Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Sameera R. Wijayawardana	Date

Statistical Methods for Robust Estimation of Differential Protein Expression

By

Sameera R. Wijayawardana

Doctor of Philosophy

Biostatistics

John J. Hanfelt, Ph.D.
Advisor

Tianwei Yu, Ph.D.
Advisor

Junmin Peng, Ph.D.
Committee Member

Amita K. Manatunga, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Robust Estimation of Differential Protein Expression

By

Sameera R. Wijayawardana
B.S., University of Colombo, Sri Lanka, 2001

Advisors: John J. Hanfelt, Ph.D. and Tianwei Yu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Abstract

Statistical Methods for Robust Estimation of Differential Protein Expression

By

Sameera R. Wijayawardana

In the post-genome world, where the sequences of most human genes are known, proteomics has taken up the mantle of being the most promising field of research for *bio-marker* discovery. Proteomics studies yield multi-layered data that pose challenges for statistical analyses due in part to the inherent complexity of the proteomes of organisms, and due to the variability of mass spectrometry based methods that form the back bone of modern proteomics methodologies. An active area of research in proteomics is the assessment of differential expression of proteins in different biological samples, with careful attention being paid to the issue of multiple testing.

To date, little attention has been paid to ensuring the robustness of the statistical results of proteomics data analyses. Nor have there been rigorous attempts to adjust statistical results to account for the high technical variability found in proteomics data. There is also a lack of methods that address the issue of missing values in a model based framework.

In this dissertation, we develop statistical methods for estimating the relative expression level of proteins that are derived from isotopically labeled protein mixtures. We develop an estimator for the overall relative protein expression using a random effects model that uses a variant of the minimum norm quadratic unbiased estimation method to estimate the associated variance components. By assuming different distributional choices for a two-groups model underlying the mechanisms generating the relative expression values, we develop a robust and flexible finite mixture modeling approach for the estimation of the posterior probability of each protein to be non-differentially expressed. In this context, we further investigate the utility of several non-standard statistical distributions: skew-normal, skew Student's t, and the generalized hyperbolic distribution, as suitable candidate distributions for the mixture components fitted to each two-groups model.

We are also interested in adjusting statistical estimation procedures to account for latent error processes that generate a majority of the technical variability in proteomics data. With this regard, we conduct a reliability analysis of the data to remove a subset of the original data that are deemed less reliable, and then use a peptide ion current area based method to estimate relative protein expression at the peptide level. We then develop a novel class preserving nested resampling strategy, and a Huber regression based error resampling strategy, to construct a bootstrap partial likelihood estimator of the overall relative expression level of each protein. A significance assessment of the estimated expression levels is obtained through the construction of nested-bootstrap p-values and the specification of a Beta mixture model to locally

estimate a false discovery rate.

Furthermore, we illustrate the application of model based estimation strategies when proteomics data are assumed to be missing at random. We present a multivariate t model to robustly estimate the mean and covariance matrix of an incompletely observed peptide level data set of a given protein. We also look at the same estimation problem, when there is only a single peptide available to quantify a protein's relative expression level. This problem is handled in the context of a bivariate normal model having a monotone missingness pattern, and a hierarchical Bayesian scheme for constructing small sample confidence intervals.

We demonstrate the use of our proposed methodologies on three proteomics data sets. Two of the data sets are derived from the yeast proteome and are technical replicates of each other. These two data sets are mixed in a 1:1 ratio using the SILAC (Stable Isotope Labeling using Amino acids in Cell culture) labeling strategy, and are therefore 'pure null' control samples. That is, since we expect each protein in these two data sets to have a relative expression ratio of one, we have a convenient means of evaluating the performance of our proposed methods, using measures such as the estimated proportion of non differentially expressed proteins, the false discovery rate, and the observed number of false positives and negatives. The proteins in the third SILAC data set are derived from the mammalian cellular proteome, in particular the Hela cell line, where the cells were stimulated with epidermal growth factor for two hours prior to mass spectrometric analysis.

Statistical Methods for Robust Estimation of Differential Protein Expression

By

Sameera R. Wijayawardana
B.S., University of Colombo, Sri Lanka, 2001

Advisors: John J. Hanfelt, Ph.D. and Tianwei Yu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Acknowledgement

I would like to take this opportunity to thank everyone who has supported me through the completion of this degree. I am extremely grateful to my co-advisors John Hanfelt, Ph.D., and Tianwei Yu, Ph.D., for their patience, guidance, and direction. I would also like to thank Amita Manatunga, Ph.D., for her belief in me, encouragement and mentorship throughout my time at Emory. Many thanks also to my committee member Junmin peng, Ph.D., for his helpful comments and suggestions and expert knowledge in proteomics, without which I would not have been able to complete this research.

I must also thank the faculty, staff and my fellow students in the Department of Biostatistics and Bioinformatics who have over the years all helped me to become a good statistician, and more importantly a better person. Last but not least, I would like to thank my entire family for their continued love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The totality of the proteins expressed in a specific cell, given a particular set of conditions, is defined as the *proteome*. The study of the proteomes of organisms is called *proteomics*, a term defined analogously to *genomics* or *metabolomic*, which are concerned with the study of the *genome* and the *metabolome*, respectively.

Proteomics is a relatively recent field (the term was coined as recent as 1997) that is primarily concerned with investigating how proteins are affected by cell processes or the external environment. In particular, *Expression proteomics* is concerned with the analysis of expression and differential expression of proteins and highlighting differences between them under different settings. For example, the protein content and expression levels of a cancerous cell is often different from that of a healthy cell. Certain proteins in the cancerous cell may be up-regulated (more abundant) or down-regulated (less abundant) compared to the healthy cell. Identifying these differentially expressed proteins can lead to the discovery of *bio-markers* which have the potential to assist in early diagnosis and formulation of new therapies, and a better understanding of the cellular level functionality of those proteins.

The actual realization of these goals is not simple. The purification and identification of proteins in any organism is hindered by a multitude of biological and technical factors. Proteomics experiments are high-throughput, often generating giga-bytes of data. Extracting meaningful information from these data is fraught with both technical and statistical data analysis issues related to high variability, robustness, and multiple testing. Fortunately, recent developments in high-throughput proteomics methods such as *tandem mass spectrometry* and quantification by stable isotope labeling have greatly advanced our ability to effectively make use of these large amounts data.

In this dissertation, we develop robust statistical methods, which when taken together form a proteomics data analysis 'pipe-line'. This pipe-line receives the peptide level expression profiles of one or more protein samples at its front-end, and outputs a list of proteins that are significantly up or down regulated.

We begin this chapter with an introduction to the field of proteomics, and to the history and background of high-throughput proteomics methods. In Section 1.4, we briefly review some of the more frequent statistical questions that arise in the context of analyzing proteomics data. In Section 1.5, we discuss some of the analytical challenges posed by the inherent complexity of proteomics data. In Section 1.6, we describe the data sets and problem formulations that motivated our research, and in Section 1.7, we set out these research goals within the statistical framework in which each research question will be addressed.

The second half of this chapter, starting with Section 2, will present relevant background information on the use of Mass Spectrometry (MS) based techniques for identifying proteins, with particular emphasis on labeling based methodologies, and the use of MS based techniques for quantifying protein expression levels. In addition, we present a brief overview of some of the common statistical methodologies used in identifying and quantifying proteins.

1.2 An Introduction to Proteins, the Proteome and Proteomics

The word *protein* is derived from the Greek word prôtos, meaning primary or first rank of importance. As its name implies, proteins perform and regulate a number of vital tasks in all organisms, from regulating the cellular machinery to determining the phenotype. Examples of proteins include whole classes of important molecules: among them enzymes, hormones, and antibodies. There are an estimated $20,000 \sim 25,000$ genes in the human genome. These genes possibly code for as many as one million proteins. This great variety in the number and types of proteins comes from a phenomenon known as *alternative splicing*, where by a particular gene in a cell's DNA can create multiple protein types, based on the demands of the cell at a given time.

The term *proteome*, a portmanteau of *proteins* and *genome*, refers to the entire complement of proteins expressed by a genome, cell tissue or organism. It is larger and in many ways a more complicated entity than the genome. More importantly, unlike an organism's genome which is more or less constant, the proteome regularly differs among individuals, cell types, and within the same cell depending on cell activity, disease, or external stimuli.

