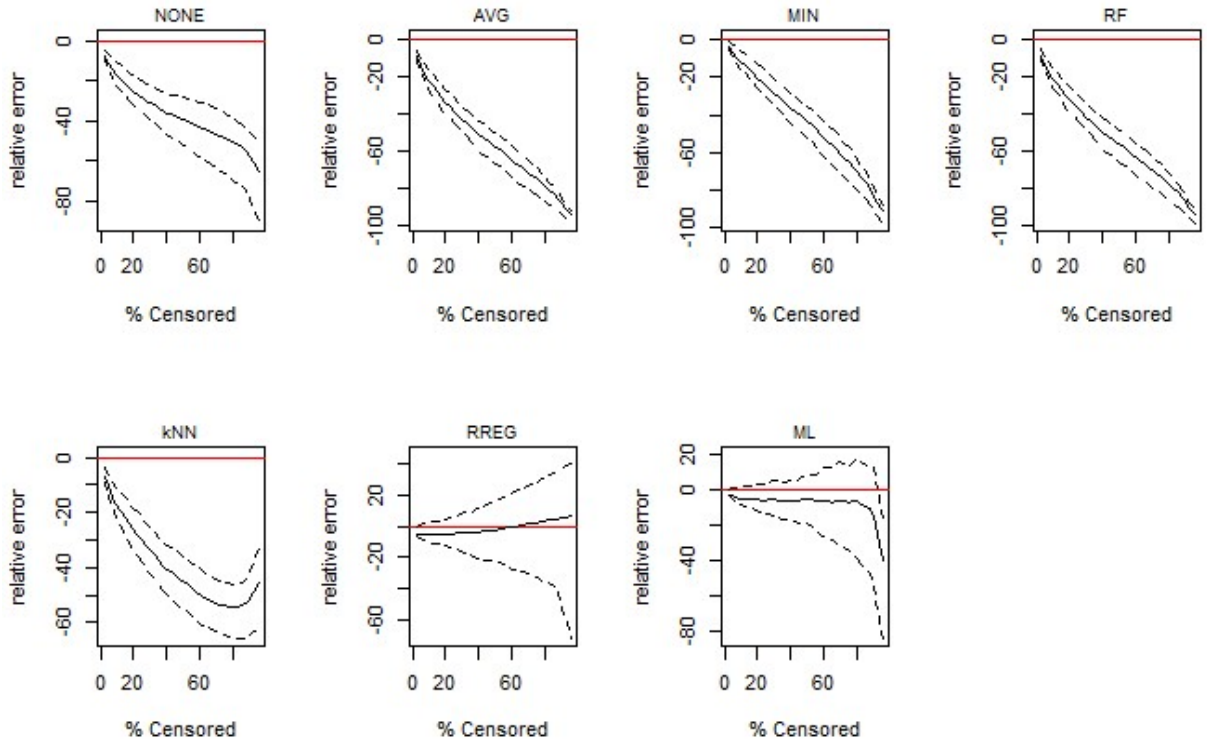


Figure 3.68: Trend plot for percent bias of SD parameter in Plasma. Trend lines are based on splinal fits. Solid line is the average while dotted are first and third quartiles.

likely to have a severe impact but a 50% reduction in the standard deviation would double transformation, implying metabolites derived from plasma, CSF or urine samples of human subjects can be reasonably regarded as log-normal.

Among data processing steps the log transformation is straightforward as it does not require estimating any tuning parameters, such as λ for Box-Cox or δ and α for the generalized log (Chapter 2.3). The log, therefore, is invariant to the removal or addition of any samples or ion-features. Relative comparisons between the two mean and standard deviation indicate that imputation is most impactful on the variance of the metabolites. This is highly relevant to z-scores, as a 5-10% change in the mean, which is about the worst seen in this parameter, isn't the magnitude of a metabolite's z-score.

Both kNN and random forest showed weak performance, particularly the latter which was virtually indistinguishable from average imputation in the real datasets. Yet both have previously been identified as strong performers [22, 64, 66, 105]. It is important to remember that individual datasets may have vastly different characteristics and the most effective method is likely to depend on the analysis of interest. As previously assessed datasets have focused on identifying group differences in the presence of missing values, those datasets have come with an artificial group structure. The datasets here consist of a single population, implying that kNN and random forest may be effective at picking up differences between groups of samples but weak at discriminating between individual values of a homogenous population. Meanwhile both maximum likelihood and rankit regression demonstrate near unbiased performance for the mean and standard deviation up to a missing rate of 70%. Consistency is another matter as the variability of these methods is roughly twice that of the traditional approaches, consequences of which are most pertinent at moderate to high levels of missing values. In the three metabolomic data sets, maximum likelihood estimates were found to be more stable in general and especially for missing proportions above 50%. Lower variation in ML estimates compared to RREG was also consistently observed in the computer simulations. Therefore, between the two ML is the preferred method.

Another advantage to the proposed methods is that they can easily be adapted to handle right censored values as well. Erroneous values could occur in the upper tail of the distribution with ion suppression, in which heavily abundant features are observed but with ion counts that do no reflected the true concentration [106]. In certain situations, it may be advisable to ignore values above a certain point. More generally, metabolomics data can be subject to extreme high outliers.

Intentionally censoring a certain amount of the highest values, say 1-5%, may be useful in large studies to avoid the impact of such outliers or extreme subjects.

A major untested assumption is that missing values are due to limit of detection. While this is a common assumption in the field, our search of the literature returned no papers investigating the veracity of this assumption. Further work is required to fully understand the sources of MVs, the prevalence of these sources and, ideally, the ability to identify the source of specific missing observations. However, when a missing value can reasonably be attributed to LOD, Gaussian models are an advisable approach to handling MVs in metabolomics.

REFERENCES

[1] Metabolon, "Metabolon and Baylor College of Medicine Partner to Launch Global Metabolomic-Assisted Pathway Screen (MAPS) for Inborn Errors of Metabolism," ed, 2014.

[2] M. J. Miller *et al.*, "Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism," *J Inherit Metab Dis,* Apr 2015.

[3] A. Kennedy *et al.*, "Utilizing Metabolomics of Human Urine to Screen for Multiple Inborn Errors of Metabolism," *Genetic Testing and Molecular Biomarkers,* vol. Under Review, 2016.

[4] S. H. Elsea *et al.*, "Elucidation of the complex metabolic profile of cerebrospinal fluid using an untargeted biochemical profiling assay," *Mol Genet Metab,* Apr 2017.

[5] T. R. Donti *et al.*, "Diagnosis of adenylosuccinate lyase deficiency by metabolomic profiling in plasma reveals a phenotypic spectrum," *Mol Genet Metab Rep,* vol. 8, pp. 61-6, Sep 2016.

[6] M. F. Wangler *et al.*, "A Metabolomic Map of Zellweger Spectrum Disorders Reveals Novel Disease Biomarkers," *Genetics in Medicine,* vol. in press, 2017.

[7] J. Tumas *et al.*, "Metabolomics in pancreatic cancer biomarkers research," *Med Oncol,* vol. 33, no. 12, p. 133, Dec 2016.

[8] L. Fan *et al.*, "Use of Plasma Metabolomics to Identify Diagnostic Biomarkers for Early Stage Epithelial Ovarian Cancer," *J Cancer,* vol. 7, no. 10, pp. 1265-72, 2016.

[9] F. Farshidfar *et al.*, "Serum metabolomic profile as a means to distinguish stage of colorectal cancer," *Genome Med,* vol. 4, no. 5, p. 42, May 2012.

[10] K. A. Lawton *et al.*, "Biochemical alterations associated with ALS," *Amyotroph Lateral Scler,* vol. 13, no. 1, pp. 110-8, Jan 2012.

[11] N. Greenberg, A. Grassano, M. Thambisetty, S. Lovestone, and C. Legido-Quigley, "A proposed metabolic strategy for monitoring disease progression in Alzheimer's disease," *Electrophoresis,* vol. 30, no. 7, pp. 1235-9, Apr 2009.

[12] L. M. de Lau, P. J. Koudstaal, A. Hofman, and M. M. Breteler, "Serum uric acid levels and the risk of Parkinson disease," *Ann Neurol,* vol. 58, no. 5, pp. 797-800, Nov 2005.

[13] E. Baraldi, S. Carraro, G. Giordano, F. Reniero, G. Perilongo, and F. Zacchello, "Metabolomics: moving towards personalized medicine," *Ital J Pediatr,* vol. 35, no. 1, p. 30, Oct 2009.

[14] J. van der Greef, T. Hankemeier, and R. N. McBurney, "Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials?," *Pharmacogenomics,* vol. 7, no. 7, pp. 1087-94, Oct 2006.

[15] Mayo Clinic Staff. (2015). *Personalized medicine and pharmacogenomics*. Available: https://www.mayoclinic.org/healthy-lifestyle/consumer-health/in-depth/personalized-medicine/art-20044300

[16] US Food and Drug Administration. (2017). *Precision Medicine*. Available: https://www.fda.gov/ScienceResearch/SpecialTopics/PrecisionMedicine/default.htm

[17] E. O. Lillie, B. Patay, J. Diamant, B. Issell, E. J. Topol, and N. J. Schork, "The n-of-1 clinical trial: the ultimate strategy for individualizing medicine?," *Per Med,* vol. 8, no. 2, pp. 161-173, Mar 2011.

[18] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: improving the biological information content of metabolomics data," *BMC Genomics,* vol. 7, p. 142, 2006.

[19] S. Bijlsma *et al.*, "Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation," *Anal Chem,* vol. 78, no. 2, pp. 567-74, Jan 2006.

[20] R. M. Salek *et al.*, "A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human," *Physiol Genomics,* vol. 29, no. 2, pp. 99-108, Apr 2007.

[21] P. Masson, K. Spagou, J. K. Nicholson, and E. J. Want, "Technical and biological variation in UPLC-MS-based untargeted metabolic profiling of liver extracts: application in an experimental toxicity study on galactosamine," *Anal Chem,* vol. 83, no. 3, pp. 1116-23, Feb 2011.

[22] P. S. Gromski *et al.*, "Influence of missing values substitutes on multivariate analysis of metabolomics data," *Metabolites,* vol. 4, no. 2, pp. 433-52, 2014.

[23] P. Aula, A. Jalanko, and L. Peltonen, "Aspartylglucosaminuria," in *OMMBID*, D. Valle, Ed., ed. New York: McGraw-Hill, 2001.

[24] Mayo Medical Laboratories, "Pipecolic Acid, Urine," in *Test Catalog*, M. F. f. M. E. a. Research, Ed., ed, 2017.

[25] M. R. Baumgartner *et al.*, "Atypical refsum disease with pipecolic acidemia and abnormal catalase distribution," *Ann Neurol,* vol. 47, no. 1, pp. 109-13, Jan 2000.

[26] National Center for Biotechnology Information, "S-Methylcysteine," in *PubChem Compound Database* vol. CID=24417, April 5th, 2018 ed, 2018.

[27] M. A. El-Magd *et al.*, "High doses of S-methylcysteine cause hypoxia-induced cardiomyocyte apoptosis accompanied by engulfment of mitochondaria by nucleus," *Biomed Pharmacother,* vol. 94, pp. 589-597, Oct 2017.

[28] T. Koal and H. P. Deigner, "Challenges in mass spectrometry based targeted metabolomics," *Curr Mol Med,* vol. 10, no. 2, pp. 216-26, Mar 2010.

[29] D. F. Conrad *et al.*, "Variation in genome-wide mutation rates within and between human families," *Nat Genet,* vol. 43, no. 7, pp. 712-4, Jun 2011.

[30] R. C. Pinto, "Chemometrics Methods and Strategies in Metabolomics," *Adv Exp Med Biol,* vol. 965, pp. 163-190, 2017.

[31] L. Yi *et al.*, "Chemometric methods in data processing of mass spectrometry-based metabolomics: A review," *Anal Chim Acta,* vol. 914, pp. 17-34, Mar 2016.

[32] R. Madsen, T. Lundstedt, and J. Trygg, "Chemometrics in metabolomics--a review in human disease diagnosis," *Anal Chim Acta,* vol. 659, no. 1-2, pp. 23-33, Feb 2010.

[33] P. Hopke, "The evolution of chemometrics," *Analytica Chimica Acta,* vol. 500, no. 1-2, pp. 365-377, 2003.

[34] I. E. Frank and J. H. Friedman, "A Statistical View of Some Chemometrics Regression Tools," *Technometrics,* vol. 35, no. 2, pp. 109-135, 1993.

[35] Mayo Clinic. (2018). *Prediabetes*. Available: https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284

[36] Joslin Diabetes Center. (2018). *Diagnosing Impaired Glucose Tolerance (IGT)*. Available: http://www.joslin.org/info/diagnosing_impaired_glucose_tolerance_IGT.html

[37] J. Cobb *et al.*, "A novel test for IGT utilizing metabolite markers of glucose tolerance," *J Diabetes Sci Technol,* vol. 9, no. 1, pp. 69-76, Jan 2015.

[38] S. A. Hills *et al.*, "The EGIR-RISC STUDY (The European group for the study of insulin resistance: relationship between insulin sensitivity and cardiovascular disease risk): I. Methodology and objectives," *Diabetologia,* vol. 47, no. 3, pp. 566-570, Mar 2004.

[39] M. Sinnott *et al.*, "Fasting plasma glucose as initial screening for diabetes and prediabetes in irish adults: The Diabetes Mellitus and Vascular health initiative (DMVhi)," *PLoS One,* vol. 10, no. 4, p. e0122704, 2015.

[40] R. Fastenau, "Metabolon to Develop accuGFR™ Kidney Function Test with Johns Hopkins and Tufts Medical Center," ed: Metabolon, 2016.

[41] G. Anton *et al.*, "Pre-analytical sample quality: metabolite ratios as an intrinsic marker for prolonged room temperature exposure of serum samples," *PLoS One,* vol. 10, no. 3, p. e0121495, 2015.

[42] D. S. Wishart *et al.*, "HMDB 4.0: the human metabolome database for 2018," *Nucleic Acids Res,* vol. 46, no. D1, pp. D608-D617, Jan 2018.

[43] S. G. Wannamethee, A. G. Shaper, and I. J. Perry, "Serum creatinine concentration and risk of cardiovascular disease: a possible marker for increased risk of stroke," *Stroke,* vol. 28, no. 3, pp. 557-63, Mar 1997.

[44] MedlinePlus, "Tryptophan," ed, 2018.

[45] DietaryFiberFood.com. (2012). *Tryptophan: food sources high in tryptophan*. Available: https://www.dietaryfiberfood.com/amino-acids/tryptophan-sources.php

[46] S. Steinberg, G. V. Raymond, N. E. Braverman, and A. B. Moser, "Zellweger Spectrum Disorder," in *GeneReviews*, M. P. Adam, Ed., ed. Seattle, WA: University of Washington, Seattle, 2003.

[47] S. M. Gospe, "Pyridoxine-Dependent Epilepsy," in *GeneReviews*, M. P. Adam, Ed., ed. Seattle, WA: University of Washington, Seattle, 2001.

[48] M. C. Laboratories, "Pipecolic Acid, Serum," ed, 2018.

[49] D. A. Applegarth, J. R. Toone, and R. B. Lowry, "Incidence of inborn errors of metabolism in British Columbia, 1969-1996," *Pediatrics,* vol. 105, no. 1, p. e10, Jan 2000.

[50] S. Sanderson, A. Green, M. A. Preece, and H. Burton, "The incidence of inherited metabolic disorders in the West Midlands, UK," *Arch Dis Child,* vol. 91, no. 11, pp. 896-9, Nov 2006.

[51] H. Moammar, G. Cheriyan, R. Mathew, and N. Al-Sannaa, "Incidence and patterns of inborn errors of metabolism in the Eastern Province of Saudi Arabia, 1983-2008," *Ann Saudi Med,* vol. 30, no. 4, pp. 271-7, 2010 Jul-Aug 2010.

[52] S. A. Latheef, "A database for inborn errors of metabolism in the Indian state of Andhra Pradesh," *Bioinformation,* vol. 4, no. 7, pp. 276-7, Jan 2010.

[53] J. Barwell and E. Boskey. (2016). *Alkaptonuria*. Available: https://www.healthline.com/health/alkaptonuria

[54] National Organization for Rare Disorders. (2005). *Medium Chain Acyl CoA Dehydrogenase Deficiency*. Available: https://rarediseases.org/rare-diseases/medium-chain-acyl-coa-dehydrogenase-deficiency/

[55] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, 2nd ed. (Wiley series in probability and statistics). Hoboken, N.J.: Wiley, 2002, pp. xv, 381 p.

[56] J. L. Schafer, *Analysis of incomplete multivariate data*, 1st ed. London; New York: Chapman & Hall, 1997, p. 430.

[57] G. Molenberghs and M. G. Kenward, *Missing data in clinical studies*. Chichester: Wiley, 2007.

[58] D. W. Hosmer and S. Lemeshow, *Applied survival analysis : regression modeling of time to event data* (Wiley series in probability and statistics Texts and references section). New York: Wiley, 1999, pp. xiii, 386 p.

[59] US Food and Drug Administration. (2017). *Laboratory Developed Tests*.

[60] US Food and Drug Administration. (2014). *Clinical Laboratory Improvement Amendments (CLIA)*.

[61] Centers for Disease Control and Prevention. (2017). *Clinical Laboratory Improvement Amendments (CLIA)*. Available: https://wwwn.cdc.gov/clia/Regulatory/

[62] College of American Pathologists. (2016). *College of American Pathologists*. Available: www.cap.org

[63] J. S. Shah, G. N. Brock, and S. N. Rai, "Metabolomics data analysis and missing value issues with application to infarcted mouse hearts," *BMC Bioinformatics,* vol. 16, no. 15, p. 16, 2015.

[64] E. G. Armitage, J. Godzien, V. Alonso-Herranz, Á. López-Gonzálvez, and C. Barbas, "Missing value imputation strategies for metabolomics data," *Electrophoresis,* vol. 36, no. 24, pp. 3050-60, Dec 2015.

[65] R. Wei *et al.*, "Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data," *Sci Rep,* vol. 8, no. 1, p. 663, Jan 2018.

[66] O. Hrydziuszko and M. R. Viant, "Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline," *Metabolomics,* vol. 8, no. 1, pp. 161-174, 2012.

[67] C. L. Parr, A. Hjartåker, I. Scheel, E. Lund, P. Laake, and M. B. Veierød, "Comparing methods for handling missing values in food-frequency questionnaires and proposing k nearest neighbours imputation: effects on dietary intake in the Norwegian Women and Cancer study (NOWAC)," *Cambridge University Press,* vol. 11, no. 4, pp. 361-370, 2008.

[68] M. Steinfath, D. Groth, J. Lisec, and J. Selbig, "Metabolite profile analysis: from raw data to regression and classification," *Physiol Plant,* vol. 132, no. 2, pp. 150-61, Feb 2008.

[69] D. Albrecht, O. Kniemeyer, A. A. Brakhage, and R. Guthke, "Missing values in gel-based proteomics," *Proteomics,* vol. 10, no. 6, pp. 1202-11, Mar 2010.

[70] J. Xia, N. Psychogios, N. Young, and D. S. Wishart, "MetaboAnalyst: a web server for metabolomic data analysis and interpretation," *Nucleic Acids Res,* vol. 37, no. Web Server issue, pp. W652-60, Jul 2009.

[71] N. L. Kuehnbaum, J. B. Gillen, M. J. Gibala, and P. Britz-McKibbin, "Personalized metabolomics for predicting glucose tolerance changes in sedentary women after high-intensity interval training," *Sci Rep,* vol. 4, p. 6166, 2014.

[72] J. J. Gooley, "Applications of Circadian Metabolomics," *Current Metabolomics,* vol. 2, no. 1, pp. 2-14, 2014.

[73] R. Dallmann, A. U. Viola, L. Tarokh, C. Cajochen, and S. A. Brown, "The human circadian metabolome," *Proc Natl Acad Sci U S A,* vol. 109, no. 7, pp. 2625-9, Feb 2012.

[74] Y. C. Yuan, "Multiple Imputation for Missing Values: Concepts and New Developments," in *SUGI*, Indianapolis, Indiana, 2000.

[75] 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.

[76] S. van Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Stat Methods Med Res,* vol. 16, no. 3, pp. 219-42, Jun 2007.

[77] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A multivariate technique for multiply imputing
missing values using a sequence of regression models," *Survey Methodology,* vol. 27, no. 1, pp. 85-95, 2001.

[78] R. Steuer, K. Morgenthal, W. Weckwerth, and J. Selbig, "A gentle guide to the analysis of metabolomic data," *Methods Mol Biol,* vol. 358, pp. 105-26, 2007.

[79] H. J. Issaq, Q. N. Van, T. J. Waybright, G. M. Muschik, and T. D. Veenstra, "Analytical and statistical approaches to metabolomics research," *J Sep Sci,* vol. 32, no. 13, pp. 2183-99, Jul 2009.

[80] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrom Rev,* vol. 26, no. 1, pp. 51-78, 2007 Jan-Feb 2007.

[81] K. H. Liland, "Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis," *TrAC, Trends in analytical chemistry (Regular ed.),* vol. 30, no. 6, pp. 827-841, 2011.

[82] K. Pearson, "On Lines and Planes of Closest Fit to Systems of
Points in Space," *Philosophical Magazine,* vol. 2, no. 7-12, pp. 559-572, 1901.

[83] H. Hotelling, *Analysis of a complex of statistical variables into principal components*. Warwick & York Inc., 1933.

[84] A. Christoffersson, "The one component model with incomplete data," Uppsala University, 1970.

[85] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Comput,* vol. 11, no. 2, pp. 443-82, Feb 1999.

[86] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics,* vol. 19, no. 16, pp. 2088-96, Nov 2003.

[87] M. Scholz, F. Kaplan, C. L. Guy, J. Kopka, and J. Selbig, "Non-linear PCA: a missing data approach," *Bioinformatics,* vol. 21, no. 20, pp. 3887-95, Oct 2005.

[88] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics,* vol. 17, no. 6, pp. 520-5, Jun 2001.

[89] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[90] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News,* vol. 2, no. 3, pp. 18-22, 2002.

[91] R. Wei, J. Wang, M. Su, E. Jia, T. Chen, and Y. Ni, "Missing Value Imputation Approach for Mass Spectrometer-based Metabolomics Data," *bioRxiv,* 2017.

[92] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R. C. Team, "Linear and Nonlinear Mixed Effects Models," R package version 3.1-131 ed, 2017.

[93] A. C. Cohen, "Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples," pp. 557-569, 1950/12 1950.

[94] M. Halperin, "Maximum Likelihood Estimation in Truncated Samples," *The Annals of Mathematical Statistics,* vol. 23, no. 2, pp. 226-238, 1952.

[95] A. C. Cohen, "Simplified Estimators for the Normal Distribution When Samples Are Singly Censored or Truncated," *Technometrics,* vol. 1, no. 3, pp. 217-237, 1959.

[96] E. T. Whittaker and G. Robinson, *The calculus of observations*, 2d ed. London and Glasgow,: Blackie & son limited, 1929, pp. xvi, 395 p.

[97] T. P. Royston, "An Extension of Shapiro and Wilk W Test for Normality to Large Samples," *Journal of the Royal Statistical Society. Series C (Applied Statistics),* vol. 31, no. 2, pp. 115-124, 1982.

[98] P. Royston, "Approximating the Shapiro-Wilk W-test for non-normality," *Statistics and Computing,* vol. 2, no. 3, pp. 117-119, 1992/09/01 1992.

[99] G. Blom, *Statistical estimates and transformed beta-variables*. New York,: Wiley, 1958, p. 176 p.

[100] M. M. W. B. Hendriks, F. A. v. Eeuwijk, R. H. Jellema, and J. A. Westerhuis, "Data-processing strategies for metabolomics studies," *TrAC, Trends in analytical chemistry (Regular ed.),* vol. 30, no. 10, pp. 1685-1698, 2011.

[101] M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart, and O. Yanes, "A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data," *Metabolites,* vol. 2, no. 4, pp. 775-95, 2012.

[102] B. M. Warrack *et al.*, "Normalization strategies for metabonomic analysis of urine samples," *J Chromatogr B Analyt Technol Biomed Life Sci,* vol. 877, no. 5-6, pp. 547-52, Feb 2009.

[103] D. Ryan, K. Robards, P. D. Prenzler, and M. Kendall, "Recent and potential developments in the analysis of urine: a review," *Anal Chim Acta,* vol. 684, no. 1-2, pp. 8-20, Jan 2011.

[104] A. J. Chetwynd, A. Abdul-Sada, S. G. Holt, and E. M. Hill, "Use of a pre-analysis osmolality normalisation method to correct for variable urine concentrations and for improved metabolomic analyses," *J Chromatogr A,* vol. 1431, pp. 103-110, Jan 2016.

[105] R. Di Guida *et al.*, "Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling," *Metabolomics,* vol. 12, p. 93, 2016 2016.

[106] A. Furey, M. Moriarty, V. Bane, B. Kinsella, and M. Lehane, "Ion suppression; a critical review on causes, evaluation, prevention and applications," *Talanta,* vol. 115, pp. 104-22, Oct 2013.

CHAPTER 4: MERGING METABOLOMIC DATASETS

4.1. Introduction

The LC-MS workflow requires periodic maintenance that requires turning off the machine. These periods in which the machine is actively processing samples is referred to here as a "runs" or "batches". Instrument performance can vary from run to run. A major obstacle in global LC-MS based metabolomics is drawing comparisons between samples processed across different instrument runs or different instruments. There are several reasons for wanting to compare samples from different instrument runs or different instruments. Before being analyzed on a mass spectrometer, samples are prepared and a portion is then transferred to an analysis well in a multi-well plate. The number of wells available depends on the type and size of wells used but is generally some multiple of 6 [1]. Even instrumentation that can accommodate large plates or multiple small plates are generally restricted to at most a few hundred wells [2]. Large epidemiologic studies with thousands of samples can easily exceed this capacity. In another example, a longitudinal clinical experiment or cohort study may not have all samples available at one time for analysis. In particular, the clinical environment is analogous to these situations as new patients are regularly being admitted and evaluated.

Mass spectrometry itself is inherently semi-quantitative, with the observed variable returned by the instrument being the number of ion counts associated with the feature, i.e. "ion peak", which depend not just on the concentration in the sample but also biochemical and instrument characteristics. For a given weight, a compound with higher molar mass will have fewer

molecules than a compound with a lower molar mass. Fewer molecules implies fewer ions, but the ionization efficiency, which relates the number of ions formed to the molecules available, varies by compound as well as with instrument performance. Hence the number of ions measured relates to the amount in the sample but is not sufficient to determine concentration. Instead, biochemical and instrument calibration are fixed for a given instrument run, which allows for inference of relative concentrations for samples within the run. One sample, for example, may have 30% more ions than another sample for a given biochemical. While the exact amount of the biochemical cannot be determined, assuming the sample types are the same and the concentrations are within the linear range of the instrument, one can infer that whatever the concentrations may be the first sample is 30% greater than the second. An implicit assumption is that ion recovery is linear over the concentrations being measured, which is explained in the following paragraph.

Exact concentrations are derived though calibration curves, aka standard curves, in which known concentrations of the target compound are included to provide a reference point for the ion counts and calculate the levels in samples of interest according to their position on the curve. For a thorough review of calibration curves see the five-part series by Dolan [3-7], but calibration curves are most conveniently treated as a linear function. In a dataset consisting of m features and n samples, let c_{ji} be the true concentration of the i^{th} subject and j^{th} analyte. For y_{ji} the ion intensity associated with the same sample and ion, the assumption under the linear calibration is that

$$y_{ji} = \beta_{0j} + \beta_{1j} c_{ji} + \varepsilon_{ji}$$

which is a simple linear regression model. The parameters β_{0j} and β_{1j}, which vary by analyte, are estimated via a set of c_{ji}'s with known concentration. β_{1j} is strictly greater than 0 while the

intercept β_{0j} is largely a nuisance parameter and generally close to 0 for any concentration range that extends well above the limit of detection of the instrument. In fact, single point calibration curves, in which all $c_{ji}'s$ have the same concentration, remove β_{0j} entirely and can perform well when the concentration level is chosen appropriately [8]. Multi-point calibration curves involve multiple concentration levels and are better suited for examining the linear assumption. When results suggest non-linearity, polynomial models or other non-linear function may be used [9]. Alternatively, analysis may be restricted to limits of quantitation in which the assay is found to be linear. Producing a sample of known concentration obviously requires being able to measure the exact quantity of target metabolite. This will usually require being able to obtain a purified amount of the metabolite to measure that, because of chemical reactions that can occur in solution, occupies a well. Allocating an entire well to a single compound is clearly infeasible in an untargeted analysis as (1) the metabolites to be captured may not all be known a-priori and (2) the number of metabolites totals hundreds or even thousands and easily exceeds the available sample wells per instrument run. Lacking full quantitation, one must find some way to relate the ion counts in different batches to each other.

This problem is not unique to metabolomics as similar LC-MS instrumentation is used in proteomics as well. Normalization is popular in proteomics and represents a potential solution for metabolomic data [10-12]. Normalization attempts to correct each individual sample for systematic effects due to collection, processing and instrumentation. In theory such correction would adjust for instrument effects as well, although metabolites span a much wider range of physical properties and chemical classes than seen in proteomics, leaving the effectiveness of normalization in global MS metabolomics unknown.

An alternative to normalization is to include related samples, referred to as *anchoring*, across

multiple instrument runs by which metabolite specific adjustment can be performed. Doing so provides a reference point with which to orient the batches to one another. Similar approaches have been shown to perform well against standard normalization schemes in GC-MS data [13].

Through a mix of targeted and untargeted data, this chapter examines the concept in LC-MS data. The results show that addressing metabolites individually generates global data that is more consistent with targeted concentration and less variable over the whole metabolome versus typical normalizations which tend to ignore the breadth of physical properties spanned by global LC-MS metabolomics. Furthermore, the number of anchors required to perform this approach is investigated and shown to be minimal for the vast majority of metabolites and is therefore not burdensome to instrument time and resources.

4.2. Traditional Normalization

Normalization is a commonly used method to remove a significant portion of sample to sample variation from the data. This purpose distinguishes normalizations from transformations, which are used to manipulate quantitative characteristics of metabolite features. In a practical manner, in a metabolomics set where the samples are the rows and the features are the columns, normalization acts on the rows while transformation acts on the columns. The most common normalization is total ion count (or total ion current) normalization (TIC) in which all metabolites in a sample are divided by the total number of ions observed in the sample [12]. Although commonly used, TIC is susceptible to being overly influenced by a small number of features with very large ion counts. Various adjustments to this basic premise include median normalization, MS-total useful signal (MSTUS) [14], median absolute deviation [15], probabilistic quotient normalization (PQN) [16] and cyclic locally weighted regression (Cyclic

Lowess)[17] among others [18, 19]. However, most of these normalizations are built on the assumption that on "average" the ion count of each sample should be more or less equal.

In this paper normalizations are separated into three classes depending on the mechanism of action. The first class involves dividing ion intensities by a function of the sample's ion spectra. The second class of normalization relies on Minus-Average (MA) plots. The third class are those normalizers that do not fit into either of the first two classes.

4.2.1. Class I – Spectral Division

Normalizers of the first class are defined as the ratio of the sample's raw intensity values and a function of the sample spectra. That is to say, letting $\boldsymbol{y}_j = \{ y_{j1}, \ldots, y_{jm} \}$ be the vector of observed ion counts and $\boldsymbol{y}_j^N$ the resulting normalized vector, the relationship is such that:

$$\boldsymbol{y}_j^N = \left\{ \frac{y_{j1}}{f_j(\boldsymbol{y}_j)}, \ldots, \frac{y_{jm}}{f_j(\boldsymbol{y}_j)} \right\}'$$

where $f_j(\bullet)$ is some function. Table 4.1 summarizes f_j for the first class of normalizations.

Several of these methods are variations on TIC, such as MSTUS which restricts the summed signal to only those features that are common to all samples. Vector Normalization takes TIC into two dimensions by measuring the Euclidean distance of the observed vector from the origin **0**, and for this reason is sometimes referred to as Euclidean Norm. Both TIC and Vector Normalization are specific versions of the more general form $\sqrt[p]{\Sigma\, y_{ji}^p}$. Mean is simply TIC adjusted for the number of features. Median Absolute Deviation (MAD) takes median a step further by finding the absolute deviations from the median within a sample and using the median of these to normalize. Some Spectral Division normalization methods use a baseline or control spectrum correction. Such spectra can be determined a-priori or chosen from the available

Table 4.1: Class I Normalizers

TIC	$f_i = \sum_{j=1}^{m} x_{ij}$
MSTUS	$f_i = \sum_{A} x_{ij}$ $A = \{k\}\ such\ that\ x_{ik}\ not\ missing\ \forall\ i \in \{1, \dots, n\}$
Vector	$f_i = \left(\sum_{j=1}^{m} x_{ij}^2\right)^{1/2}$
Mean	$f_i = \sum_{j=1}^{m} \frac{x_{ij}}{n}$
Median	$f_i = median(Y_i)$
MAD	$f_i = median(\lvert Y_i - median(Y_i)\rvert)$
LB[a]	$f_i = {median(Y_i)}\Big/{median(Y_{Baseline})}$
PQN[b]	$f_i = median\{q_{i1} \quad \dots \quad q_{im}\}$ $q_{ij} = {x_{ij}^{TIC}}\Big/{x_{control,j}^{TIC}}$

[a,b] Baseline / Control spectrum may be taken from a designated sample or calculated from available data, such as sample with median TIC.

samples such as the sample with the median TIC. Linear Baseline scaling (LB) and PQN are examples of this. In LB, each sample is normalized so that the TIC of the resulting normalized sample is equal to that of the "baseline". LB assumes a constant linear relationship between the sample and the baseline. Non-linear extensions are available. Although the name includes "scaling", the intent is consistent with normalization which seeks to adjust all spectrum of each sample to the same level in some sense and the computation is consistent with the Class I definition. PQN, which involves a four-step process, is the most computation intensive of Class I normalizers listed here. In the first step TIC normalization is performed. Second, a control spectrum is calculated – this may be based upon a designated sample or the median spectra from

all samples may be used. Third, for each feature the ratio, i.e. quotient, of the TIC normalized intensity of the sample and control spectrum is found. The final normalizer is then the median of all quotients. Most class I normalizers are straightforward to calculate and are not computationally time intensive. Hence, these are popular and common choices for normalizing.