Proteomics is the study of the proteome. It was defined by Marc Wilkins in 1994 as "the study of proteins, how they're modified, when and where they're expressed, how they're involved in metabolic pathways and how they interact with one another". In the past, the determination of the set of proteins produced in a cell was done by mRNA analysis. However, it is now known that mRNA is not always translated into protein, and that there is an insufficient correlation between mRNA and protein abundance (Gygi *et al.*, (1999)) [53]. Proteomics on the other hand, attempts to confirm the presence of the protein and provide a direct measure of its quantity present. The

scope of proteomics studies have now evolved from simple biochemical analysis of single proteins to measurements of complex protein mixtures. In addition to protein expression studies, other key areas of proteomics include: the characterization of the 3-D structure of proteins, as illustrated by the 3-D crystalline structure of a protein shown in Figure 1.1, protein/protein and protein/DNA interactions, and the analysis of post-translational modifications.

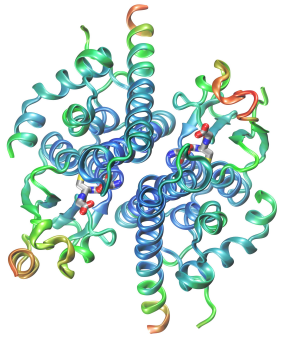


Figure 1.1: 3D crystalline structure of the yeast prion protein Ure2P.

1.3 Mass Spectrometry (MS) based Proteomics

In proteomics, before analyzing the expression levels of proteins in a given sample, the protein complement of that sample needs to be accurately and efficiently identified. This is afforded by a suite of technological tools broadly called *mass spectrometry*.

Mass spectrometry is a powerful technique that is used to identify unknown compounds, to quantify known compounds, and to elucidate the structure and chemical properties of molecules. The introduction of two ionization techniques in the mid-1980s: ESI (ElectroSpray Ionization) and MALDI (Matrix Assisted Laser Desorption/Ionization), transformed MS into an enabling technology in proteomics by allowing ionization of large intact macromolecules such as peptides and proteins. Mass spectrometry based proteomics has emerged as the preferred and most powerful technological approach in proteomics studies. many improvements over the last

decade in both instrumentation and associated computing tools now allow the rapid processing of high-throughput data, leading to accurate and routine protein identification, quantification, and determination of sites of post-translational modification (Aebersold *et al.*, (2003)) [1].

In proteomics, the typical output from a mass spectrometer are spectra that comprise of digitized arrays of the intensities of all discrete mass-to-charge ratios detected over a discrete mass range. A specialized MS procedure known as *tandem mass spectrometry* (MS/MS) is widely used to sequence peptides in real-time during the MS operation. Interpretable MS/MS spectra are usually produced after only about 1-2 seconds of data acquisition time, and the entire process of pre-cursor peptide selection and MS/MS analysis can be fully automated. This allows the high-throughput, large-scale analysis of complex protein mixtures in a fairly rapid and sensitive manner.

The MS method of first ionizing intact proteins using either ESI or MALDI, and then introducing them to a mass analyzer, is referred to as the "top-down" strategy of protein analysis. Conversely, the "bottom-up" strategy first digests protein analytes into smaller peptides, enzymatically or by chemical cleaving, before being fed into the mass spectrometer.

1.4 Statistical Methods in Proteomics Data Analysis

Tandem MS coupled with database searching, has become the de facto standard for identifying and quantifying proteins in complex mixtures. In general, a protein mixture of interest is enzymatically digested, and the resulting peptides are further fragmented through CID (Collision Induced Dissociation). The resulting tandem MS spectrum contains information about the constituent amino acids of the peptides, which in turn provide information about their pre-cursor or parent proteins. The

format of the data that results from MS or MS/MS consists of a two dimensional grid of paired data points of mass-to-charge ratio (m/z) and signal intensity. Figure 1.2, shows an example of a raw MALDI-TOF (Time-of-Flight) MS spectrum. Note that the total number of measured data points is usually extremely large ($\tilde{1}0^6$ for a conventional MALDI-TOF instrument). The primary challenge in quantitative proteomics, especially with respect to *bio-marker discovery*, is making efficient use of this richness of data to find a few peptides/proteins that can distinguish between case and control samples.

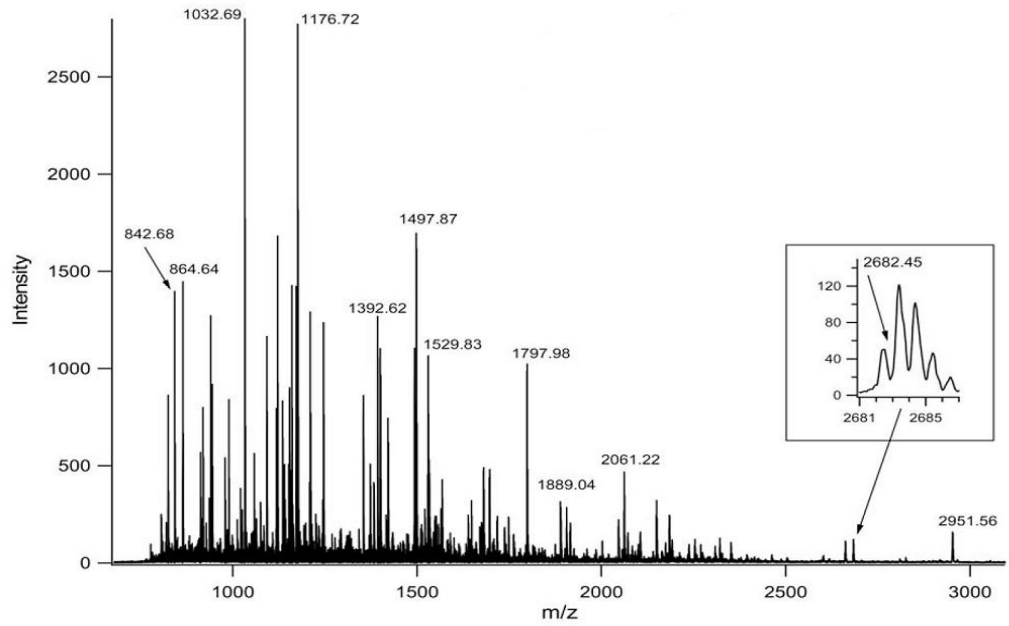


Figure 1.2: An example MALDI-TOF MS spectrum. The horizontal axis denotes the m/z (mass/charge) ratio and the vertical axis denotes the intensity value.

The typical work flow of analyzing proteomics data consist of the following steps: [1] preprocessing of the raw spectral data; [2] sequence identification; [3] translation of the peptide information into protein expression levels; and [4] determining whether the observed protein expression changes between two or more samples are accurate, repeatedly observable, and are statistically significant under an assumed null hypothesis. The focus of our research is only on steps 3 and 4. However, for completeness,

we give a brief review of current statistical methods used in steps 1 and 2, in Sections 2.4 and 2.5.

A number of important statistical questions that arise in steps 3 and 4 are: how should systematic shifts in MS/MS profiles across multiple samples be addressed?; should ratios be calculated from the MS, or MS/MS level data?; how should peptide ratios be combined to calculate protein levels?; should all peptides be treated the same in calculating protein levels?; how should observed intensity levels be adjusted to correct for background noise?; and how should account for multiple testing?. Any solution or set of solutions that effectively answers any or all of these questions would be a significant advancement in the field of quantitative proteomics. There have been many attempts to answer each of these questions on their own, and a few attempts at addressing more than one of them within a unified framework. However, to our knowledge there is no one method or a collection of methods that has garnered acceptance as the 'gold standard'. We present a brief review of existing statistical methods that have been widely adopted by the proteomics community in Chapter 2.

1.5 Proteomics - Analytical Challenges

The relative complexity of the proteome compared to the genome or the metabolome, has so far meant that the development of proteome-wide analyses technologies has lagged and proven more difficult than the development of DNA analyses technologies. A number of other reasons contribute to the relative complexity of proteomics data. For a start, the basic alphabet for encoding proteins consists of twenty amino acids, whereas there are only four different nucleotides in the alphabet of DNA. Also, many proteins undergo modifications after they have been synthesized. These *post-translational* modifications (e.g., phosphorylation, ubiquitination, methylation, etc.) have the potential for profoundly affecting the functional activity of the protein.