4.2.2. Class II – Minus Average (MA) normalizers

The second class of normalizers involve the plot of the minus versus average, which are essentially Altman-Bland plots on the log scale [20, 21]. For any two samples j and j' the MA are the scatter plot where each feature has coordinates $\left(minus_{ji}, avg_{ji}\right)$ given by:

$$minus_{ji} = \log_2\left(y_{ji}\right) - \log_2\left(y_{j'i}\right)$$

$$avg_{ji} = \frac{\log_2\left(y_{ji}\right) + \log_2\left(y_{j'i}\right)}{2}$$

The minus (M) can be viewed as being the log of the ratio while average (A) is the log of the product. Orienting the two spectra in this way is intended to magnify systemic effects, both linear and non-linear. From this plot a curve can be fit to the data to determine normalized values for the two samples. Cyclic Locally Weighted Regression (Cyclic Lowess) and Contrast Normalization [22] (CN) are normalizers of this type. Under Cyclic Lowess, a non-linear local regression curve (lowess) is fit to the MA plot for a given pair of samples. The process is then repeated for all possible pairwise combinations of samples in the dataset. Following a complete iteration over all samples, a model tolerance parameter is calculated and the cycle is repeated until some tolerance is achieved between the latest cycle and preceding one. The complete set of all ion features for all samples can be expressed as

$$\boldsymbol{Y}_{\mathrm{n\times m}} = \begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_n} \end{bmatrix}$$

in which is $\boldsymbol{y_j}=\{y_{1i}, \ldots, y_{mi}\}$ is the vector of ion intensities for the j^{th} subject. Under CN, $\boldsymbol{Y}_{n\times m}$ is log transformed such that

$$\log(\boldsymbol{Y}) = \begin{bmatrix} \log(\boldsymbol{y_1}) \\ \log(\boldsymbol{y_2}) \\ \vdots \\ \log(\boldsymbol{y_n}) \end{bmatrix}$$

with $\log(\boldsymbol{y_j})=\{\log(y_{j1}), \ldots, \log(y_{jm})\}$ then linearly transformed using a m by m orthonormal matrix $\boldsymbol{M}$ to produce a new set of orthogonal vectors:

$$\boldsymbol{Y^O} = \log(\boldsymbol{Y})\,\boldsymbol{M}.$$

The first row of $\boldsymbol{M}$ is the repetition of the constant $1/\sqrt{m}$. The other rows of $\boldsymbol{M}$ are not uniquely defined, except in the case of $m = 2$ which gives

$$\boldsymbol{M_2} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

For $m > 2$, $\boldsymbol{M}$ is not unique which requires some consideration in the next step in which $\boldsymbol{y_1^0}$, the first row of $\boldsymbol{Y^O}$, is used to predict the remaining rows $\boldsymbol{y_j^0}$ for $j = 2, \ldots, n$. Referring to these predictions as $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{0}}$, using loess regression with weighted least squares produces $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{0}}$'s that are invariant to the choice of $\boldsymbol{M}$. Estimation of $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{0}}\boldsymbol{'s}$ is iterated until some tolerance between the previous and newest estimate is achieved. The final normalized matrix is then given by

$$\boldsymbol{Y^N} = \begin{bmatrix} \boldsymbol{y_1^0} \\ \boldsymbol{y_2^0} - \hat{\boldsymbol{y}}_{\boldsymbol{2}}^{\boldsymbol{0}} \\ \vdots \\ \boldsymbol{y_n^0} - \hat{\boldsymbol{y}}_{\boldsymbol{n}}^{\boldsymbol{0}} \end{bmatrix}.$$

From this point the dataset may be analyzed or mapped back to the original space by taking the anti-log transformation of each element of $\boldsymbol{Y^N M}$. The similarity to cyclic lowess may not be immediately obvious; however, notice when $m = 2$ the contrast matrix $\boldsymbol{M}$ coupled with the log

transformation is analogous to the orientation of the MA plot. Contrast normalization essentially generalized the MA concept to higher dimensions. From here, both methods utilize lowess to smooth out the normalized values.

4.2.3. Class III – Other

Normalizations that don't fit the criteria of Class I or Class II are considered. One example of this is Quantile Normalization (Quant) which rescales the dataset so that the distribution of intensities within each sample is the same across all samples. Let $\boldsymbol{y}_{j,\,ord}$ be the ordered set of intensities for sample j:

$$\boldsymbol{y}_{\boldsymbol{j},\,ord} = \left\{ y_{j[1]}\,,\ \ldots,\ \ y_{j[m]} \right\}.$$

Consider next the vector of average ordered statistics across all $\boldsymbol{y}_{\boldsymbol{j},\,ord}$'s:

$$\bar{\boldsymbol{y}}_{\boldsymbol{ord}} = \left\{ \bar{y}_{[1]},\ \ldots, \bar{y}_{[m]} \right\} = \left\{ \sum_{\mathrm{j}=1}^{\mathrm{n}} \frac{y_{j[1]}}{n},\ \ldots,\ \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{y_{j[m]}}{n} \right\}$$

This process essentially orders the intensities from lowest to highest within each row of the dataset and then takes the average of each rank. The normalized value of any given sample and feature is then

$$\boldsymbol{y}_{\boldsymbol{ji}}^{\boldsymbol{N}} = \bar{y}_{[q]}$$

where $q \in \{1, \ldots, m\}$ is the rank of y_{ji} among all the intensities for sample j. Notice that the set of normalized values is the same for all j. Standardizing the spectra is one advantage to Quant in that it directly puts the intensities of each sample on the same scale, making sample-to-sample comparisons easier. A significant drawback is that features with missing values must be removed or imputed. Second, biochemicals that are significantly more abundant than all other features, such as oleate in plasma or creatinine in urine, may be normalized to a near static state. The same

problem would occur in metabolites that are significantly lower in abundance than all other metabolites; however, this case is uncommon in untargeted metabolomics because a significant proportion of features have intensities approaching the instrument's LOD.

4.3. Proposed Alternatives for Instrument Run

This chapter is concerned with alternatives to adjusting for instrument runs that involve correcting each metabolite separately. The methodology for such adjustment is accomplished through the use of common samples included in each instrument run, or *batch*, being combined. Samples in the same batch are normalized against these common samples, called *anchors*. Computationally, this anchor normalization approach is similar to performing LB but with normalization taking place within each batch and acting on the biochemicals rather than on the samples. As a result, the proposed method is more accurately classified as scaling rather than normalization since the manipulation is performed on each feature (i.e. column) in the data independently disregarding any sense of overall spectral equivalence. The intent is not to induce any sort of equivalence between the features, such as scaling each metabolite against its own mean to put each feature on the same scale or performing log / GLOG transformation to induce constant variance. Instead, the purpose of this anchoring approach is to account for systematic changes in the samples due to instrument run. The term "normalization" is used for the application of the following proposed methods to make clear that the intent of these methods is to adjust for instrument variability which has traditionally been the role of normalizations.

The main challenge to this anchoring approach is finding a baseline between samples across different batches. Solutions to such a reference point are considered here in two approaches. The first approach, referred to as *anchor normalization*, uses pooled technical replicates which are incorporated each time a new batch of samples is analyzed. The second approach, here referred

to as *batch normalization*, involves balancing the experimental design equally across all instrument runs. It should be pointed out that batch normalization described here is not the same as Batch Normalizer described by Wang *et al*. [27], which utilizes the total ion count in its normalization of the dataset. Both approaches are described in the subsequent sections. It is useful to emphasize once more, that in either case the adjustment is taking place on the metabolites (i.e. columns) of the data rather than on the samples (i.e. rows), as is the case with typical normalizations.

4.3.1. Anchor Normalization

Returning to the linear calibration model, consider two separate instrument runs in which k technical replicates of the same sample are run in both batches. The ion intensity of the i^{th} metabolite of batch $b \in \{1, 2\}$ for any replicate $j \in \{1, \ldots, k\}$ is given as

$$y_{jib} = \beta_{ib} x_i e^{\eta_{jib}} + \varepsilon_{jib}.$$

The subscript of the sample concentration, x_i, is dependent only on the biochemical since these k samples are technical replicates and thus have the same concentration. Recall that $\eta_{ib} \sim N(0, \sigma^2_{\eta ib})$ and $\varepsilon_{ib} \sim N(0, \sigma^2_{\varepsilon ib})$ so that both terms depend on the batch and metabolite, but not the individual replicates. The expected value of any such replicate is then:

$$\mu_{ib} = \mathrm{E}\left[y_{jib}\right] = \beta_{ib} x_i e^{\sigma^2_{\varepsilon ib}/2}.$$

Rearranging these terms gives

$$\mu_{\mathrm{ib}} = \mathrm{E}\left[y_{jib}\right] = \left(\beta_{ib} e^{\sigma^2_{\varepsilon ib}/2}\right) \mathrm{x_i}.$$

As both the IE factor β_{ib} and the term resulting from log-normal error component $e^{\sigma^2_{\varepsilon ib}/2}$ are both fixed, but unknown, parameters depending only on the batch and metabolite, they can be

combined into a single parameter term. Letting $\beta^*_{ib} = \beta_{ib} e^{\sigma^2_{\varepsilon ib}/2}$ it is easy to see that mean ion count for the batch is proportional to true concentration level:

$$\mu_{ib} = \beta^*_{ib} x_i$$

Hence, the mean ion count for the two batches are proportional:

$$\frac{\mu_{i1}}{\beta^*_{i1}} = \frac{\mu_{i2}}{\beta^*_{i2}}$$

By the central limit theorem, there exists a k such that average of the replicates within a batch

$$\bar{y}_{ib} = \sum_{j=1}^{k} \frac{\beta_{ib} x_i e^{\eta_{jib} + \varepsilon_{jib}}}{k}$$

is reasonably close to $\beta^*_{ib} x_i$.

Consider experimental samples analyzed on two independent instruments or over two batches of the same instrument. Assuming a sufficient number of replicates are used, dividing each batch by the average of the batch's technical replicates will anchor the samples to a common scale, namely the relative concentration to the replicate. Hence the term *anchor samples*.

Anchor normalization is appropriate for batches with small number of experimental samples, when the experimental design is unknown, or when the experimental design prevents all sample types of interest from being available at the same time, such as a time course or longitudinal study. Of course, finding a source of material to use for the inter-batch technical replicates can be non-trivial. QC samples, which are included as part of many metabolomic workflows to monitor instrument performance [23, 24, 25, 26], are convenient sources of material when available to serve as anchor samples. Another option would be to pool a small amount of each sample in the first batch, and then inserting aliquots of these into each batch to be combined. Aliquoting into each batch requires a-priori knowledge of the situation and requires significantly more material

depending on the number of batches forecasted, however.

4.3.2. Batch Normalization

Consider a large number of samples drawn from the same experimental design and randomly assigned to separate instrument runs. The ion intensity of the i^{th} metabolite of batch $b \in \{1, 2\}$ for any replicate $j \in \{1, \ldots, n_b\}$ is given as

$$y_{jib} = \beta_{ib} x_{jib} e^{\eta_{jib}} + \varepsilon_{jib}$$

with n_b the number of experimental samples in the batch. In this situation the concentration x_{jib} depends on the metabolite, sample and batch, since each sample in the experiment represents a different individual being tested. However, since each batch is composed of the same experimental design and contains a large, randomly selected subset of samples, it follows from the Law of Large Numbers that the average concentration of the i^{th} metabolite should be roughly the same and equal to the mean concentration of the metabolite μ_{xi}. Hence, a similar technique can be employed as before for the expected average ion count μ_{yi} in terms of the mean batch concentration μ_{xi}:

$$\mu_{yi} = E(\bar{y}_{jib}) = \beta_{ib} e^{\sigma^2_{\varepsilon ib}/2} \bar{x}_{jib} = \beta^*_{ib} \mu_{xi}.$$

This leads to a similar result as in anchor normalization. In a sense, the use of technical replicates has been replaced with experimental samples at the cost of requiring for larger sample sizes and identical batch designs, which are required to connect the two ends in above equation.

In many settings, batch normalization may be more feasible than obtaining enough material for technical replicates to serve as the anchor samples for anchor normalization. Additionally, analysis of anchor samples increases the number of samples per batch, increasing the cost and time required for large experiments. Thus, batch normalization can be considered appropriate for large experiments that replicate the study design across multiple batches. Therefore, this

approach may be preferable as being more cost effective. And while it does require large numbers of samples, the concept can be extended to groups of samples within the batch. For example, suppose an experiment of control and stage I disease subjects are run on the first instrument batch. After some time, a second batch comparing the same type of control subjects to stage II disease subjects. Although the entire experimental design differs between the two runs, one could conceivable normalize the two sets based on the common control group. It is important to emphasize that when saying the controls are the same "type", this requires more than simply being from the same population. It involves all facets of analysis, including sample collection, sample storage, additional preanalytical methodology, and instrumentation. Replicating so many stages of the metabolomics workflow could be impractical in many situations. Therefore, when all samples of a large study are not available for analysis at once and the entire design cannot be replicated at a later date or is too detailed to , the anchor normalization approach should be considered.

4.4. Methods

The goal of this chapter is to compare anchor (ANCH) and batch (BAT) normalization to standard -omic normalizations that might be considered for a metabolomic dataset. The available data, which is described in the following section, contains technical replicates of pooled plasma included in each batch which serve as the anchor samples. ANCH was performed by median scaling each metabolite of each batch against the anchor samples. BAT was performed by scaling each batch against the median of the experimental samples contained within that batch. In general, the sample mean is more a consistent estimator than the sample median, but the mean can be impaired by skewed distributions when the number of samples in small. The results of Chapter 2 indicate metabolite ion counts tend to be heavily right skewed; thus, the median was

chosen instead out of precaution.

Normalization methods compared in this section include: total ion count (TIC), median absolute deviation (MAD), probabilistic quotient normalization (PQN) and cyclic lowess (CLOW). This list includes a representative mix of commonly used Class I and II normalizers.

4.4.1. Relative Standard Deviation

All considered normalizations were applied and the relative standard deviation (RSD) was compared for each biochemical between the raw (un-normalized) and normalized data. RSD, also known as coefficient of variation (CV), calculates the observed variance in an experimental group as a percentage of the group average and is often used in clinical chemistry to judge the performance of an assay. The total variance will be a combination of biological and instrument effects. In theory, lower instrument error should lead to lower RSDs.

4.4.2. Variance Components of Experimental Samples

A variance components model using instrument run as a random effect was also performed on the experimental samples in the various normalized versions of the data. This model, which was fit to each metabolite, is given by

$$y_{jib} = \mu_i + D_{ib} + \varepsilon_{jib}$$

in which μ_i is the overall mean ion count of the i^{th} feature, $D_{ib} \sim N(0,\sigma_{Di}^2)$ is the random error associated with the b^{th} instrument batch and $\varepsilon_{jib} \sim N(0, \sigma_i^2)$ is the error associated with the individual measurement representing the population variation where D_{ib} and ε_{jib} are independent. The variance of an individual sample is then

$$var\left(y_{jib}\right) = \sigma_{Di}^2 + \sigma_i^2$$

By finding estimates $\hat{\sigma}_{Di}$ and $\hat{\sigma}$, the proportion of the overall variance due to the instrument batch can be estimated as $\hat{\sigma}_{Di}^2 / \left(\hat{\sigma}_{Di}^2 + \hat{\sigma}_i^2\right)$. Normalizations that effectively address batch-to-batch

variation in instrument performance should, ideally, reduce this variance component close to 0.

4.4.3. Variance Components of Anchor Samples

A useful result of decomposing the variance into batch and instrument error components is the ability to quantify the number of anchor samples needed to reduce the overall error below an acceptable threshold. By fitting a variance component model on the anchor samples gives

$$y_{anch\,j,\,ib} = \mu_{anch,\,i} + D_{anch,\,ib} + \varepsilon_{anchj,\,ib}$$

with $\mu_{anch,\,i}$, $D_{anch,\,ib}$ and $\varepsilon_{anchj,\,ib}$ the parameters specific to the anchor samples but analogous to the previously described components model. The variance of the batch anchor average is

$$var\left(\bar{y}_{anch,\,ib}\right) = \sigma^2_{anch,\,Di} + \frac{\sigma^2_{anch,i}}{k}$$

where k is the number of anchor samples used in batch b. By increasing k, the contribution of batch component to the variance can be reduced to an acceptable level. Reducing this error as much as possible is ideal and it is the total variance, which depends greatly on $\sigma^2_{anch,\,Di}$, which may be most important. However, acceptable limits of variation are not immediately available across the entire metabolome relative to the ion-count levels of a given batch.

As an alternative, limiting the total variance to within a certain percentage of $\sigma^2_{anch,\,Di}$ is considered. The formula to determine the minimum size k which keeps the variance of the anchor batch average within p% of the instrument error is

$$\sigma^2_{anch,\,Di} + \frac{\sigma^2_{anch,i}}{k} \leq (1+p)\sigma^2_{anch,\,Di}\,.$$

Solving for k gives

$$k \geq \frac{100}{p}\frac{\sigma^2_{anch,i}}{\sigma^2_{anch,\,Di}}\,.$$

There is no clinically accepted or otherwise immediately obvious value of p from which to investigate, but values of 5-20% appear to be a reasonable place to start.

4.4.4. Global vs. Targeted

Quantitative measurements using stable isotope dilution LC-MS/MS assays for the true concentration of 7 biochemicals detected in the global profiling dataset are also available. Normalized versions of the global dataset were compared to the targeted results using Pearson's R^2 and mean square error (MSE). For the MSE comparisons, both the global and untargeted data were centered and scaled to account for the different scales between the ion counts and concentration.

4.5. Data

Plasma samples from participants enrolled in the Insulin Resistance Atherosclerosis Study (IRAS) were obtained for metabolomic profiling. IRAS was a seventeen-year multicenter, tri-ethnic observational study sponsored by the National Heart, Lung and Blood Institute to examine the relationship between insulin resistance and cardiovascular disease [28]. The initial IRAS cohort enrolled 1,625 African Americans, Hispanics, and non-Hispanic whites between the ages of 45 and 65 in four regions across the country with baseline entry from 1992 and 1994 and a five-year follow up. The IRAS Family Study enrolled approximately 1,280 additional family members of selected Hispanic and African American members from the original IRAS cohort with baseline enrollment between 1999 and 2002 and proceeded by another five-year follow up period [29]. The total number of participants for IRAS Family Study was about 1,440 with ages ranging from 18 to 81. Material from 1,718 plasma samples collected during IRAS and IRAS Family Study were analyzed with global LC-MS/MS metabolomic profiling.

These samples, which are primarily used for monitoring performance of the instrument, serve

as the anchor samples. The entire instrument platform contains four arms, two versions of both positive and negative ion modes. The two positive ions modes split the chromatogram in half and thusly designated Pos Early and Pos Late. Both negative modes survey the entire chromatogram with one specifically intended for polar compounds. These arms are designated Neg and Polar. Accommodating the number of samples in the study required between 13 and 15 instrument runs per arm. Between 6 and 24 aliquots of pooled plasma were included on each batch. TIC was performed on each arm of the platform separately, while the other normalizations were performed across the combined arms.

The resulting analysis measured 1,780 features across the four arms yielding 1,276 unique metabolites. Features included for study here are limited to those that were observed in at least 99% of participant samples, a criterion satisfied by 767 features. Since anchor normalization is dependent on measurement of a feature in the anchor matrix, and can be unreliable if not well detected, a detection requirement of at least 2/3 in the anchor samples was imposed in order to include a feature in this normalization process. Under this requirement, an additional 24 features were lost because the anchor samples did not reliably measure these features. Three features were lost because the anchor was insufficiently filled in every instrument run, while the other 21 features were lost due to low fill in at least one but not all runs.

Seven analytes had previously been shown to be markers for impaired glucose tolerance as described by Cobb, Eckhart, Perichon, Wulff, *et al*. [30]. The markers in question are α-hydroxybutyrate (AHB)(Polar), β-hydroxybutrate (BHB)(Polar and Pos Early), 4-methyl-2-oxopentanoic acid (4MOP)(Neg), 1-linoleoylglycerolphosphocholine (LGPC)(Neg, Polar, Pos Late), oleic acid (Neg), pantothenate (Neg and Polar) and serine (Neg, Polar, Pos Early). These analytes were measured quantitatively using a separate quantitative mass spectrometric assay in

all 1,718 subjects.

4.6. Software

All analysis was performed in R version 3.4.3 [31]. The following packages were used: limma package [32], nlme package [33] and Data Normalization R-script by Hochrein *et al*. [34].

4.7. Results

4.7.1. Global data

Results from the variance components model, shown in Table 4.2, show that standard -omic normalizations do not effectively reduce the impact of instrument run. Across the 743 features, on average 35% of the observed variation is attributed to instrument run. MAD, PQN and CLOW do not improve upon this number much whilr MAD actually increases, though only slightly and likely to be random noise level, in instrument effect. Overall profiles of the five-number summary for MAD, PQN and CLOW are more or less the same as the profile of the raw data as well. TIC does manage to address instrument variation somewhat, lowering the average contribution across the metabolites down to 24% and the five-number summary also shows TIC

Table 4.2: Variance components in global metabolites. Five-number summary of variance proportion due to instrument run across all 743 metabolites.

DATA	Min	1st Quartile	Median	3rd Quartile	Max	Mean
Raw	0.4%	17.6%	33.2%	49.8%	94.1%	35.0%
TIC	0.4%	10.0%	19.0%	34.2%	90.6%	24.1%
MAD	0.6%	19.9%	33.2%	49.8%	90.8%	35.4%
PQN	0.2%	16.2%	33.1%	48.8%	90.7%	34.0%
CLOW	0.2%	15.7%	31.7%	48.7%	89.7%	33.4%
ANCH	0.0%	1.7%	3.6%	9.9%	63.8%	7.3%
BAT	0.0%	0.0%	0.0%	0.0%	13.0%	0.2%

to be consistently lower than Raw. ANCH and BAT on the other hand dramatically reduce the variation due to instrument run. This is not surprising for BAT which centers based on the participant samples from which the D_{ib}'s are calculated. Thus, the variance component is not highly insightful in terms of what to expect in a future dataset for this method. ANCH, on the other hand, is completely independent of participant samples, and the average instrument component here is just 7.3%. Additionally, half of the metabolites in this experiment have a component proportion $\leq$ 3.6% and three quarters have a component proportion below 10%.

Next, RSDs of participant samples across all metabolites are shown in Table 4.3. Strictly speaking, lower RSD need not, as a general rule, automatically signal more effective removal of instrument effect. This is because ion counts for the experimental samples include both biological and instrument variation, as opposed to technical replicates which contain only instrument variation. However, with around one third of the ion count variation being due batch effects on average, it seems reasonable that normalizations effectively removing instrument run will generally be lower. On average the overall standard deviation is roughly 63.5% of the

Table 4.3: Coefficient of Variation in global metabolites. Five-number summary of RSDs across all 743 metabolites.

DATA	Min	1st Quartile	Median	3rd Quartile	Max	Mean
Raw	11.4%	39.9%	54.2%	72.6%	332.9%	63.5%
TIC	9.2%	36.3%	48.8%	69.2%	248.2%	58.3%
MAD	19.7%	39.1%	51.6%	68.3%	239.5%	59.3%
PQN	15.9%	37.7%	51.3%	69.6%	253.0%	59.9%
CLOW	12.3%	37.2%	51.1%	68.6%	244.4%	58.8%
ANCH	7.5%	32.4%	44.7%	61.8%	328.8%	54.6%
BAT	7.1%	31.2%	43.2%	61.1%	291.9%	52.8%

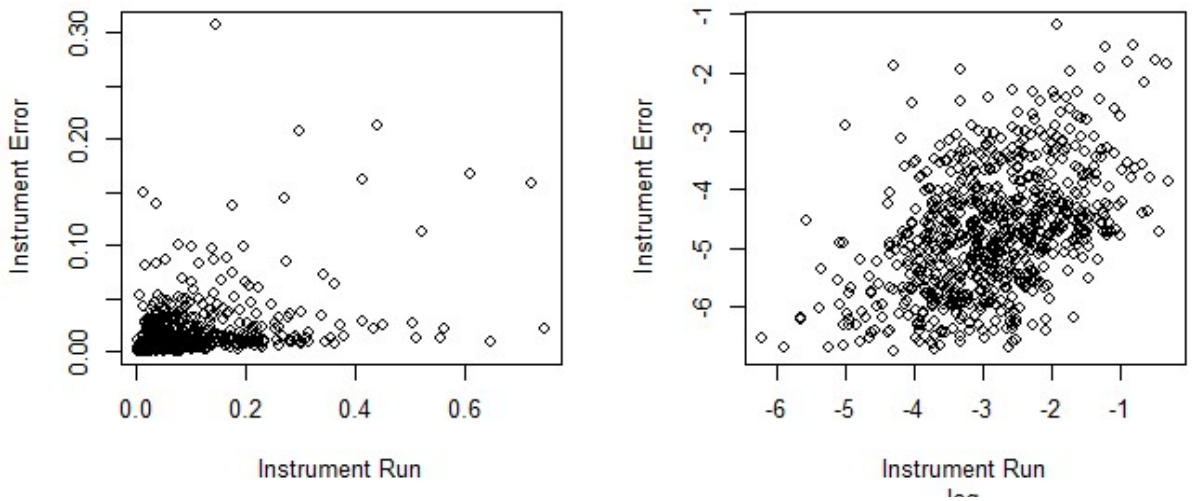


Figure 4.1: Instrument run variation versus instrument error variation. Based on estimates from variance components model

metabolite mean in the uncorrected data. All of the normalization methods generally reduce the RSD, both on average and across the quantiles of the five-number summary. The standard -omic normalizations show a small, but consistent decrease in the RSD with an average value around 59%. ANCH and BAT are even lower, with ANCH less than 55% on average and BAT less than 53%.

Moving on to the variance components model in the anchor samples, the five-number summary for the proportion due to the batch component, $\hat{\sigma}_{anch,\,Di}$, in the raw data is as follows: Minimum = 8.3%, 1st Quartile = 75.8%, Median = 87.9%, 3rd Quartile = 93.3%, Maximum = 98.9%. The average $\hat{\sigma}_{anch,\,Di}$ is 82.5%. So, as expected variation between instrument runs account for a great deal of variability, with an average 82.5% of the variation in the raw ion counts of the anchor samples being due to this batch effect. *Figure 4.1* plots the individual $\hat{\sigma}_{anch,\,Di}$*'s* against their respective $\hat{\sigma}_{anch,i}$*'s*. Interestingly, the estimated standard deviations of the two components are somewhat similar in scale, though batch component does tend to be larger the replicate component, which fits with the average percentage due to batch being so high. The plot is also populated with several outliers in both components. Taking a log transformation helps to better

Table 4.4: Summary of minimum anchor replicate limiting instrument error to 5%. Based on estimates $\hat{\sigma}^2_{anch,\,i}$ and $\hat{\sigma}^2_{anch,\,Di}$ from variance components model.

Min	1st Quartile	Median	3rd Quartile	Max	Mean
0.0024	0.1018	0.3799	2.036	2469.1	9.299

see the individual standard deviation estimates and reveals a weak, positive correlation between the two components. Table 4.4 shows the five-number summary across all metabolites for sample size required to minimize instrument error to no more than 5% of the batch component. On average the metabolites require just over 10 replicates, but this is inflated by the presence of a few metabolites requiring extremely large sample sizes. Three quarters of the metabolites surveyed actually require just over 2 replicates and 86.4% require 5 or less (not shown).

4.7.2. Targeted Versus Global

For the 7 metabolites measured by both the quantitative and semi-quantitative assays, raw ion counts are plotted against their concentration levels from the targeted analysis in *Figures 4.2* and *4.3*. Individual points are colored by instrument run and there is noticeable banding of colors in most of the features. Among the metabolites that are measured on multiple arms, instrument effect can vary by arm. Plots for LGPC for instance indicate a handful of rather poor performing batches in Polar and Pos Late, but in Neg batch differences are less pronounced. Similarly, serine has clear differences between instrument days on Neg and Polar, but such effects are not as apparent on Pos Early. Thus, neither arm is likely to be superior overall, but certain arms may perform better for certain metabolites. Table 4.5 and *Figure 4.4* show the R^2 between the clinically measured concentrations and raw intensities by normalization method. Certain metabolites correlate better than others. Oleate has raw R^2 of 0.777 while 4MOP has a raw R^2 of 0.584 for instance. Additionally, for metabolites observed on multiple arms the performance may

Figure 4.2: Targeted vs raw ion counts. Points are colored to highlight separate instrument runs.

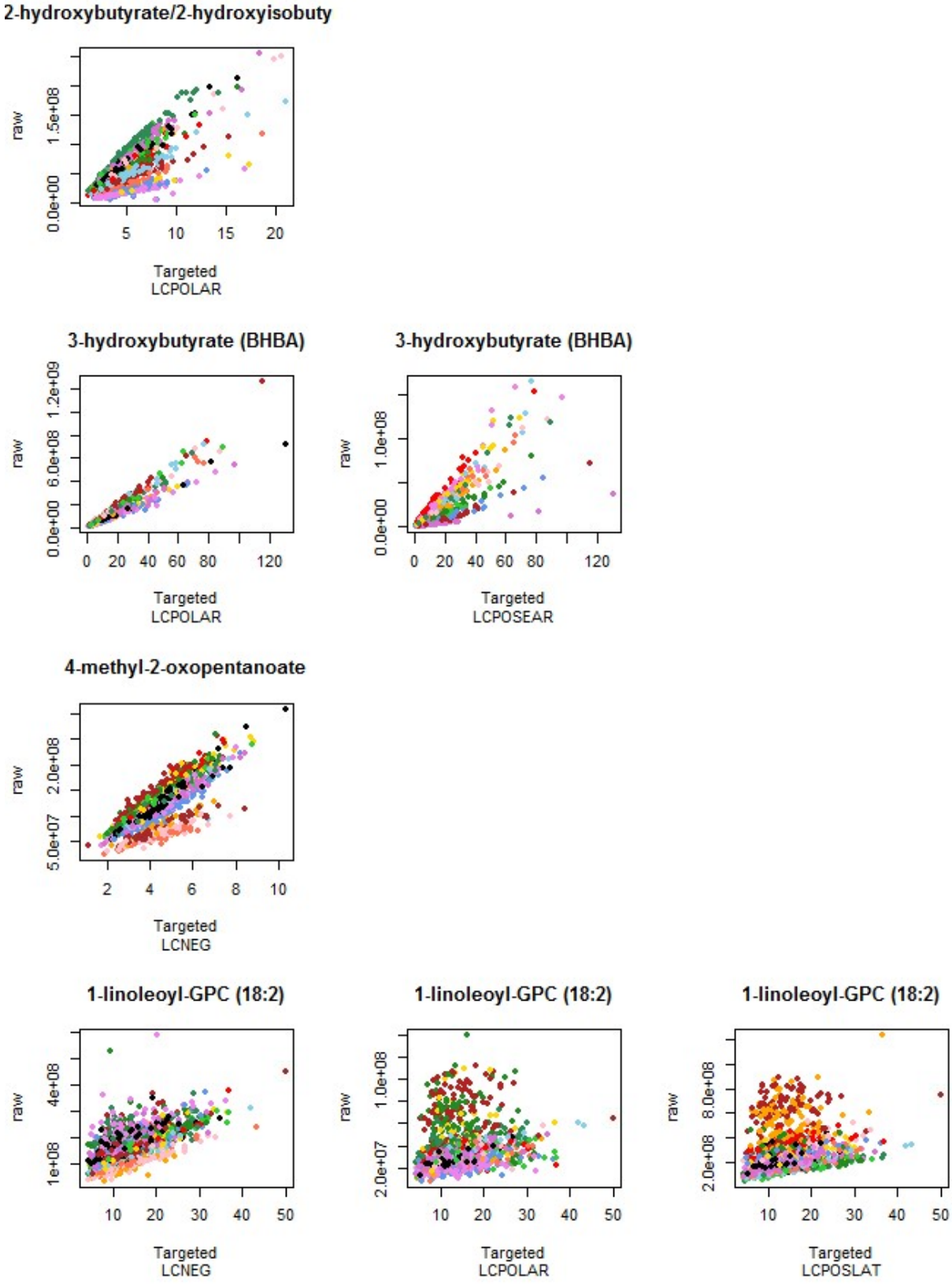


Figure 4.3: Targeted vs raw ion counts continued. Points are colored to highlight separate instrument runs.

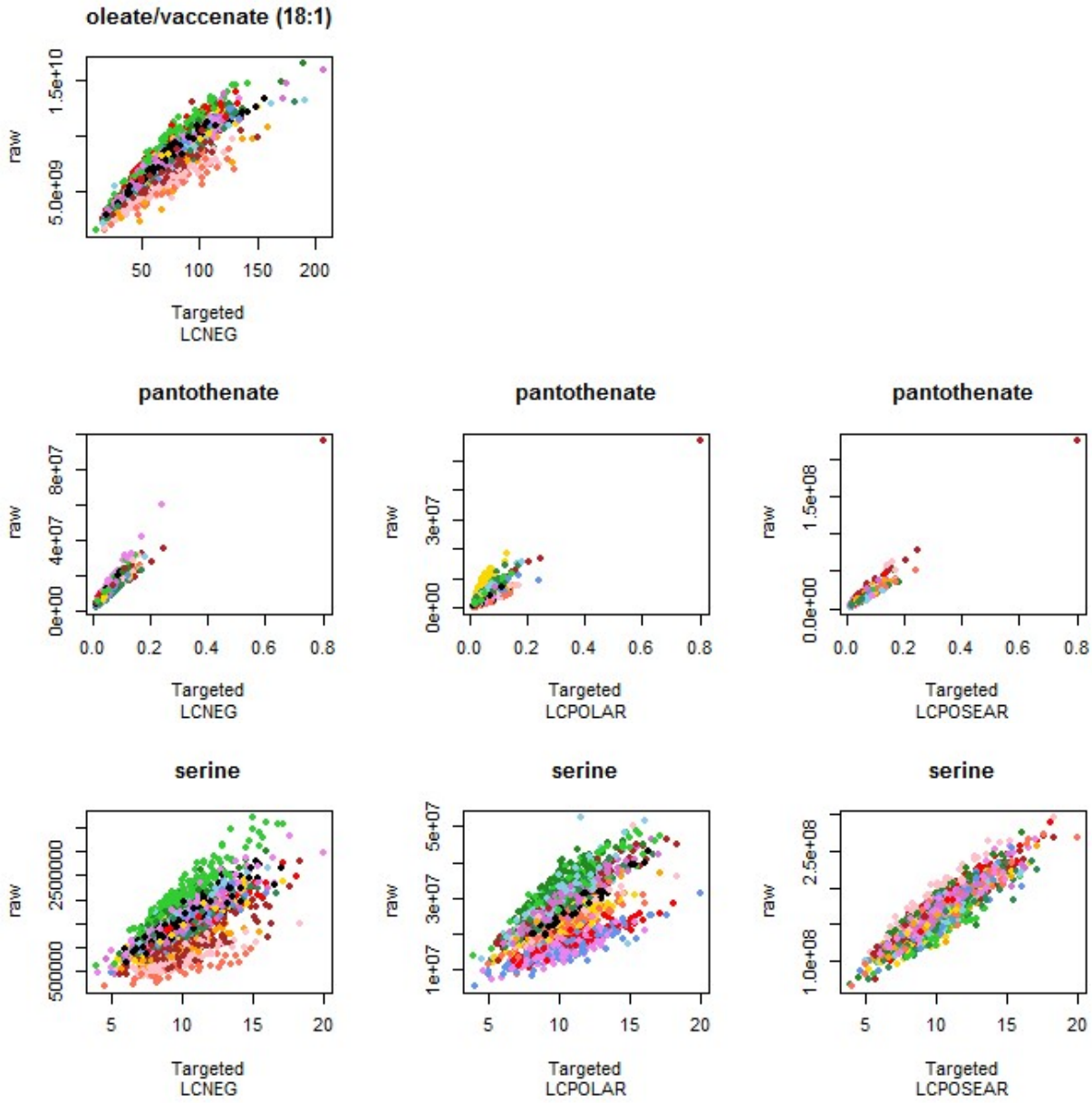


Table 4.5: R^2 in targeted metabolites. Targeted concentrations from clinical assay versus global ion counts.