There are also a number of analytical challenges posed by the technical difficulties associated with MS and MS/MS based high-throughput proteomics methods. The most intractable of these being the *ionization efficiency* of the target molecules. Basically, the ionization efficiency of a particular molecule in a droplet eluting from the LC system is directly related to its hydrophobicity and its susceptibility to receiving a charge. For example, in ESI (Electrospray Ionization) and MALDI (Matrix-Assisted Laser-Desorption Ionization), ionization efficiency can be quite variable for peptides of different sequence, identical peptides from different MALDI spots, or even for ESI under different HPLC (High Performance Liquid Chromatography) conditions. This means that it is virtually impossible to perform tandem MS on every ion presenting in a chromatographic window at any given time. Furthermore, while a specific peptide ion is being selected for CID; other co-eluting ions cannot be selected for tandem MS. These excluded ions may or may not be selected for tandem MS in subsequent iterations of the process, leading to a set of ions that will be completely undetected during the experiment. Karas *et al.*, (2003) [62] demonstrated these and other factors impairing ionization efficiency and quantification using model metabolites. However, to our knowledge; there are currently no methods available that can accurately predict the severity of ion suppression in any given MS run.

The quality of peptide MS/MS spectra is dependent on a number of factors: the sequence location of amino acids; amino acid side chain basicity; amino acid side chain structure; and charge state of the fragmented peptide ion. Also, one of the prerequisites of sequencing by MS/MS is that the peptide undergoes complete, or near complete fragmentation. If there is incomplete fragmentation, or the products of other fragmentation pathways constitute a large portion of the spectra, the data becomes difficult to decipher. Fortunately, the advent of high-resolution, high-speed mass spectrometers have improved things greatly in this context. These high-sensitivity instruments can separate signal from background as well as from co-eluting compounds

with similar mass, and can capture signals associated with much weaker peaks while quantifying the stronger peaks with greater accuracy.

In addition to the above complexities in protein samples and MS based technologies, there are questions of robustness of quantitative methods used in the analysis of peptide and protein data. Statistical validity of reported quantification results is often based on the standard deviations obtained in separate runs, from different peptides quantifying the same protein, and for a single peptide, from consecutive scans of co-eluting peptides or peptide pairs. Ideally, all sources of variability in quantification should be tracked, and a compound standard deviation determined. This compound standard deviation should account for instrument error, sample preparation error, and the inherent variation in biological samples. The biological sample variation often goes un-investigated even though Molloy *et al.*, (2003) [88] had shown a large contribution of the biological sample variation on top of instrumental errors, that may reach a total error of up to 70% percent CV in some proteomics applications.

Furthermore, many of the statistical challenges in proteomics are exacerbated by the inconsistency of the detected peptide complement between different MS runs of the same protein sample. The failure to identify or detect a peptide in any given MS run, does not always indicate the absence of its pre-cursor protein. Therefore, it is essential that we do not rely entirely on the boolean nature of peptide identifications in formulating a robust protein identification or quantification scheme. Stable isotope labeling methods, described in Section 2.3, can serve to reduce the complexity of protein mixtures by minimizing the level of systematic and biological variation introduced during sample preparation. The caveat being that, if we only use standard post-hoc mean expression level comparisons to analyze the data, we then lose potentially useful information that could have been associated with the biological sample variation.

1.6 Motivating Examples

Our motivation behind this research is the development of robust and flexible methodologies for identifying differentially expressed proteins. Towards this end, we make use three data sets: a mammalian cellular proteomics data set that is publicly available, and two proprietary yeast data sets collected from a control experiment.

1.6.1 Data Sets

The best benchmark for evaluating the efficiency and effectiveness of a particular methodology for identifying differentially expressed proteins is to apply that methodology on a data set with known behavior, or on a control sample. In essence, all protein expression ratios derived from these control data sets should be known in advance, with allowance made for random noise contamination. In this work, we make use of LC-MS/MS labeling based protein expression data derived from a yeast protein mixture. This mixture consists of *light* and *heavy* isotope labeled yeast samples that were mixed in a 1:1 ratio and digested in solution using trypsin. Each of six resulting gel fractions were then run in duplicate as technical replicates. Peptide/Protein identifications were made using the SEQUEST algorithm (cf. Section 2.5.1). We note here that the use of specially designed control samples to validate statistical methods is quite rare in proteomics.

The third data set we use was first published in Cox and Mann (2007) [28]. The data come from a SILAC experiment, where HeLa cells were stimulated with EGF (Epidermal Growth Factor) for 2 hours prior to mass spectrometric analysis. The signal pairs (*heavy*, *light*), correspond to the EGF stimulated, and control samples, respectively. The combined protein mixture was digested in solution using trypsin, and the resulting peptides were separated into 24 gel fractions. Each gel fraction was then purified and analyzed using LC-ESI (liquid chromatography - electrospray

ionization) combined with MS/MS. Peptide/protein identifications were made using the MASCOT algorithm (cf. Section 2.5.2).

1.7 Proposed Research

Typical proteomics studies yield large data sets that consist of multiple layers of data corresponding to: MS/MS scans $\rightarrow$ peptides $\rightarrow$ proteins. Primarily, our focus is on the top most layer of data, i.e., the proteins. The goal of statistical analysis of such data is to quantify the expression level of proteins, by making maximum use of the data available at all layers. Effectively combining data across multiple layers is hampered by a number of factors: high variability in protein expression levels; inherent variability in mass spectrometry based methods; lack of replicate data in most cases; and the simultaneous testing of thousands of hypotheses.

Current statistical methods in proteomics have lagged behind the technological evolution of MS and other 'omics' methodologies. This is partly due to the relative complexity of the proteome compared to the genome. But, to a great extent is also due to a dearth of methods that explicitly address robustness and repeatability of statistical results. This lack of robustness can be attributed primarily to the ubiquitous use of unvalidated parametric assumptions about the distribution of protein expression levels, and/or the failure to explicitly account for the error variation present at each layer of the data. The use of empirical hierarchical Bayes methods that capture variability within and between each layer of data, and nonparametric methods that estimate the distribution of protein expression levels from the data itself, may help to mitigate these issues.

A vast majority of labeling based proteomics experiments are run as non-replicated experiments. This absence of repeat measures means that, when there are missing values, a complete case only analysis of the data is highly inefficient. To date, the

extent to which this issue has been addressed is limited to simple multiple imputation methods that rely on the missing completely at random (MCAR) assumption. Analyses based on a more rigorous and realistic setting, i.e., based on robust model based estimation approaches under the less restrictive missing at random (MAR) assumption, may provide a more solid platform for point estimation and/or conducting inference on protein expression levels.

1.7.1 Robust Estimation of Labeling Based High-Throughput Relative Protein Expressions

When the distribution of the relative protein expression levels contain one or more of the issues: non-Gaussian tails, regions of data sparsity, excess kurtosis, or asymmetry, standard distributional approximations such as the $N(0,1)$, or $N(\mu,\sigma^2)$, do not perform adequately. We propose to develop methodologies that make maximum use of the available data; account for both within and between variations of each data layer; and robustly capture the empirical distribution of the relative protein expression levels. To this end, we work within the framework of finite mixture models, and maximize the flexibility in data modeling by empirically inferring the full shape of the class-conditional probability distribution of each mixture component.

Furthermore, we propose to investigate the utility of skew normal, skew Student's t, and Generalized Hyperbolic (GH) distributions, as suitable approximations for the distribution of relative protein expression levels, when the said distribution suffers from one or more of the afore mentioned irregularities. To our knowledge, this is the first work to explore the utility of these distributions from an 'omics' data analysis perspective. The fitting of these distributions is rigorously controlled for both goodness-of-fit; the number of fitted mixture components; and for stable estimates of mixture parameters. We discuss these methodologies using a step by step data analysis procedure; where for each protein, we start with estimating its relative expression

ratio, and end with calculating its posterior probability of being non-differentially expressed.

1.7.2 Resampling Based Methods for Identifying Differentially Expressed Proteins Using XIC Area

The observed protein profiles of two technical replicates of the same protein mixture can differ both with respect to the number of proteins identified and their expression levels. The latent error mechanisms leading to these discrepancies are not well understood. However, it is believed that experimental and/or physicochemical properties of the constituent peptides of the proteins themselves might account for much of this variability. This variability can be reduced to some degree through careful quality control at each stage of the experiment; by removing 'unreliable' data points; and by using a more stable method to quantify the peptide level relative expression ratios. In this context, we propose to use a bivariate mixture model based clustering analysis to cluster the data into 'reliable' and 'unreliable' groups; and to use Savitzky-Golay filtered ion current (XIC) profile area ratios to quantify the relative expression ratio of peptides. We propose to account for the latent error mechanisms by a novel method that relies on drawing weighted nested-bootstrap samples, and using a bootstrap partial maximum likelihood estimator (BPMLE) to estimate the overall relative expression ratio of each protein. We also develop a novel method that constructs a model - based BPMLE, through resampling the residuals from a robust Huber regression of peptide level relative expression ratios. Finally, we assign a significance value to each estimated protein expression ratio using a nested-bootstrap p-value calculation, and then control the false discovery rate locally by modeling the p-value distribution as a two component Beta mixture model.