			Normalizations					
Compound	Arm	Raw	TIC	MAD	PQN	CLOW	ANCH	BAT
2-hydroxybutyrate	Polar	0.471	0.356	0.385	0.446	0.434	0.915	0.890
3-hydroxybutyrate (BHBA)	Polar	0.920	0.892	0.907	0.916	0.917	0.976	0.940
3-hydroxybutyrate (BHBA)	Pos Early	0.710	0.709	0.642	0.710	0.710	0.944	0.941
4-methyl-2-oxopentanoate	Neg	0.584	0.533	0.469	0.586	0.543	0.897	0.899
1-linoleoyl-GPC (18:2)	Neg	0.396	0.430	0.258	0.326	0.324	0.483	0.576
1-linoleoyl-GPC (18:2)	Polar	0.064	0.109	0.059	0.054	0.058	0.132	0.182
1-linoleoyl-GPC (18:2)	Pos Late	0.067	0.078	0.063	0.064	0.067	0.244	0.313
oleate/vaccenate (18:1)	Neg	0.777	0.570	0.510	0.707	0.672	0.900	0.904
pantothenate	Neg	0.837	0.573	0.562	0.685	0.649	0.864	0.887
pantothenate	Polar	0.724	0.643	0.523	0.680	0.666	0.928	0.921
pantothenate	Pos Early	0.901	0.877	0.756	0.819	0.818	0.939	0.932
serine	Neg	0.460	0.413	0.364	0.459	0.454	0.802	0.808
serine	Polar	0.331	0.513	0.298	0.330	0.332	0.664	0.755
serine	Pos Early	0.796	0.721	0.510	0.571	0.620	0.854	0.841

Figure 4.4: R^2 in targeted metabolites. X-axis are individual biochemical versions listed in ascending order of raw R^2. Platform names have been shortened such that N = Neg, P = Polar, PE = Pos Early and PL = Pos Late.

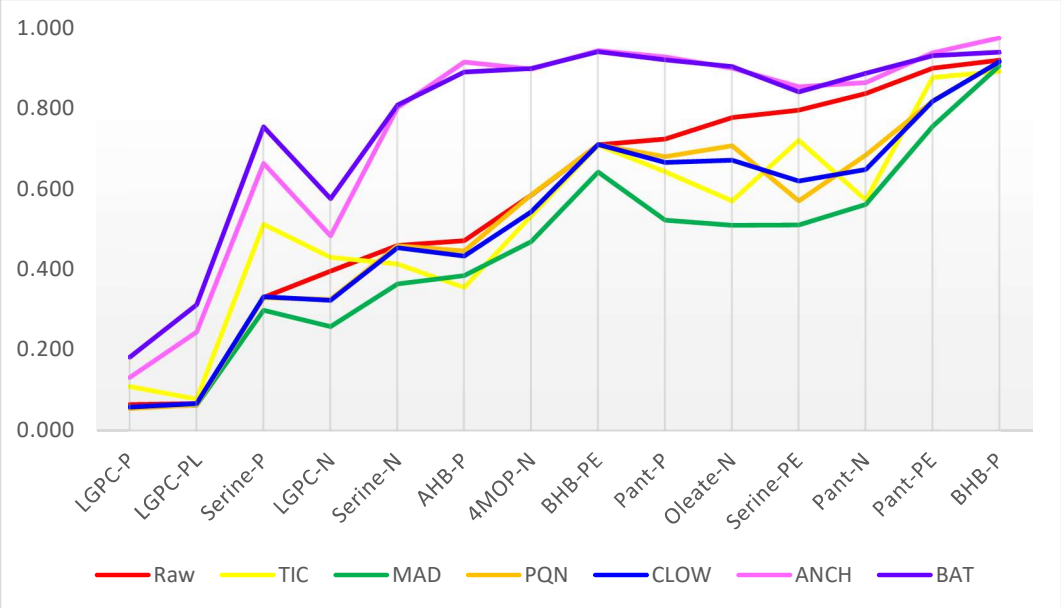
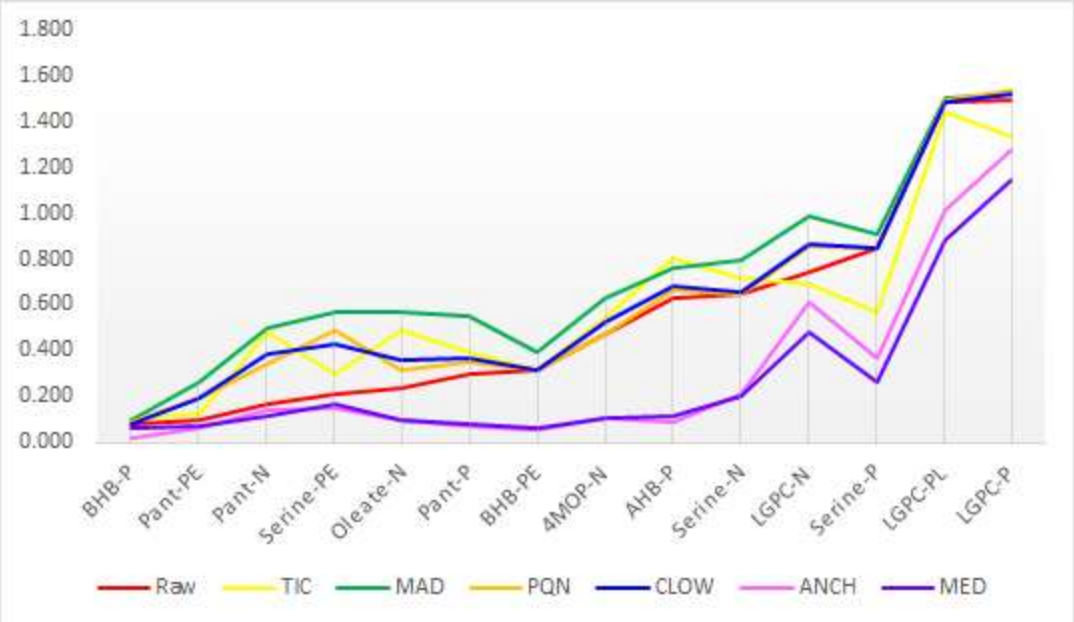


Table 4.6: MSE for targeted metabolites. Targeted concentrations and global ion counts are mean centered and scaled against the standard deviation.

			Normalizations					
Compound	Arm	Raw	TIC	MAD	PQN	CLOW	ANCH	BAT
2-hydroxybutyrate	Polar	0.627	0.807	0.759	0.664	0.683	0.087	0.113
3-hydroxybutyrate (BHBA)	Polar	0.082	0.111	0.096	0.086	0.085	0.024	0.061
3-hydroxybutyrate (BHBA)	Pos Early	0.315	0.316	0.397	0.315	0.314	0.056	0.060
4-methyl-2-oxopentanoate	Neg	0.471	0.539	0.630	0.469	0.526	0.105	0.103
1-linoleoyl-GPC (18:2)	Neg	0.741	0.689	0.983	0.858	0.862	0.610	0.483
1-linoleoyl-GPC (18:2)	Polar	1.492	1.339	1.515	1.533	1.518	1.274	1.147
1-linoleoyl-GPC (18:2)	Pos Late	1.482	1.442	1.498	1.494	1.481	1.012	0.882
oleate/vaccenate (18:1)	Neg	0.237	0.490	0.572	0.318	0.361	0.102	0.098
pantothenate	Neg	0.170	0.486	0.501	0.345	0.389	0.141	0.116
pantothenate	Polar	0.298	0.396	0.554	0.350	0.368	0.073	0.081
pantothenate	Pos Early	0.102	0.127	0.261	0.190	0.191	0.062	0.070
serine	Neg	0.644	0.714	0.793	0.645	0.653	0.209	0.202
serine	Polar	0.850	0.568	0.907	0.851	0.847	0.370	0.263
serine	Pos Early	0.215	0.302	0.571	0.489	0.426	0.152	0.165

Figure 4.5: MSE for targeted metabolites. X-axis are individual biochemical versions listed in ascending order of raw MSE. Platform names have been shortened such that N = Neg, P = Polar, PE = Pos Early and PL = Pos Late.



better on certain arms than on others, which is consistent with the plots in *Figures 4.2* and *4.3*. This phenomenon is most evident in LGPC, where the raw ion counts have an R^2 on the Neg arm of .396 while the other two arms for this biochemical are just above .06. Similarly, serine has a R^2 in the raw data on Pos Early of nearly 0.8 while the Neg and Polar versions are below 0.5. Some versions are more prone to instrument effects and higher instrument effect will introduce more variation. In extreme cases, as with LGPC on Polar and Pos Late, the instrument's ability to measure the biochemical is lacking to the point that the version of this compound is not trustworthy. Examining the normalized versions of the data, it is often the case that typical normalizations actually result in a lower R^2 compared to the raw version of the data. MAD consistently has the lowest R^2 of any data version. TIC, PQN and CLOW occasionally have a higher R^2, but, as *Figure 4.4* clearly shows, for the most part typical normalizations provide no obvious improvement to this measure. Anchor and Median normalization on the other hand always have a higher R^2 than the raw data. Furthermore, with the exception of LGPC, the R^2 levels are consistently around 0.9, which represents a very strong improvement for AHB, 4MOP and serine (Neg version).

Examination of the MSE, shown in Table 4.6 and *Figure 4.5*, is consistent with the R^2 results. While MSE levels vary by biochemical and platform arm, traditional normalizations generally increase the amount of error between global and targeted versions. In contrast, ANCH and MED have consistently lower MSE over the raw version and in some cases, most notably AHB and serine, the decrease is quite large.

The various versions for each analyte are plotted in *Figures 4.6-4.18*. These help to illustrate visually the superior performance of ANCH and BAT at reducing the instrument run effect. Color bands associated with the separate run days collapse together while the points tighten

Figure 4.6: Normalization comparisons to targeted levels in 3-hydroxybutyrate (Polar Platform).

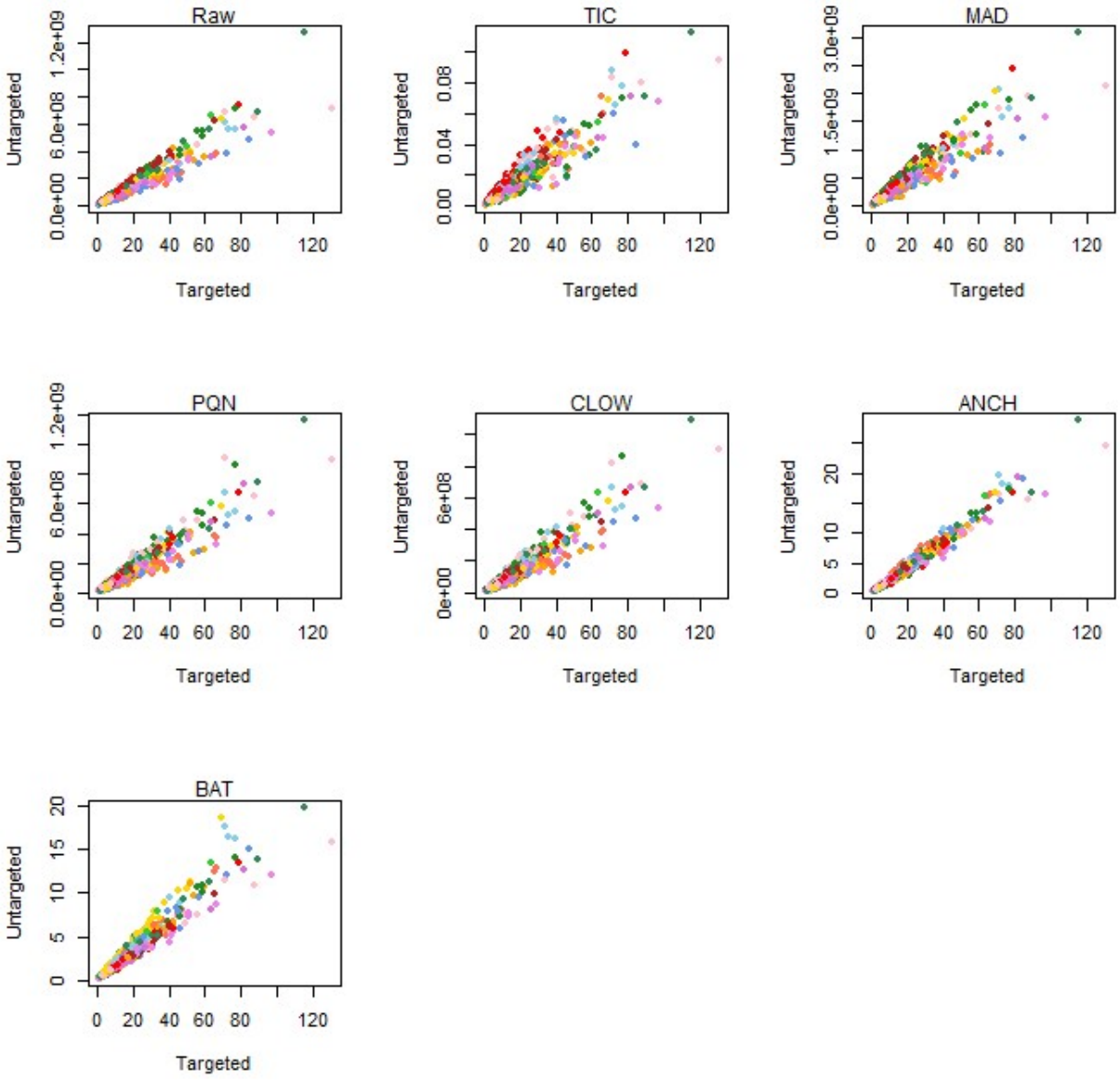


Figure 4.7: Normalization comparisons to targeted levels in 3-hydroxybutyrate (Pos Early).

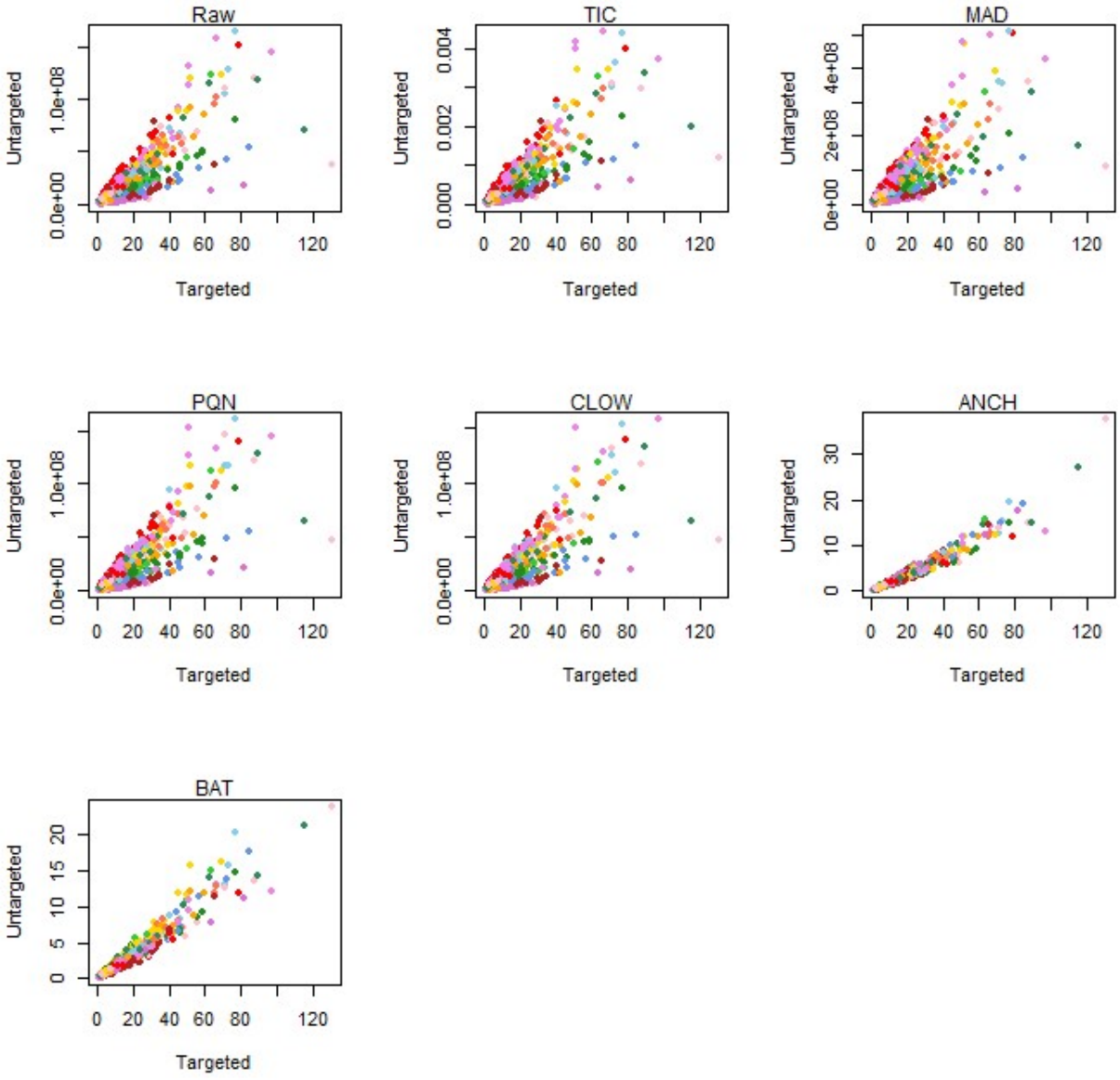


Figure 4.8: Normalization comparisons to targeted levels in 4-MOP.

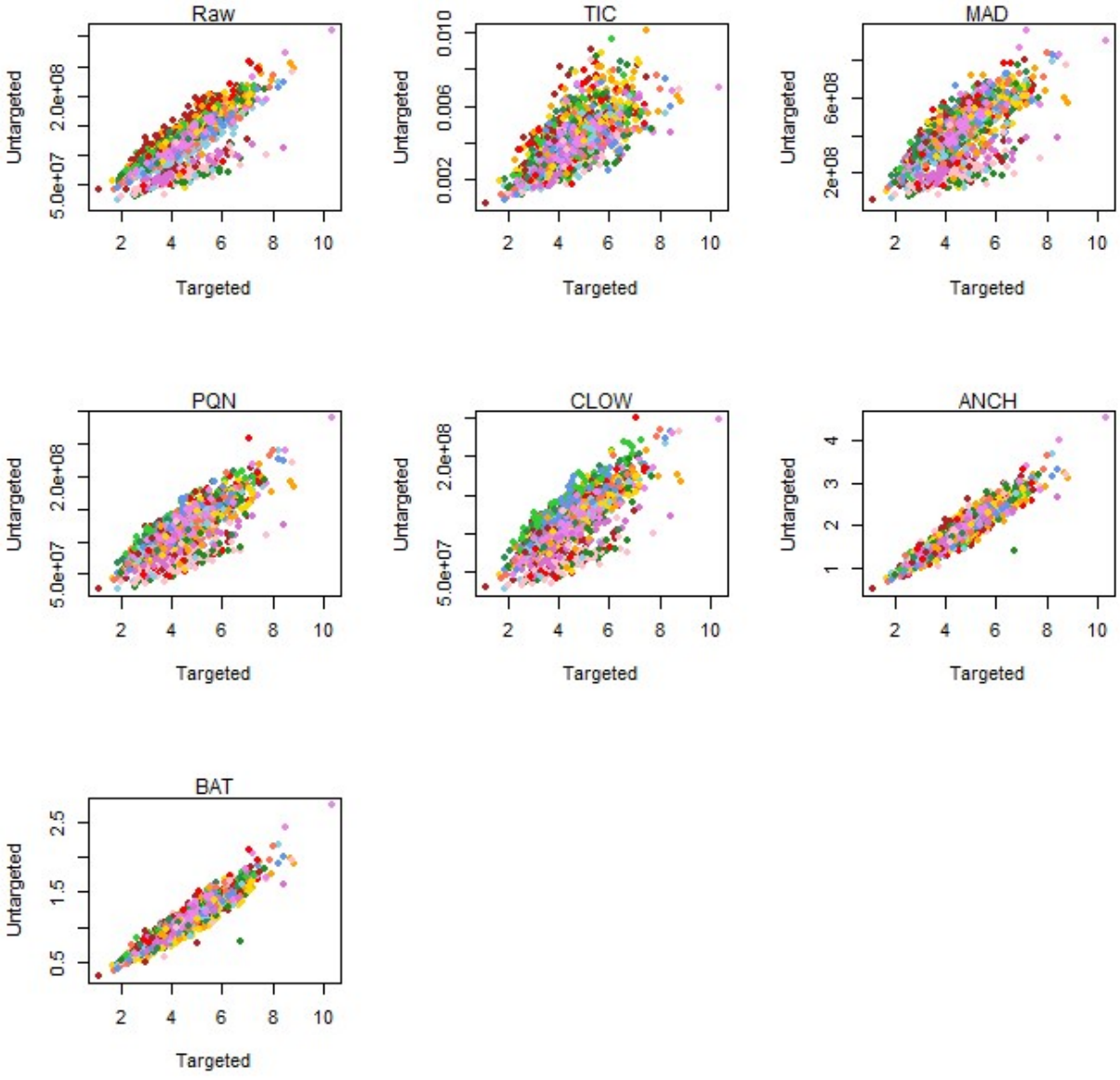


Figure 4.9: Normalization comparisons to targeted levels in L-GPC (Neg).

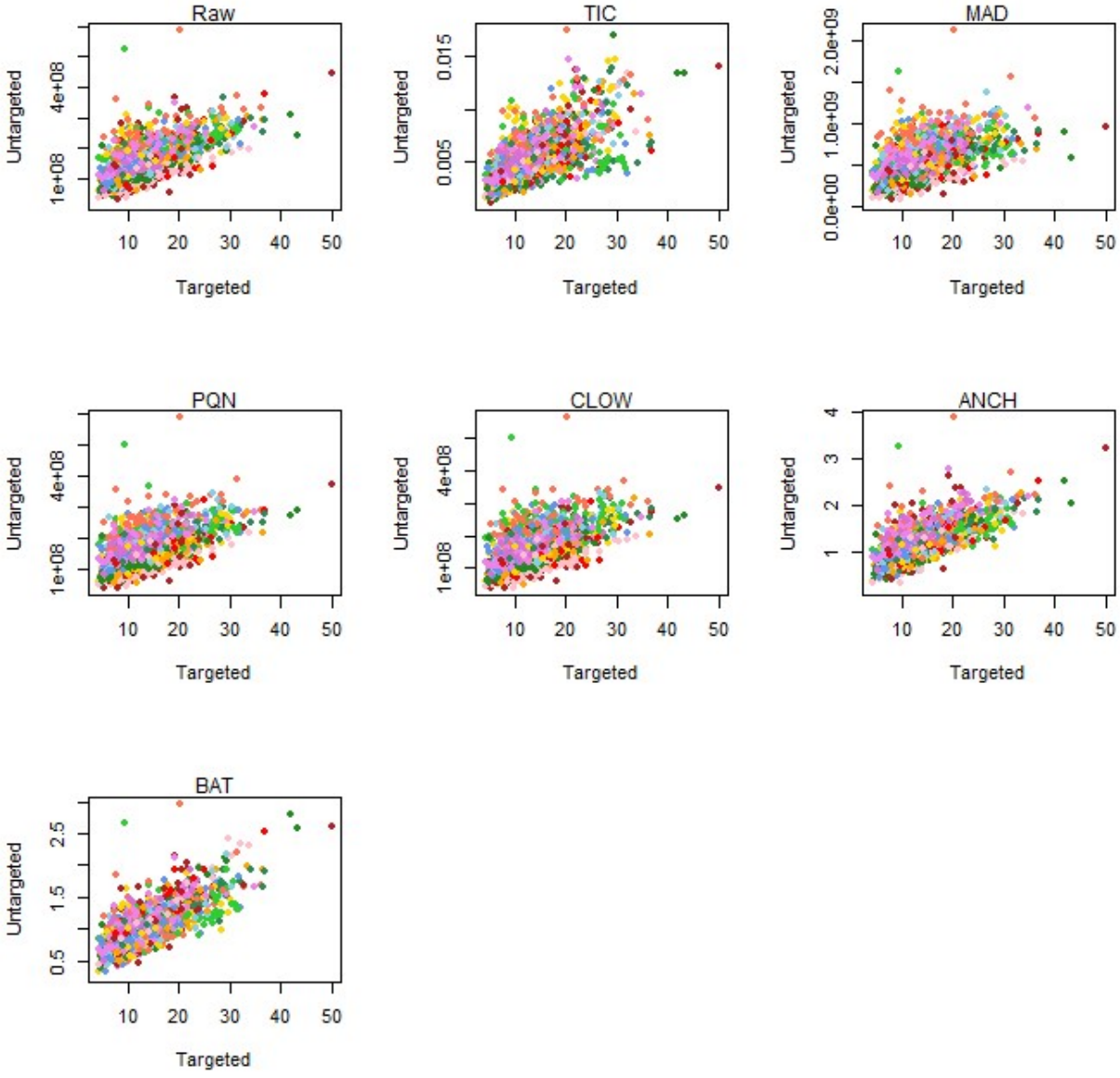


Figure 4.10: Normalization comparisons to targeted levels in L-GPC (Polar).

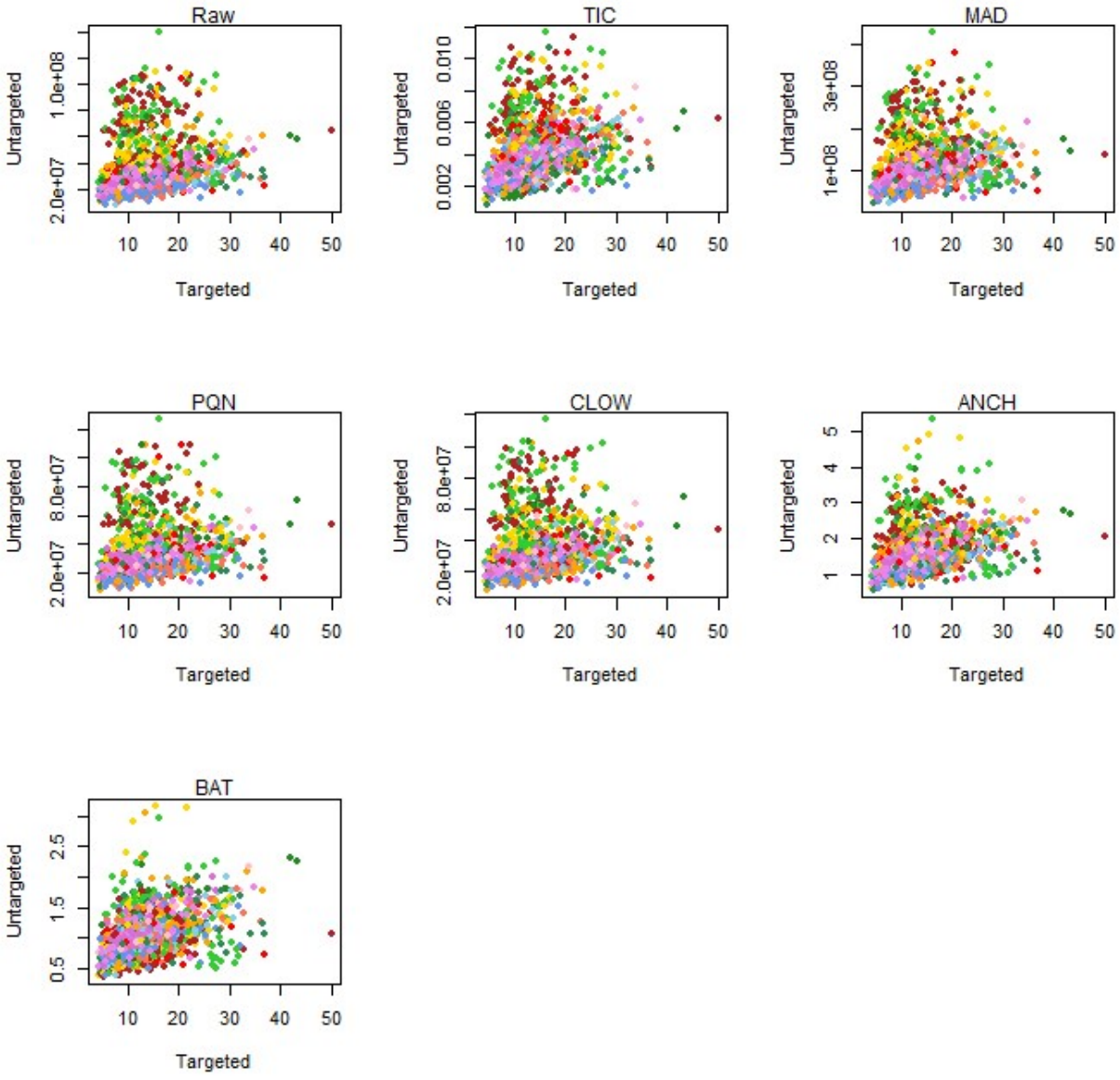


Figure 4.11: Normalization comparisons to targeted levels in L-GPC (Pos Late).

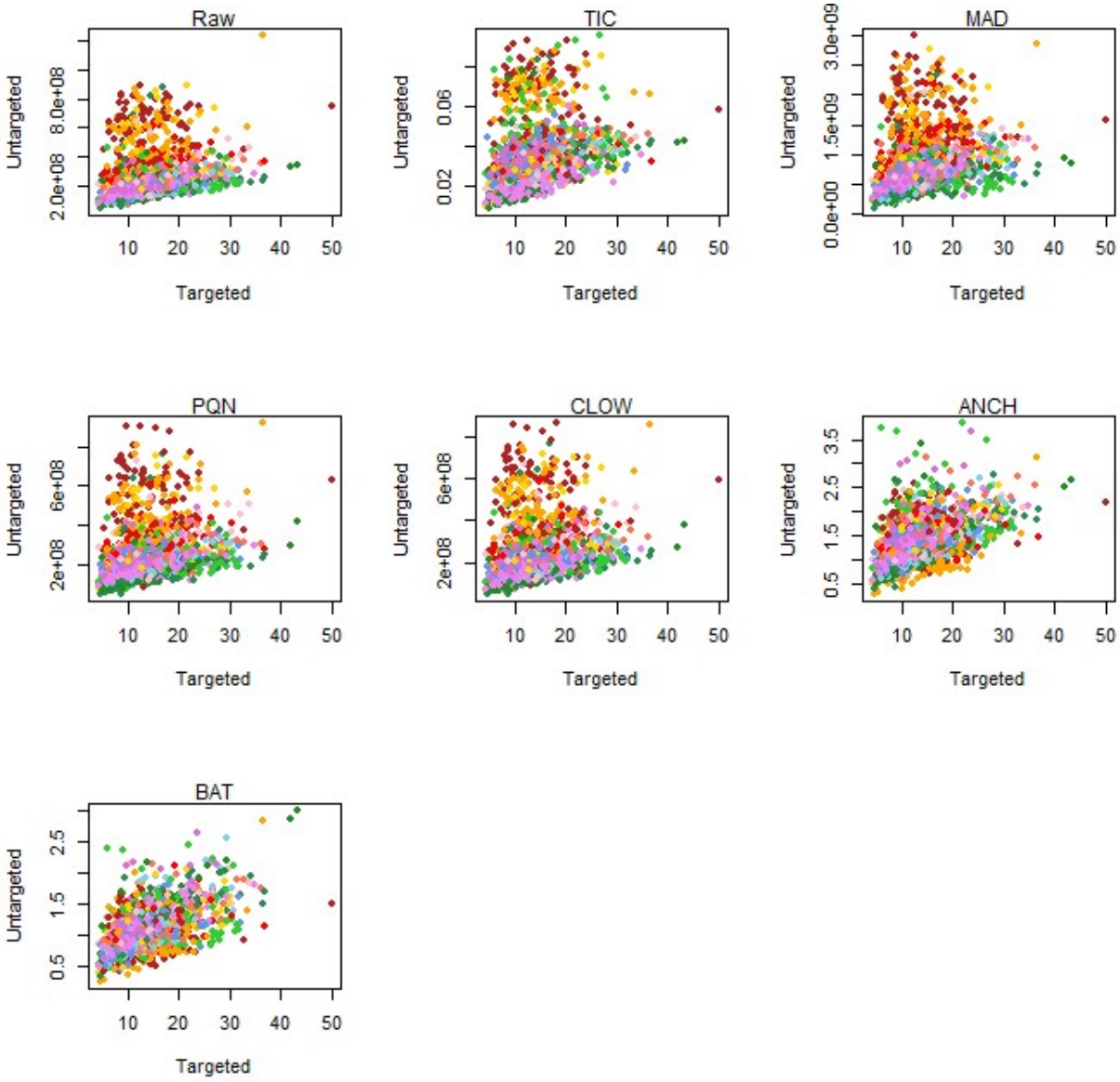


Figure 4.12: Normalization comparisons to targeted levels in Oleate.

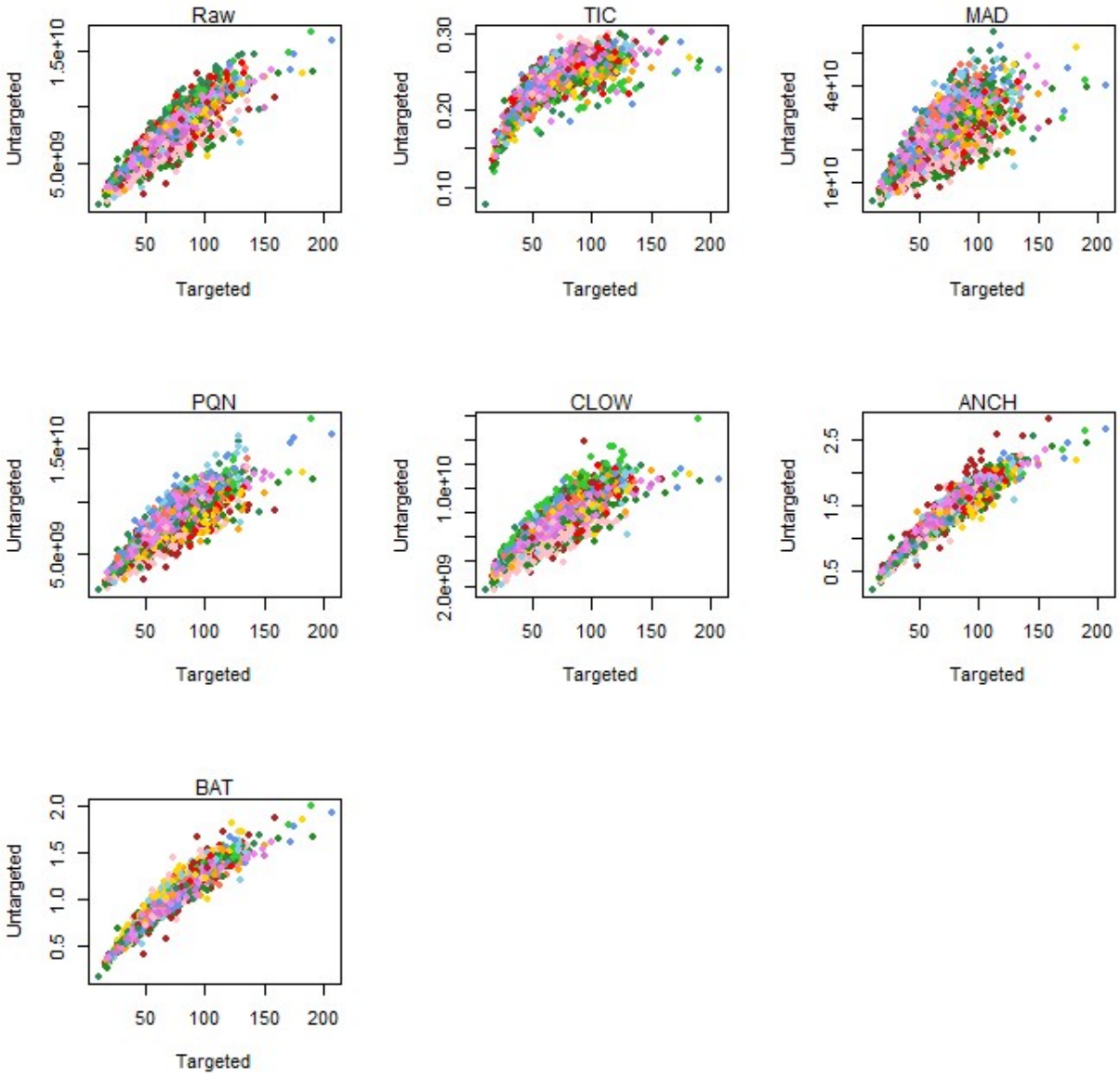


Figure 4.13: Normalization comparisons to targeted levels in Pantothenate (Neg).