1.7.3 Estimating Relative Protein Expression Levels from Incomplete Data

Most statistical analyses of proteomics data do not address the issue of missing values beyond simple missing value imputation schemes, under the MCAR assumption. We believe that with proteomics data, MCAR is both too restrictive and unrealistic. We propose model based estimation strategies that are based on the less restrictive MAR assumption in two settings. First, we look at robustly estimating the true relative expression ratio of a protein based on incompletely observed peptide level data, using a multivariate t model. Secondly, we will look at a valid estimation strategy when only one peptide is available to uniquely identify a protein, using a bivariate normal model with a monotone missingness pattern. Since the sample sizes associated with single peptide proteins is relatively small, inferences based on the inverse of the observed information matrix are not ideal. Therefore, we also propose a Bayesian scheme for constructing confidence intervals for the estimated parameters under the bivariate normal model. We believe that our work is the first to investigate the use of these types of model based approaches for estimating the relative protein expression under MAR.

Chapter 2

Background

In this chapter, we present briefly, relevant background information on some key MS based methodologies; data generating mechanisms; and associated statistical techniques for pre-processing and identification of protein data.

2.1 Mass Spectrometric Methods for Protein Identification

In proteomics, the interpretation of MS or tandem MS spectra requires the combined use of many techniques. The standard strategy for identifying an unknown compound is to compare its observed mass spectrum against a database of theoretical mass spectra. In general, database searching methods compare the experimentally observed tandem MS spectra with features predicted for hypothetical spectra from candidate peptides in the database, and assigns a score that is proportional to the degree of matching. Because of the large number of MS/MS spectra that are generated in a typical MS-based protein analysis, highly automated searching methods are an absolute necessity (Eng *et al.*, (1994)[44], Perkins *et al.*, (1999)[95]). These high-throughput automated searches were only made possible with the development of

computer algorithms that can correlate peptide fragmentation spectra with sequence databases using fragment ion pattern and mass of the parent (or precursor) ion as input for sequence database searching. The best known of these types of algorithms are the SEQUEST (Eng et al. (1994)[44]), MASCOT (Perkins *et al.*, (1999)[95]), PeptideProphet (Keller *et al.*, (2002)[63]) and ProteinProphet (Nesvizhskii *et al.*, (2003)[89]). A brief discussion of these algorithms is given in Section 2.5.

In a typical MS run, it is not unusual to have peptides identified with high scores that are identified simply due to random sequence matching. These matches are False Positives (FPs). In recent years, many researchers have adopted an approach based on the use of a "target-decoy" database search for peptide identifications. In this approach, spectral data are searched against a protein sequence database (the target) and a database comprised of reversed or random amino acid sequences (the decoy). The number of positive identifications from the decoy database is used to estimate the expected number of FPs in the target database search, under the assumption that the probability of an incorrect peptide spectrum match is the same for both target and decoy databases. This assumption underlying the target-decoy strategy was demonstrated to be valid by Elias and Gygi (2007)[43].

An alternative approach to database searching is *de novo* MS/MS sequencing (Taylor and Johnson (1997) [114], Dančík et al. (1999)[30], Chen et al. (2001)[25]), which attempts to derive the peptide sequence directly from tandem MS data. The primary advantage of *de novo* sequencing is that it can handle situations where a target sequence is not found in the protein database being searched. However, the utility of the approach is dependent upon the quality of tandem MS data, such as the level of noise and the number of predicted fragment ion peaks that are observed.

2.2 Mass Spectrometric Methods for Protein Expression Profiling

Accurate quantification of proteins of different samples, such as diseased vs. healthy tissue, plays a vital role in proteomics. This quantification can be either absolute or relative. Absolute quantification involves the determination of the exact quantity of a protein (peptide) in a given sample. Relative quantification involves estimating the quantity of a protein (peptide) in relation to the quantity of the same protein (peptide) in a different sample, or the same sample in an altered state.

The absolute signal intensity of a peptide ion does not always reflect the true abundance of the peptide in the analyzed sample. This is due to differences in ionization efficiency, ion suppression, and the inconsistency in the detection of analytes across different MS runs. However, differences in the relative peak intensities of the same analyte do accurately reflect differences in its expression. Various methodologies have been developed to quantify relative changes in protein abundance between samples. These methods fall broadly into two categories: label-free methods, and methods based on stable isotope labeling. In this work, we focus exclusively on the latter category.

2.3 Quantification based on Stable Isotope Labeling

A reliable internal standard is often required to normalize the quantitative variations across different MS measurements. Such a standard should ideally be as similar as possible to the analyzed peptide both chemically and physically. A known analyte of unknown concentration can be absolutely quantified by spiking in an isotopic variant of the same analyte to the sample of interest at a pre-determined concentration. As

the two compounds typically co-elute, a comparison of the highest peaks or the peak areas of the two compounds, will enable the absolute quantification of the analyte.

In proteomics, a similar quantification strategy is based on the addition of an isotopically labeled synthetic peptide to the digestion reaction as a internal standard or reference. Quantitative protein profiling is accomplished by comparing a reference protein sample to a second sample containing the same proteins that are labeled with heavy stable isotopes. Theoretically, all the peptides in the mixture of the two samples then exist in pairs of identical sequence but different mass. As the peptide pairs have the same physicochemical properties, they are expected to behave identically during isolation, separation, and ionization. Thus, the ratio of intensities of the lower and higher mass components provides an accurate measure of the relative expression of the peptides (and hence of their parent proteins) in the original protein samples. Since the ion intensities of each pair is measured simultaneously, much of the systematic variations present in different MS runs along with variabilities in measured intensities due to; ion-suppression, dynamic exclusion, and differing amounts of injected sample, are eliminated. This allows an 'apples-to-apples' comparison of intensities, resulting in more accurate profiling of the relative protein expressions.

Currently, several methods exist for stable isotope based quantification that differ mainly in the technical process used in the actual labeling. The labeling is done either: *in vivo* through metabolic incorporation, where labeling of the peptide/protein is done by growing cells in a media enriched with stable isotope-containing anabolites; or *in vitro* through chemical incorporation of the stable isotopes and relying on the use of re-agents for chemical modification of proteins in a site-specific manner.

2.3.1 In Vitro Labeling via Chemical Incorporation

In vitro labeling involves incorporation of stable isotopic tags at selective sites on peptides via *in vitro* chemical reactions. These site-specific incorporations include

labeling of target peptides at their amino-(N-) or carboxyl-(C-) termini or on specific amino acid residues, such as *cysteine*, *lysine*, *tyrosine* etc. The most popular of these methods to data is ICAT (Isotope-Coded Affinity Tags), first introduced by Gygi *et al.*, (1999) [53]. In a typical ICAT experiment, proteins from two samples are labeled at their *cysteine* residues with either isotopically light (^{1}H) or heavy (^{2}H) ICAT reagents. The light- and heavy- labeled samples are then combined, proteolyzed to peptides, fractionated by multidimensional chromatography and then quantitatively analyzed by MS. Then the ratio of ion intensities of the co-eluting ICAT-labeled peptide pairs (with a mass shift of 8 Da per labeled *cysteine* residue) allows the quantification of the relative expression of the peptides in the two samples. A subsequent MS/MS scan is used to generate CID spectra that will be used for protein identifications.

2.3.2 In Vivo Labeling via Metabolic Incorporation

In vivo labeling approaches involve metabolic incorporation of stable isotopes into proteins of cells grown in special media containing these isotopes. We describe here only one such approach, SILAC, that has recently garnered popularity due to the high level of predictability of the mass shifts generated by the approach.