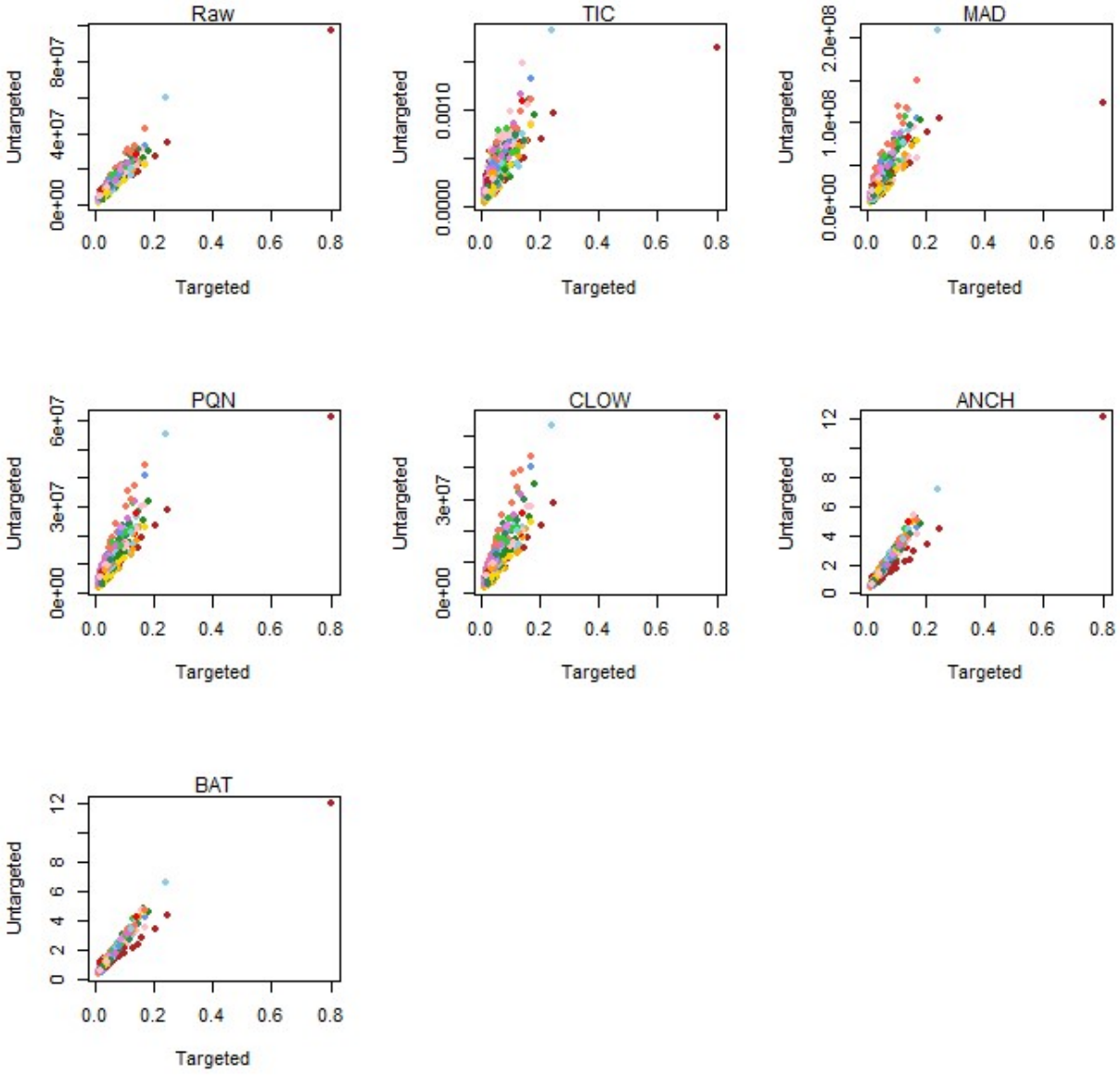


Figure 4.14: Normalization comparisons to targeted levels in Pantothenate (Polar).

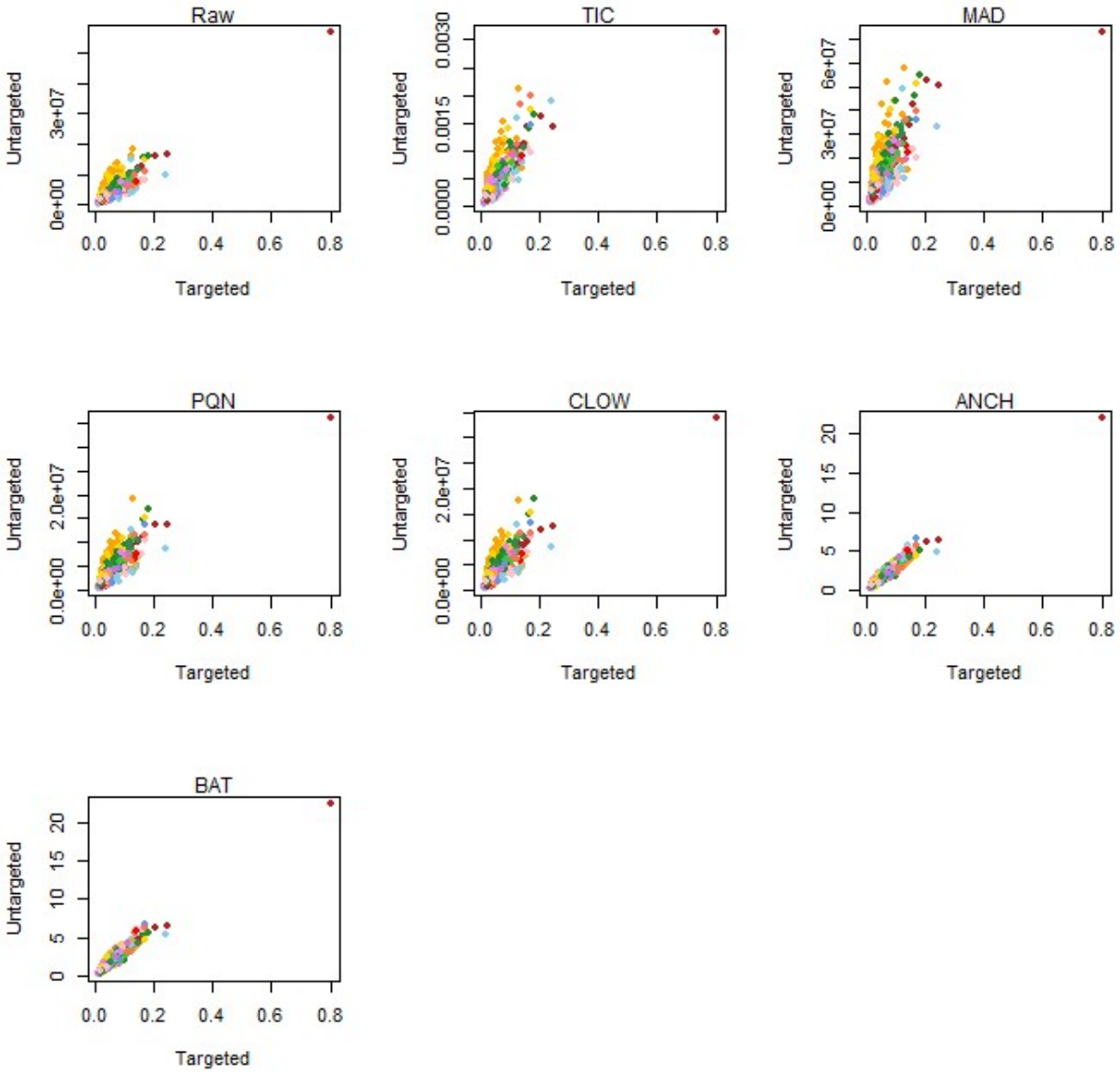


Figure 4.15: Normalization comparisons to targeted levels in Pantothenate (Pos Early).

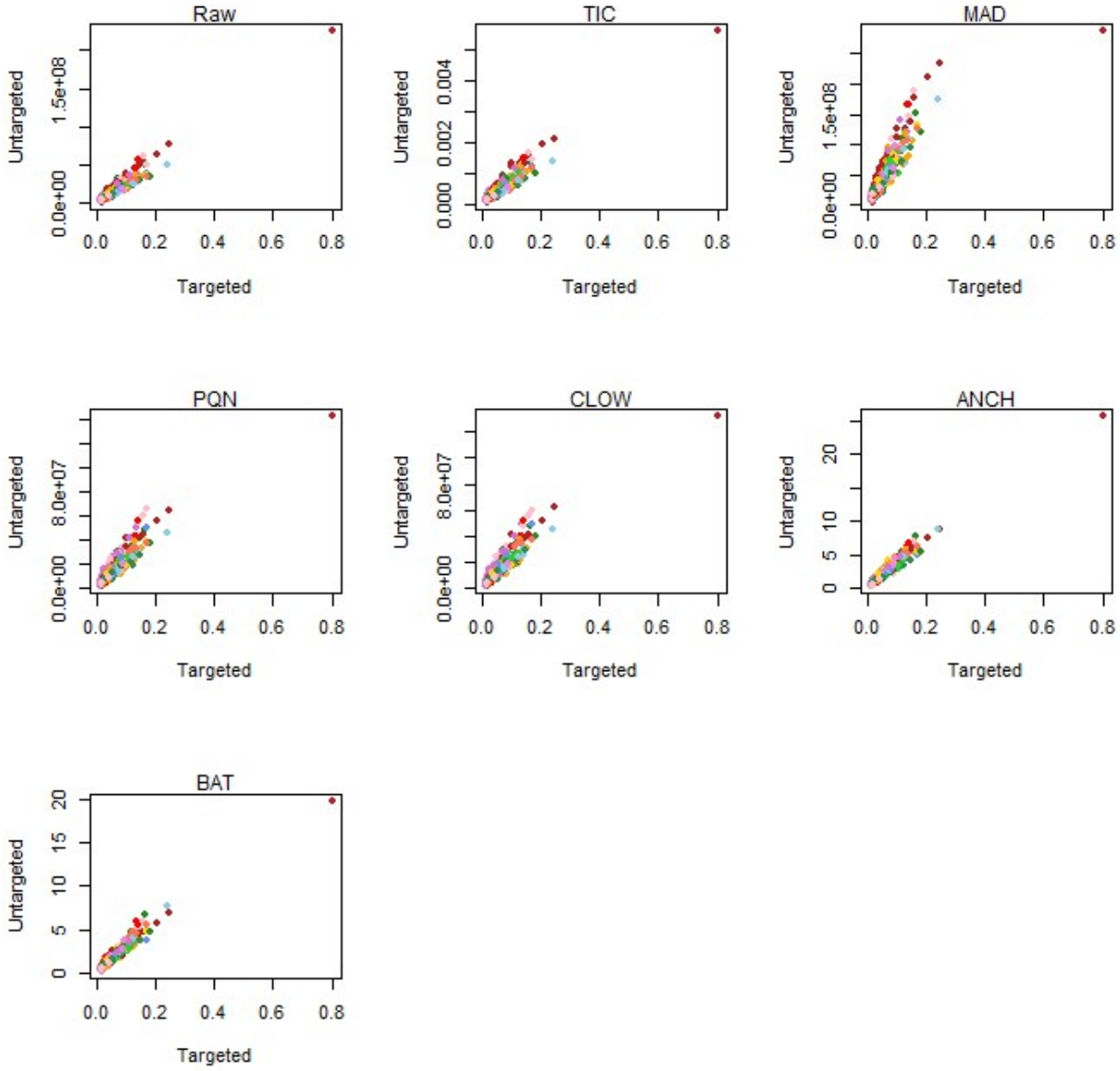


Figure 4.16: Normalization comparisons to targeted levels in Serine (Neg).

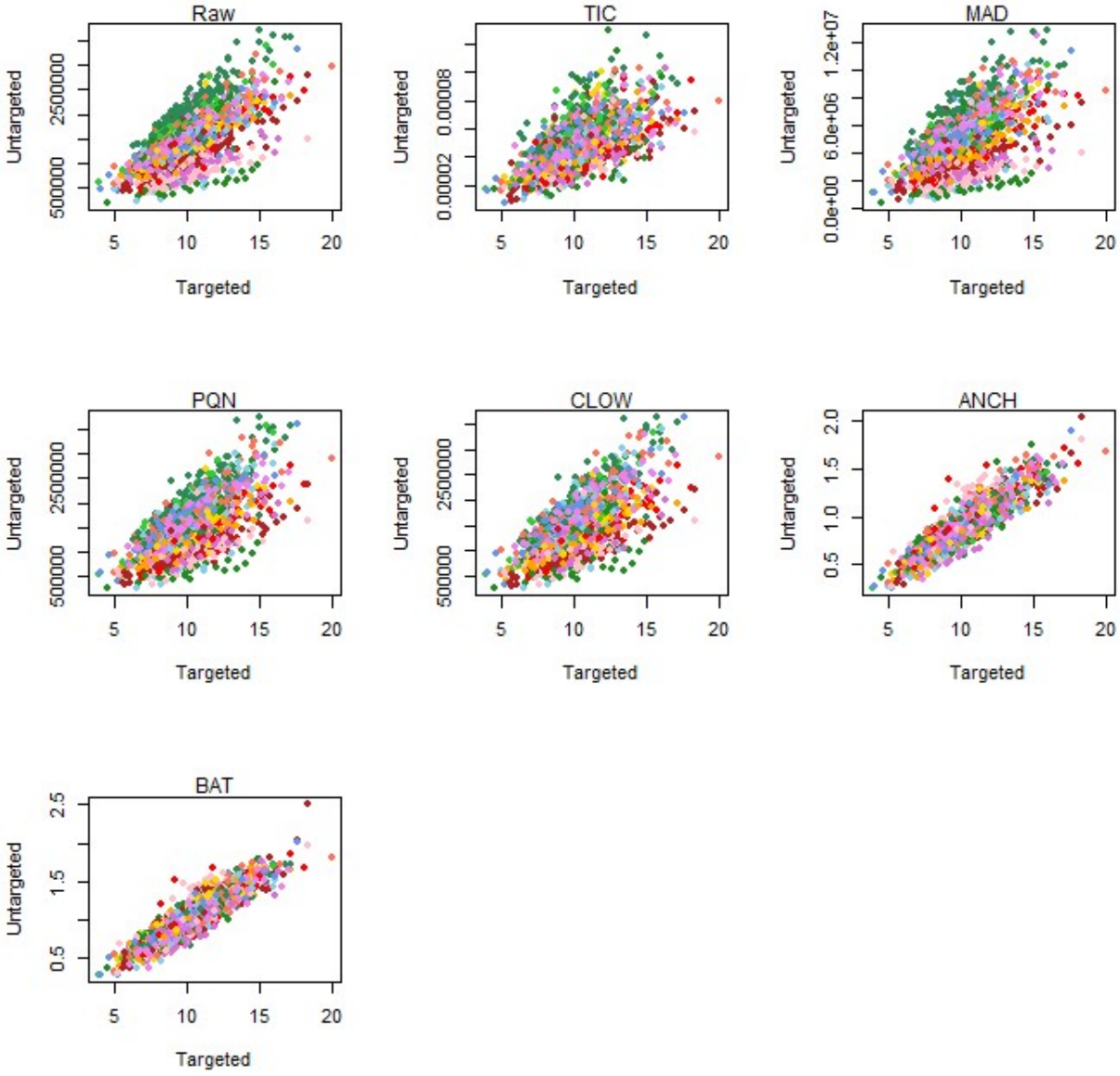


Figure 4.17: Normalization comparisons to targeted levels in Serine (Polar).