SILAC (Stable isotope labeling with amino acids in cell culture) is a straightforward approach for *in vivo* incorporation of a 'label' into proteins through the metabolic incorporation of given 'light' or 'heavy' forms of amino acids with substituted stable isotopic nuclei (e.g., ^{2}H, ^{13}C, ^{15}N). In a typical experiment, two cell populations are grown in culture media that are identical except for the fact that one population contains a 'light' and the other a 'heavy' form of a particular amino acid (e.g., ^{12}C and ^{13}C labeled L-lysine). When the labeled analog of an amino acid is supplied to cells in culture instead of the natural amino acid, it is incorporated into all newly synthesized proteins. After a number of cell divisions, each instance of this

particular amino acid will be replaced by its isotope labeled counterpart. Since there is virtually no chemical difference between the labeled amino acid and the natural amino acid isotopes, the cells behave exactly like the control cell population grown in the presence of the normal amino acid.

There are several advantages of the SILAC approach compared to chemical incorporation methods. The accuracy of the approach is only limited by the quality of the peptide signals observed; and the level of systematic and biological variation introduced during sample preparation and MS processing is considerably less. The labeling process is straightforward and highly efficient since usually 100% of the sample is available for MS analysis. SILAC also allows unlabeled and labeled samples to be combined prior to lysis of the cells and to be treated as a single sample in all subsequent steps. This affords the experimenter the flexibility to choose any method of protein or peptide purification (after enzymatic digestion), without introducing additional error components into the final compound standard error of the protein expression levels.

The typical work flows of the ICAT and SILAC strategies are shown in Figure 2.1.

2.4 Statistical Methods in Preprocessing Proteomics Data

The typical first step in proteomics data analysis is the removal or reduction of systematic artifacts introduced by the instrumentation and experimental protocols, and by the random background fluctuations produced by chemical and electronic noise. Secondly, spectra from multiple MS runs need to be aligned with respect to retention time of the compounds. These prerequisite data adjustment steps are known as *preprocessing*, and at a minimum, involve the following steps: spectrum calibration; base-line correction; smoothing; peak identification; and intensity normalization and

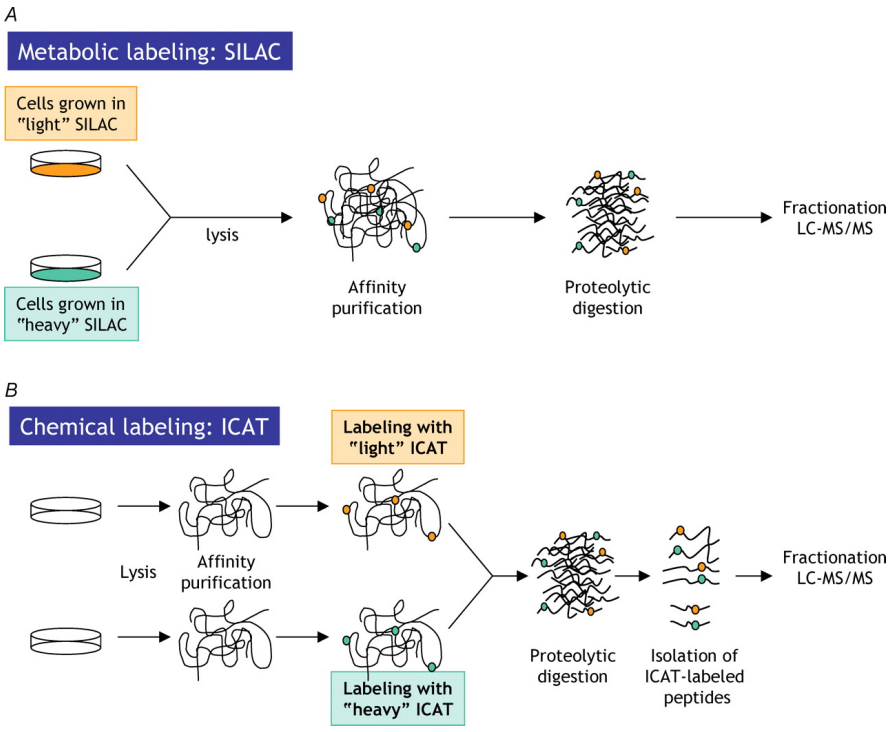


Figure 2.1: Stable isotope labeling - (A) SILAC and (B) ICAT flow diagrams.

peak alignment.

Preprocessing generally starts with aligning individual spectra based on the maximum observed intensity of an internal calibrant. Removal of high frequency background noise is usually achieved with some form of local smoothing. For example, with MALDI data, Wu *et al.*, (2003)[119] uses a local linear regression method to estimate the background intensity values, and then subtract the fitted values from the local linear regression result. Baggerly *et al.*, (2003)[13] consider a semi-monotonic baseline correction method in their analysis of SELDI data. Liu *et al.*, (2003)[76] compute the convex hull of the peak spectrum, and subtract the estimated convex hull from the original spectrum to get the baseline noise-corrected spectrum.

In the following sections, we give a brief overview of peak selection and alignment methods, which are arguably the two most important preprocessing steps.

2.4.1 Peak Selection

Most peak selection methods are based on simple heuristics. In general, spectral peaks are identified as the local maxima in predefined neighborhoods of intensity values. It is also common that these local maxima are required to be higher than the average intensity level of its neighborhood by a certain margin. For example, Liu *et al.*, (2003)[76] declare a point in the spectrum as a selected peak if the intensity is a local maximum; its absolute value is larger than a threshold; and the intensity is larger than a threshold times the average intensity in the neighborhood surrounding the peak. Coombes *et al.*, (2003) [27] consider two peak identification procedures. In the *simple peak detection* procedure, local maxima are identified, and nearby maxima that likely represent the same peptides are merged after filtering out local maxima that are likely to be noise. In the *simultaneous peak detection and baseline correction* procedure; *simple peak detection* is first used to obtain a preliminary list of peaks, after which a baseline is calculated by removing these preliminary candidate peaks. A final set of peaks is selected by iterating the above steps, until all peaks selected are above a chosen signal-to-noise ratio threshold.

The success of these peak selection methods depend largely on whether or not the noise model is correctly specified. Often, a particular noise model is assumed without attempting to validate the assumed model. Coombes *et al.*, (2003)[27] define noise as the median absolute value of intensities. Satten *et al.*, (2004)[108] use the negative part of the normalized MS data to estimate the variance of noise. Yasui *et al.*, (2003a)[120] argue that binary peak/non-peak data is more useful than the absolute values of intensities in quantifying noise, based on the observation that the measured intensities are susceptible to substantial measurement error. Wavelets based approaches for de-noising have also been proposed in Coombes *et al.*, (2004)[**?**], and Randolph and Yasui (2006)[99].

All peak selection algorithms are limited by the resolution of the MS data and the

consequent overlapping effect of neighboring peaks. There is also the issue of false positive and false negative selections. Yasui *et al.*, (2003b)[121] address this issue by adding a peak width constraint. Randolph and Yasui (2004)[99]) choose a specific scale level after a wavelets based decomposition of the original MS data. In the case of high resolution data, Yu *et al.*, (2004)[122]) propose that more than one isotopic variant of a peptide peak should be present before a spectral peak can be selected as originating from true peptide ionization.

2.4.2 Peak Alignment

After a set of suitable spectral peaks have been selected from multiple data sets, they need to be aligned before they can be compared against each other or combined to form an average peak profile. The variations in peak location between data sets generated by the same MS protocol have been well demonstrated in several studies (Torgrip *et al.*, (2003)[117], Eilers (2004)[42]). According to Yu *et al.*, (2006)[122], this variation exists even among the selected peak profiles of technical replicates, i.e., replicates of the same sample run under similar conditions. The reasons for this variability are not well understood and are quite often not of primary interest, at least as far as the functionality of the peak alignment algorithms are concerned.

Current peak alignment algorithms range from fairly simple to extremely complicated. Johnson *et al.*, (2003) assumes that peak variation is less than the typical distance between peaks and uses a closest point matching method for peak alignment. Yu *et al.*, (2006)[122] uses the same concept of peak distance to address the alignment of multiple peak sets. Coombes *et al.*, (2003)[27] align spectral peaks by pooling the list of detected peaks that differed in location by three clock ticks in the retention time axis or by 0.05% of the mass. Yasui *et al.*, (2003a)[120] extrapolate each peak to its local neighborhood with a peak width equal to 0.4% of the m/z value of the middle point based on the observation that the shift of peaks is approx-

imately 0.1% to 0.2% of the corresponding m/z value. In a separate study, Yasui *et al.*, (2003b)[121] first calculates the number of peaks in all samples allowing certain shifts, and selected m/z values with the largest number of peaks. Next, this set of peaks is removed from all spectra and the procedure is iterated until all peaks are exhausted from all samples. Tibshirani *et al.*, (2004)[115] propose the use of complete linkage hierarchical clustering to cluster peaks. All peaks that fall into the same cluster are considered to represent the same peak. Randolph and Yasui (2006)[99] uses wavelets to represent the MS data in a multiscale frequency domain and use a coarse-to-fine decomposition method to first align peaks at a dominant scale and then refine the alignment of other peaks at a finer scale. Eilers (2004)[42] propose a parametric warping model with polynomial or spline functions to align chromatograms by adding calibration sequences into chromatograms. Peak alignment approaches based on dynamic programming (Nielsen *et al.*, (1998)[91], Torgrip *et al.*, (2003)[117]) have also been proposed.