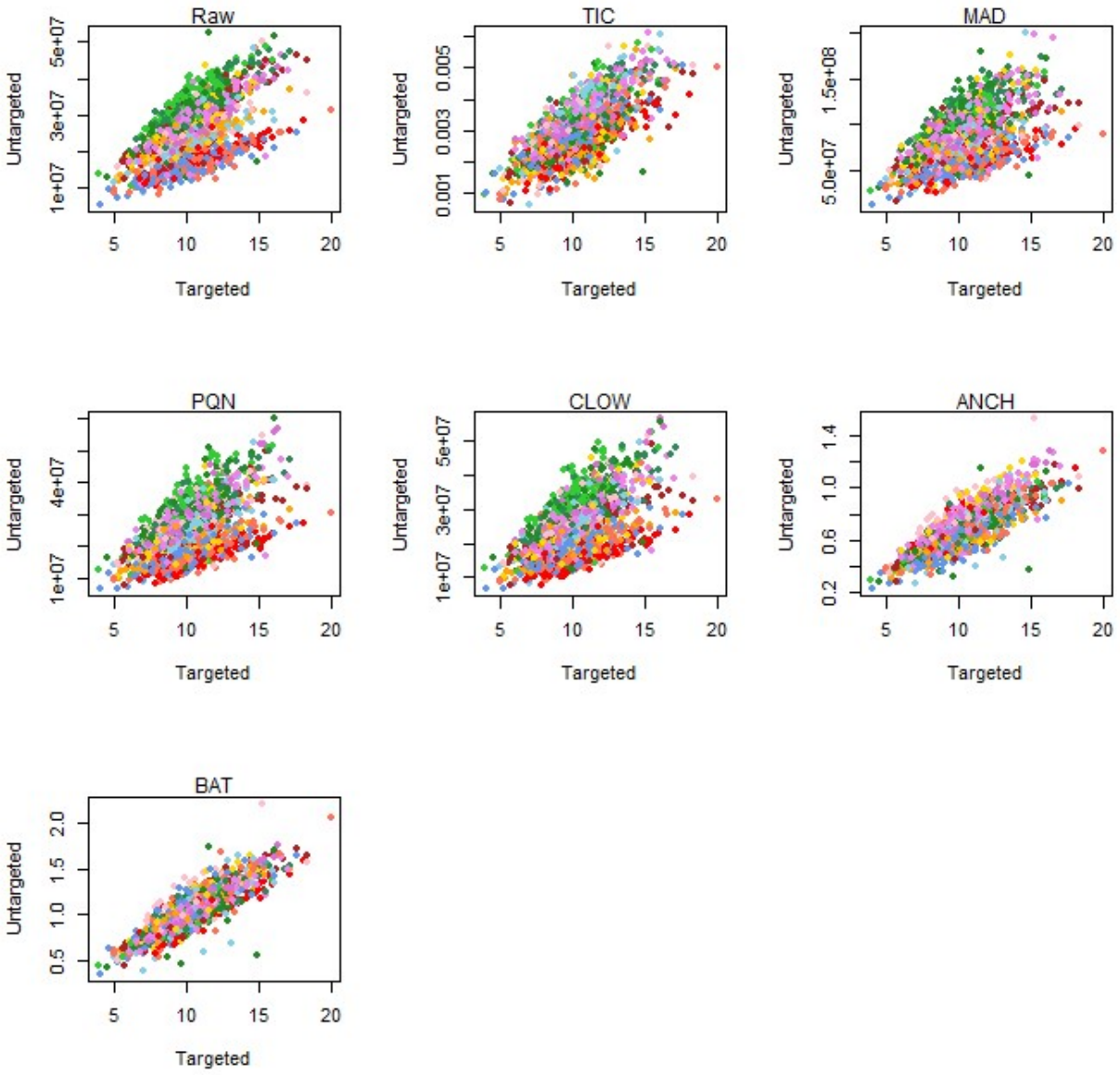
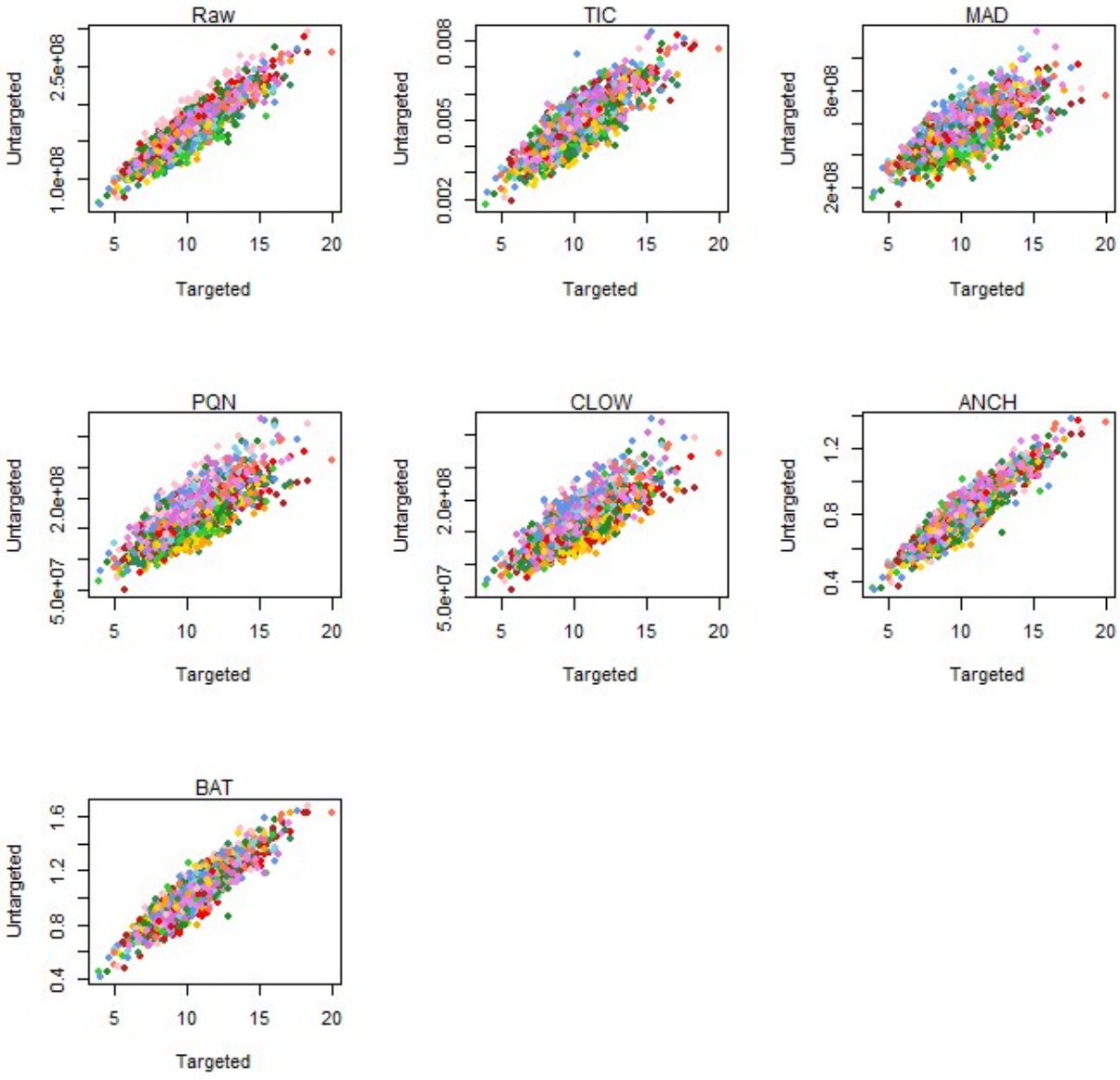


Figure 4.18: Normalization comparisons to targeted levels in Pantothenate (Pos Early).



across the whole range of clinical concentrations. Relationships between untargeted and targeted data is fairly linear. The exception to this observation is oleate which displays concavity in the form of higher values being increasingly depressed. This pattern is consistent with ion suppression, but curiously the non-linearity is present in normalized versions only, particularly TIC, and not the raw values. Other analytes from the same arm as oleate (Neg) do not display this behavior, implying that while normalization is over-correcting oleate the problem is not due the arm itself.

4.8. Conclusions

Results in both the global and targeted data support that normalization based on anchor samples or experimental samples themselves at the individual compound level are much more effective at reducing variation due to instrument run. In every instance of the 7 metabolites with available clinically derived concentrations, anchoring improved the R^2 and MSE over the raw data. The standard omic normalization frequently resulted in little to no improvement, and on occasion would worsen the association between the global and targeted values. Across all global metabolites, anchoring consistently lowered the observed variation in the data more so than the omic methods. There is some evidence that scaling based on the experimental samples themselves does better than anchoring from the independent set of technical replicates. However, in practice this may be difficult due to the number of samples available at the time of each instrument run or the number of experimental groups that must be randomized. The results presented here suggest Anchor normalization with as few as 5 technical replicates was extremely effective and can be feasibly accomplished through the use of quality control samples which are already available in many metabolomic workflows.

REFERENCES

[1] Thermo Fisher Scientific. (2017). *Thermo Scientific Plastics Consumables*. Available: https://www.thermofisher.com/us/en/home/brands/thermo-scientific/molecular-biology/thermo-scientific-plastics-consumables.html

[2] C. Tachibana. (2014) What's next in 'omics: The metabolome. *Science*.

[3] J. W. Dolan, "Calibration Curves Part I: To b or Not to b?," *Chromatography Online,* vol. 27, no. 3, pp. 224-230, 2009.

[4] J. W. Dolan, "Calibration Curves Part II: What are the Limits?," *Chromatography Online,* vol. 27, no. 4, pp. 306-312, 2009.

[5] J. W. Dolan, "Calibration Curves Part III: A Different View," *Chromatography Online,* vol. 27, no. 5, pp. 392-400, 2009.

[6] J. W. Dolan, "Calibration Curves Part IV: Choosing the Appropriate Model," *Chromatography Online,* vol. 27, no. 6, pp. 472-479, 2009.

[7] J. W. Dolan, "Calibration Curves Part V: Curve Weighting," *Chromatography Online,* vol. 27, no. 7, pp. 534-540, 2009.

[8] F. T. Peters and H. H. Maurer, "Systematic comparison of bias and precision data obtained with multiple-point and one-point calibration in six validated multianalyte assays for quantification of drugs in human plasma," *Anal Chem,* vol. 79, no. 13, pp. 4967-76, Jul 2007.

[9] L. M. Schwartz, "Nonlinear calibration curves," *Anal Chem,* vol. 48, no. 14, pp. 2287-8, Dec 1976.

[10] S. M. Kohl, M. S. Klein, J. Hochrein, P. J. Oefner, R. Spang, and W. Gronwald, "State-of-the art data normalization methods improve NMR-based metabolomic analysis," *Metabolomics,* vol. 8, no. Suppl 1, pp. 146-160, Jun 2012.

[11] J. M. Fonville *et al.*, "Robust data processing and normalization strategy for MALDI mass spectrometric imaging," *Anal Chem,* vol. 84, no. 3, pp. 1310-9, Feb 2012.

[12] S. O. Deininger *et al.*, "Normalization in MALDI-TOF imaging datasets of proteins: practical considerations," *Anal Bioanal Chem,* vol. 401, no. 1, pp. 167-81, Jul 2011.

[13] A. C. Reisetter *et al.*, "Mixture model normalization for non-targeted gas chromatography/mass spectrometry metabolomics data," *BMC Bioinformatics,* vol. 18, no. 1, p. 84, Feb 2017.

[14] B. M. Warrack *et al.*, "Normalization strategies for metabonomic analysis of urine samples," *J Chromatogr B Analyt Technol Biomed Life Sci,* vol. 877, no. 5-6, pp. 547-52, Feb 2009.

[15] B. J. Webb-Robertson, M. M. Matzke, J. M. Jacobs, J. G. Pounds, and K. M. Waters, "A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors," *Proteomics,* vol. 11, no. 24, pp. 4736-41, Dec 2011.

[16] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics," *Anal Chem,* vol. 78, no. 13, pp. 4281-90, Jul 2006.

[17] Y. H. Yang *et al.*, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Res,* vol. 30, no. 4, p. e15, Feb 2002.

[18] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, and M. Oresic, "Normalization method for metabolomics data using optimal selection of multiple internal standards," *BMC Bioinformatics,* vol. 8, p. 93, Mar 2007.

[19] M. R. Nezami Ranjbar, Y. Zhao, M. G. Tadesse, Y. Wang, and H. W. Ressom, "Gaussian process regression model for normalization of LC-MS data using scan-level information," *Proteome Sci,* vol. 11, no. Suppl 1, p. S13, Nov 2013.

[20] D. G. Altman and J. M. Bland, "Measurement in medicine: the analysis of method comparison studies," *The Statistician,* vol. 32, pp. 307-317, 1983.

[21] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet,* vol. 1, no. 8476, pp. 307-10, Feb 1986.

[22] M. Astrand, "Contrast normalization of oligonucleotide arrays," *J Comput Biol,* vol. 10, no. 1, pp. 95-102, 2003.

[23] L. C. Kenny *et al.*, "Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers," *Hypertension,* vol. 56, no. 4, pp. 741-9, Oct 2010.

[24] B. Zhou, J. F. Xiao, L. Tuli, and H. W. Ressom, "LC-MS-based metabolomics," *Mol Biosyst,* vol. 8, no. 2, pp. 470-81, Feb 2012.

[25] W. B. Dunn *et al.*, "Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry," *Nat Protoc,* vol. 6, no. 7, pp. 1060-83, Jul 2011.

[26] W. B. Dunn, I. D. Wilson, A. W. Nicholls, and D. Broadhurst, "The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans," *Bioanalysis,* vol. 4, no. 18, pp. 2249-64, Sep 2012.

[27] S. Y. Wang, C. H. Kuo, and Y. J. Tseng, "Batch Normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and

comparison with current calibration methods," *Anal Chem,* vol. 85, no. 2, pp. 1037-46, Jan 2013.

[28] US National Library of Medicine. (2016). *Insulin Resistance Atherosclerosis Study (IRAS)*. Available: https://clinicaltrials.gov/ct2/show/NCT00005135

[29] US National Institutes of Health Archive (2005). *Insulin Resistance Atherosclerosis Study (IRAS)*. Available: https://clinicaltrials.gov/archive/NCT00005135/2005_06_2

[30] J. Cobb *et al.*, "A novel test for IGT utilizing metabolite markers of glucose tolerance," *J Diabetes Sci Technol,* vol. 9, no. 1, pp. 69-76, Jan 2015.

[31] R. C. Team, "R: A Language and Environment for Statistical Computing," ed. Vienna, Austria: R Foundation for Statistical Computing, 2017.

[32] M. E. Ritchie *et al.*, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res,* vol. 43, no. 7, p. e47, Apr 2015.

[33] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R. C. Team, "Linear and Nonlinear Mixed Effects Models," R package version 3.1-131 ed, 2017.

[34] J. Hochrein *et al.*, "Data Normalization of (1)H NMR Metabolite Fingerprinting Data Sets in the Presence of Unbalanced Metabolite Regulation," *J Proteome Res,* vol. 14, no. 8, pp. 3217-28, Aug 2015.

CHAPTER 5: SUMMARY

The chapters comprising this dissertation have practical and immediate implications to challenges of applying untargeted, global MS metabolomics to the clinical environment, and also to the field in general. Chapter 2 informs on basic metabolite characteristics across three common human matrices: plasma, urine and cerebral spinal fluid. Across all three of these matrices, raw ion counts of the chemo-centric approach to metabolomics displayed a consistent right skew that was largely corrected using a natural log transformation. Furthermore, correlations between features using a chemo-centric approach are generally moderate at most, with average correlations of around 0.3. This knowledge is helpful when simulating metabolomic data when examining feasibility of experimental designs or estimating sample size from power calculations.

Building upon this, Chapter 3 presents new parametric methods for handling missing data in metabolomics when values are left-censored due to the true value falling below the instruments level of detection. Compared to standard imputation approaches in the field, both methods are shown to produce more accurate estimates of population parameters. This is most critical to estimates of variance, which can be severely under-estimated in imputed data. Conversely, both proposed methods remain largely unbiased in both the mean and standard deviation even when up to 70-80% of metabolite values are missing. This has clear value for establishing reference ranges of healthy populations, but also for every day metabolomic studies as the problem of missing values is inherent to global MS and researchers often employ designs with few sample sizes due to cost and availability. Parametric approaches can help to maximize power while also

providing more robust results.

The final chapter demonstrates typical omic normalizations that more or less treat all features the same, disregarding the various types of chemical classes present in a global metabolomics set, is deeply flawed. When addressing instrument run effects, such approaches are no better, and often worse, than using the raw ion counts alone. Adjusting for instrument run at the metabolite level, rather than sample level, via a group of anchor samples included in the run or balancing the experimental design across the instrument runs is far more effective at reducing batch effects. Further, when using technical replicates to anchor the batches together, the number of replicates required can be quite low implying minimal cost of instrument resources. This is useful both for clinical practice, which is continuously evaluating new samples, and for large metabolomics studies requiring multiple instrument runs.

While each of these chapters address key elements of metabolomics for clinical practice, truly implementing these techniques requires doing so simultaneously. Handling missing values across multiple merged sets presents a complication as metabolite detection may vary from batch to batch. Assuming different levels of LOD per batch, maximum likelihood is capable of handling this whereas rankit regression is not. However, in anchored data missing values may occur because the metabolite was not detected in the anchor samples. For samples in these batches, it may be more appropriate to treat them as MAR. This gets at the larger issue that a key assumption in these paper, that missing values are due to LOD, is left untested. Although it is generally assumed, it is also believed that some proportion is due to other reasons. An examination into the sources of missing values and their relative contribution to the overall missingness would be highly valuable. If a significant proportion of missingness is MAR, maximum likelihood with truncated rather than censored samples may be more advisable.