2.5 Statistical Methods in Identification of Peptides/Proteins

Most algorithms for protein identification from MS/MS spectra consist of the following three elements (Bafna and Edwards (2001))[12]: [1] *interpretation*, where the input MS/MS data are interpreted; [2] *filtering*, where the interpreted MS/MS data are used as templates in a database search to identify a set of candidate peptides; [3] *scoring*, where the candidate peptides are ranked with a score.

In the following sections, we present a brief review of the statistical methods employed by two of the most popular peptide/protein identification algorithms.

2.5.1 SEQUEST (Eng *et al.*, (1994))

SEQUEST® is perhaps the most popular sequence identification algorithm in current use. The algorithm first creates a list of peptide masses, isobaric to the observed mass on which CID was carried out, by searching the database of choice for possible amino acid sequences that could have generated peptide masses to match the mass of the parent peptide. For each of these candidate peptides, the algorithm generates a theoretical CID mass spectrum that is then compared to the observed fragment ion spectrum using a cross-correlation algorithm. Similarity of the theoretical spectrum to the observed is quantified using both a correlation coefficient and a correlation factor. The correlation coefficient represents the quality of the match while the correlation factor represents the difference between the best-matched peptide and the next possible candidate. Each comparison is then ranked relative to all other possibilities, based on the score of the correlation coefficient, the correlation factor, and other parameters such as the number of fragment ions predicted versus found. However, since the algorithm does not calculate a probabilistic significance for the cross-correlation score, it is not possible to determine the probability that the top-ranked match was not simply the result of random chance. Based on empirical evidence, Eng *et al.*, (1994) [44] suggest that a difference greater than 0.1 between the normalized cross-correlation functions of the first- and second-ranked peptides indicates a successful match between the top-ranked theoretical peptide sequence and the observed spectrum.

2.5.2 MASCOT (Perkins *et al.*, (1999))

Unlike SEQUEST, the MASCOT® algorithm is based on a probability-based scoring scheme. The probability that a match between the experimental MS/MS data and each sequence database entry is a chance event is calculated and the match with the lowest probability is reported as the best match. MASCOT iteratively searches for

the set of the most intense ion peaks that provide the highest score (reported as -10 $\log(P)$, where P is the probability of the match resulting from a chance event). In addition, MASCOT considers many other factors in its probability calculations: the number of missed cleavages; quantitative and non-quantitative modifications; mass accuracy; the particular ion series to be searched; and peak intensities. Perkins *et al.*, (1999) [95] suggest that the validity of the MASCOT probabilities be tested by repeating the search against a randomized sequence database and/or by comparing the MASCOT results with those obtained via the use of other search engines.

2.5.3 Other Methods

Mann and Wilm (1994)[78] propose a peptide sequence tag approach to extract a short, unambiguous amino acid sequence from the peak pattern that, when combined with the mass information, infers the composition of the peptide. PeptideProphet (Keller *et al.*, (2002)[63]) and ProteinProphet (Nesvizhskii *et al.*, (2003)[89]), are two methods that validate the peptide and protein identifications using robust statistical models. After scores are derived from a database search, PeptideProphet models the distribution of these scores as a mixture of two distributions, where the two mixture components correspond to the distributions of the correct and incorrect matches. ProteinProphet takes as input the list of peptides and probabilities from PeptideProphet, adjusts the probabilities for observed protein grouping information, and then discriminates correct from incorrect protein identifications.

Chapter 3

Robust Estimation of Labeling Based High-Throughput Relative Protein Expression

3.1 Introduction

Once a list of putative proteins have been identified, the next step in data analysis is the calculation of the expression level of each of the identified proteins. Recent advances in the field of quantitative proteomics have led to the development of many algorithms and methods that deal with this issue. These methods all have different advantages and disadvantages; and are often better suited to a specific setting, thereby reducing their generalizability across multiple proteomics platforms.

There are two broad categories of estimation methods in quantitative proteomics: relative, and absolute. Absolute quantification methods attempt to measure the absolute expression level of a protein using one or more characteristic peptides unique to that protein (Gerber *et al.*, (2003)[49], Beynon *et al.*, (2005)[18], Anderson *et al.*, (2004)[4]). In this work, we focus on relative quantification methods, with particular

emphasis on estimating the relative expression of proteins originating from labeling based proteomics experiments. Due to availability and accessability, we have chosen to work with proteomics data from SILAC experiments. However, the developed methodologies are easily adaptable to data from other stable isotope labeling strategies such as iTRAQ (Ross *et al.*, (2004)[105], and ICAT (Gygi *et al.*, (1999)[53]).

Previous studies analyzing proteomics data have often relied on techniques which do not account for experiment-specific data variabilities. For example, applying a universal fold-change threshold to identify differentially expressed proteins (Blagoev *et al.*, (2003)[20], Ong *et al.*, (2003)[92]). Other studies analyze differential protein expression without considering the confidence level of each estimated expression level(Han *et al.*, (2001)[54], Ranish *et al.*, (2003)[100], Blagoev *et al.*, (2003)[19]), or without fitting a distribution to all expression levels (Lin *et al.*, (2006)[68], Mertins *et al.*, (2008)[87]). A criticism of these types of analyses is the failure to appropriately isolate the expression levels that correspond to the truly differentially expressed or non-differentially expressed proteins. Another criticism is the use of Bonferroni or q-value type multiple testing adjustments, which are known to be too conservative.

In our work, we explore the application of empirical Bayes hierarchical modeling of SILAC data within the framework of finite mixture models. We adjust for multiple testing by controlling the false discovery rate locally, using an approach first proposed by Efron (Efron (2002)[37], Efron (2008)[39]) in the context of gene expression analyses. In our previous research, we found that these methods, off the shelf, could not robustly model proteomics data, when the data contained non-Gaussian tails, regions of data sparsity, excess kurtosis, or were asymmetrically distributed. Non-Gaussian tails and excess kurtosis are the norm with proteomics data due to the presence of extreme observations and the fact that a significant majority of proteins have a relative expression ratio of one. Asymmetry is typically the by product of discrete errors associated with sample preparation and mass spectrometric processing of proteomics

data. At the sample preparation stage, the main discrete error sources are the differences in the level of *trypsin* digestion and the efficiency of incorporating isotopic tags. At the mass spectrometric level, the significant discrete error sources are ionization efficiency and ion suppression. In addition, asymmetry could result if of the proteins that are differentially expressed, a majority are down-regulated with only a few that are up-regulated, or vice versa. This would lead the distribution of the calculated expression ratio statistic to be skewed to the left or to the right.

In subsequent sections of this chapter, we develop methodologies for both quantification and significance assessment of relative protein expression, when data come from non-replicated experiments. In fact, most labeling based proteomics experiments are not run with replicates due to cost and time constraints. Therefore, we discuss methodologies that make maximum use of the available data to efficiently estimate relative expression levels; and maximize the flexibility to infer the true shape of the mixture components beyond the capacity provided by standard symmetrical distributions.

3.2 Identifying Differentially Expressed Proteins in Non-replicated Experiments

3.2.1 Data Structure

In SILAC experiments, isotopically-labeled amino acids are used to 'pair' peptide signals arising from two or more samples. The hierarchical structure of the resultant data can be outlined as follows. Let $\boldsymbol{F}^1, \ldots, \boldsymbol{F}^n$ denote the data vectors originating from n gel fractions that are analyzed in an experiment. At the gel fraction level, the structure of the data can be described as follows. Assuming that we observe n_i individual proteins in gel fraction i, we have $\boldsymbol{F}^i = \{Pr^{i1}, \ldots, Pr^{in_i}\}$; $i = 1, \ldots, n$,

where Pr_j^i denotes the j^{th} protein observed from the i^{th} gel fraction. Pr^{ij} in turn represents the set of n_{ij} individual peptides $\{Pp_1^{ij}, \ldots, Pp_{n_{ij}}^{ij}\}$ that are derived from Pr_j^i. However, we do not observe the same set of proteins in each gel fraction. Similarly, the derived peptide complement of a protein will not be the same from one gel fraction to another.

Let m be the total number of individual proteins detected over all n gel fractions, and let the set of these proteins be $\boldsymbol{Pr} = \{Pr^1, \ldots, Pr^m\}$. For a given protein $Pr^j \in \boldsymbol{Pr}$; $j = 1, \ldots, m$, let m_j be the total number of individual peptides derived from that protein across all gel fractions, represented by the set $\boldsymbol{Pp}^j = \{Pp^{j1}, \ldots, Pp^{jm_j}\}$. Now for each $Pp^{jk} \in \boldsymbol{Pp}^j; k = 1, \ldots, m_j$; let m_{jk} be the number of times peptide Pp^{jk} is detected across all gel fractions. In proteomics the sum, $\mathrm{SC}_j = \sum_{k=1}^{m_j} m_{jk}$ is known as the *spectral count* of protein j. Furthermore, if we denote the r^{th} occurrence of the peptide Pp^{jk} as $Pp^{jkr}; r = 1, \ldots, m_{jk}$, then throughout the full duration of the scan in which Pp^{jkr} is detected, we observe m_{jkr} signal intensity pairs $(\boldsymbol{l}^{jkr}, \boldsymbol{h}^{jkr}) = \{(l_h^{jkr}, h_h^{jkr}) : h = 0, \ldots, m_{jkr}\}$ in which the corresponding light and heavy signals are quantified. Each of these m_{jkr} data pairs provide an independent estimate of the relative expression ratio for the r^{th} occurrence of peptide Pp^{jk}.

The observed ion intensities for a given peptide level scan are assumed to be composed of three parts: the true intensity corresponding to the light or heavy isotopologue, a background intensity that is uniformly present during the scan, and a random noise component. Then for the r^{th} occurrence of peptide k of protein j,

$$\boldsymbol{l}_{Observed}^{jkr} = \boldsymbol{l}_{True}^{jkr} + \boldsymbol{l}_{Background}^{jkr} + \boldsymbol{l}_{Noise}^{jkr} \tag{3.1}$$

$$\boldsymbol{h}_{Observed}^{ijk} = \boldsymbol{h}_{True}^{jkr} + \boldsymbol{h}_{Background}^{jkr} + \boldsymbol{h}_{Noise}^{jkr} \tag{3.2}$$

Assuming that within each full scan: background component is the same for both the light and heavy signals; and both noise components have zero-mean, we can

extract the true signal from the observed through careful preprocessing of the data. If we now denote the post-processed peptide signal pairs as L^{jkr} and H^{jkr}, then the ratios

$$\mathcal{R}^j = \left\{ \frac{L_h^{jkr}}{H_h^{jkr}}; k = 1, \ldots, m_j, r = 1, \ldots, m_{jk}, h = 1, \ldots, m_{jkr} \right\} \tag{3.3}$$

each provide an estimate of the true relative expression of the j^{th} protein between the light and heavy isotopically labeled samples. Note that signal pairs in which $H_h^{jkr} = 0$, have to be removed from this set, since we have chosen to define the relative expression ratio as light/heavy. Now for a given Pp^{jkr}, we represent its estimated relative expression ratio using a summary measure derived from its corresponding m_{jkr} pre-processed signal intensity pairs. The summary measure we consider here is the relative expression ratio derived from the signal intensity pair that has the highest signal-to-noise ratio.

Let, $\boldsymbol{s}^{jkr}$ denote the summary relative expression ratio for the r^{th} occurrence of peptide Pp^{jk}, and let the set of all m_{jk} summary ratios available for peptide Pp^{jk} be denoted by $\boldsymbol{s}^{jk\bullet} = \{\boldsymbol{s}^{jk1}, \ldots, \boldsymbol{s}^{jkm_{jk}}\}$, where the dot notation is used to indicate the aggregation of all data over the index that is being represented by the dot. At the protein level, let $\mathcal{S}^j = \{\boldsymbol{s}^{j1\bullet}, \ldots, \boldsymbol{s}^{jm_j\bullet}\}$ denote the union of all m_j aggregate peptide sets that constitute all available data for protein j. Each peptide set $\boldsymbol{s}^{jk\bullet} \in \mathcal{S}^j, k = 1, \ldots, m_j$ provides m_{jk} separate peptide level relative expression ratios, and $\mathcal{S}^j$, $M_j = \sum_{k=1}^{m_j} m_{jk}$ ratios altogether, for the estimation of the relative expression ratio of protein j. The question then is, how best to utilize $\mathcal{S}^j$, which carries peptide level information, to arrive at a summary estimator $\boldsymbol{g}(\mathcal{S}^j)$ of the true relative expression of protein j, $\boldsymbol{g}(\mathcal{S}^j)=\frac{\boldsymbol{l}_{True}^{j\bullet\bullet}}{\boldsymbol{h}_{True}^{j\bullet\bullet}}$, that has an associated measure of determining statistical significance of the estimated ratio.

3.2.2 A Random Effects Model for Estimating Relative Protein Expression

After aggregation, let the set of all scan level peptide relative expression ratios available from the experiment be $\mathcal{S} = \left\{ \frac{L^{jkr}}{H^{jkr}} : j = 1, \ldots, m;\ k = 1, \ldots, m_j;\ r = 1, \ldots, m_{jk} \right\}$. The form of $\boldsymbol{g}$ that we apply on $\mathcal{S}$ is based on the idea of a one-way random-effects ANOVA model, that is allowed to be both unbalanced and heteroscedastic. More specifically, let the log base-2 transformed elements of the set $\mathcal{S}$ corresponding to protein j be denoted by x_{jkr}, for $k = 1, \ldots, m_j$ and $r = 1, \ldots, m_{jk}$. In quantitative proteomics, log transformation makes intuitive sense since it treats the magnitude of both over and under expression symmetrically around the zero point; i.e., the point corresponding to equal expression. We assume that x_{jkr} is a realization of a random variable X_{jkr} constructed as

$$X_{jkr} = \mu_{jk} + \varepsilon_{jkr} = \mu_j + \beta_{jk} + \varepsilon_{jkr} \tag{3.4}$$

where μ_{jk} is the true relative expression ratio of the k^{th} peptide derived from protein j, and ε_{jkr} is a random error term representing the sampling error of X_{jkr} as an estimate of μ_{jk}. We can further decompose μ_{jk} into the mean μ_j of the population from which the μ_{jk}'s are sampled and the error β_{jk} of μ_{jk} as an estimate of μ_j. In this decomposition, we make the following distributional assumptions: $\beta_{jk} \sim N(0, \sigma_j^2)$, $\varepsilon_{jkr} \sim N(0, \sigma_{jk}^2)$, and the β_{jk} and ε_{jkr} are mutually independent. The two variance components σ_j^2 and σ_{jk}^2 can be thought of as the between peptide and within peptide variances. The term σ_{jk}^2 allows for heteroscedasticity of the within peptide variances derived from the same protein. This is important since the variance of peptide expression levels are typically associated with their mean expression levels, leading to differential variability for different peptides. In the formulation of model 3.4, the choice of the random effect β_{jk} for the peptide level error reflects the fact that each

protein's relative expression is quantified based only on a sample of peptides (i.e., based only on the peptides that were observed in the experiment) out of the population of all peptides that are indicative of that protein. Similarly, the random effect ε_{jkr} reflects the fact that each peptides's relative expression is quantified based only on a sample of expression level pairs out of the population of all expression pairs that are indicative of that peptide.

3.2.3 Estimation of Model Parameters

Our interest is in estimating μ_j, σ_j^2, and σ_{jk}^2 from all available data for protein j. Here we make use of a variant of the minimum norm quadratic unbiased estimation (MINQUE) method, developed by C.R. Rao [101, 102, 103] for the estimation of variance components in random effects models. P.S.R.S. Rao *et al.*, (1981)[104] called this variant the average of the squares residuals (ASR) type estimators and demonstrated that when $\sigma_j^2 > 0$, ASR estimates of σ_j^2 and μ_j have the lowest and second lowest MSE's, respectively, among eight different estimators that includes the maximum likelihood estimator. Furthermore, ASR estimators are always nonnegative and can be easily estimated from the data, without resorting to iterative procedures; which is often the case with MLE, REML or Bayesian estimators.

Let,

$$
\begin{aligned}
l_{jk} &= m_{jk}/(m_{jk}+1) \\
\bar{x}_{jk} &= \sum_{r=1}^{m_{jk}} x_{jkr}/m_{jk} \\
\hat{w}_{jk} &= m_{jk}/(\hat{\sigma}_{jk}^2 + m_{jk}\,\hat{\sigma}_j^2) \\
\hat{W}_j &= \sum_{k=1}^{m_j} \hat{w}_{jk} \\
\bar{\bar{x}}_j &= \frac{\sum_{k=1}^{m_j} l_{jk}\,\bar{x}_{jk}}{\sum_{k=1}^{m_j} l_{jk}}
\end{aligned}
$$

Then the ASR estimators are given by

$$\hat{\sigma}^2_{jk} = \frac{1}{m_{jk}} \sum_{h=1}^{m_{jk}} (x_{jkr} - \bar{x}_{jk})^2 + \frac{l^2_{jk}}{m^2_{jk}} (\bar{x}_{jk} - \bar{\bar{x}}_j)^2 \tag{3.5}$$

$$\hat{\sigma}^2_j = \frac{1}{m_j} \sum_{k=1}^{m_j} l^2_{jk} (\bar{x}_{jk} - \bar{\bar{x}}_j)^2 \tag{3.6}$$

$$\hat{\mu}_j = \frac{\sum_{k=1}^{m_j} \hat{w}_{jk}\, \bar{x}_{jk}}{\sum_{k=1}^{m_j} \hat{w}_{jk}} \tag{3.7}$$

$$Var(\hat{\mu}_j) = \hat{\sigma}^2_{\mu_j} = \frac{1}{\hat{W}_j} \tag{3.8}$$

The ASR estimate of μ_j; i.e., the true relative expression ratio of protein j, is therefore a weighted sum of the peptide specific relative expression ratios, $\bar{x}_{jk}, k = 1, \ldots, m_j$, where the weights are the reciprocal of an estimate of the total variance of $\bar{x}_{jk}$ that encompasses both between and within peptide variance components. Having obtained the estimates $\hat{\mu}_j$ and $Var(\hat{\mu}_j)$, we can construct an approximate test of whether μ_j differs from unity (i.e., equal abundances of the protein j in the two samples), using the test statistic

$$z_j = \frac{\hat{\mu}_j - log_2(1)}{\hat{W}_j^{1/2}} = \frac{\hat{\mu}_j}{\hat{W}_j^{1/2}} \sim f(z) \tag{3.9}$$

where $f(z)$ is the distribution of the test statistic Z. Relatively high or low values of the test statistic indicate proteins whose expression levels are significantly different between the light and heavy labeled samples.

3.2.4 Simultaneous Testing of Relative Protein Expression Levels

Let $\mathcal{Z}$ be the set of all test statistics Z_j; $j = 1, \ldots, m$. The goal is then to identify a relatively small set of interesting non-null proteins, after adjusting for multiple testing considerations. The simplest means of achieving this objective is to set a simple

fold-change cutoff rule (Ong *et al.*, (2002)[92], Blagoev *et al.*, (2004)[20]). In this approach, proteins with a fold-change larger than a pre-defined cut off (e.g., 1.5-fold, or 2-fold) are classified as differentially expressed. While this is an intuitively simple approach; there are a number of limitations of the approach that have been widely discussed in the context of micro-array data (Gusnanto *et al.*, (2007)[51]), which are equally relevant in quantitative proteomics. A more statistically rigorous method uses the Student's t-test and corresponding p-values to identify significant changes in expression level, after a Bonferroni or q-value type multiple testing adjustment (Cox *et al.*, (2008)[29], Chang *et al.*, (2004)[24], Hendrickson *et al.*, (2006)[55]).

Let $f(z)$ denote the distribution of all the z's, and $f_0(z)$ denote the distribution of z's under the null hypothesis of non differential expression. We would expect $f_0(z)$, to be nearly $N(0,1)$. However, the standard normal distribution rarely holds in practice. In our experience, we have encountered data where the distribution of z_j's is better approximated by a $N(\mu, \sigma^2)$, a skewed and/or kurtotic normal or even a distribution with heavier tails than the normal, such as the Student's t. On occasion, $f_0(z)$ can even be multi-modal. In addition $f(z)$, the distribution of all the z's, can also be asymmetric and have longer tails than a normal, especially if the data contain outliers. There is also the issue of sparsity of data towards the tail ends of the distribution of z_j's. This is typically the case when a majority of proteins (typically upwards of 90%) are not differentially expressed and are tightly packed around the center of the distribution with a few non-null proteins spread out thinly at the tail edges. Therefore methods based on fitting a single distribution to the z_j's, without robustly capturing the null and non-null components is likely to be too conservative in identifying significant proteins, or may not identify any significant proteins at all.

Since a large majority of proteomics experiments are not run in replicate, strong parametric assumptions are often the only recourse. However, since in high throughput experiments, m is usually large, it seems unnecessary to restrict oneself to strictly

parametric approaches. In the field of microarray gene expression analyses; empirical Bayes and mixture models based methods have been suggested as alternatives to making strong parametric assumptions about the distribution of z_j's (Efron (2002)[37], Pan *et al.*, (2003)[94], Allison *et al.*, (2002)[3], McLachlan *et al.*, (2006)[84]). In proteomics, Marelli *et al.*, (2004)[79], Kim *et al.*, (2007)[64], and Chen *et al.*, (2008)[26] have used gaussian mixtures to model protein expression levels. In this work, we adopt a similar line of thought to that of the 'two-groups model' adopted by Efron (2008)[39], for the simultaneous testing of all hypotheses concerning differential expression; $\{H_j\}_{j=1,\ldots,m}$, without the need for strong Bayesian or frequentist assumptions. In particular, we extend the empirical Bayes framework of Efron (2002, 2008), where Gaussian mixture models were used as approximations to $f(z)$ and $f_0(z)$ in the context of microarray gene expression analyses.

3.2.4.1 The Two-Groups Model

The two-groups model is a simple Bayesian construction that facilitates empirical Bayes analyses. It has been widely adopted in the Bayesian microarray literature, as in Lee *et al.*, (2000)[65], Newton *et al.*, (2001)[90], and Efron (2008)[39]. Simply put, the two-groups model assumes that the m proteins are each either null or non-null with prior probability p_0 or $p_1 = 1 - p_0$, and with the corresponding z-values having density either $f_0(z)$ or $f_1(z)$. Then the distribution of all z_j's, $f(z)$, can be given as

$$f(z) = p_0\, f_0(z) + p_1 f_1(z) \tag{3.10}$$

3.2.4.2 Local False Discovery Rate

The idea of using the mixture distribution in (3.10) is also closely related to FDR. In particular to the local false discovery rate of Efron *et al.*, (2001)[41], and Efron and Tibshirani (2002)[40]. By definition, the local false discovery rate, *locfdr*, is the

posterior probability of population $f_0(z)$, given the mixture model in (3.9), and is given by

$$\begin{aligned} locfdr_j(z^*) &= Pr(\text{protein } j \text{ is null} \mid z_j = z^*) \\ &= p_0 f_0(z^*)/f(z^*) = f_0^+(z^*)/f(z^*) \qquad (3.11) \\ &= f_0^+(z^*)/(f_0^+(z^*) + p_1 f_1(z^*)) \qquad (3.12) \end{aligned}$$

where the sub-density $f_0^+(.)$ corresponds to the distribution of the null proteins.

Why a local false discovery rate ? In proteomics data analyses, the above Bayesian definition of $locfdr$ has several advantages over the frequentist FDR. Firstly, it can be implemented at the test statistic value level, when a p-value computation is either cumbersome or not feasible. Secondly, since it only depends on the marginal distribution of the z values, independence of the z_j's is not required. Assumptions about the distribution of the z values under H_1 are also not required.

In essence, the FDR gives an estimate of the number of false positive hypotheses that a practitioner can expect if the experiment is done an infinite number of times, and as such is a less reliable estimate of the number of false discovery hypotheses in any given experiment. The q-value approach of Storey and Tibshirani (2003)[113] is an improvement in this sense since it assigns to each protein its own measure of significance. However, the q-value is not a true estimate of the probability for an individual protein, say protein A, to be a false positive since it is computed using all the proteins that are more significant than protein A. Clearly a protein whose p-value is near to a chosen cutoff, for example 0.05, does not have the same probability to be differentially expressed as a protein whose p-value is close to zero. This 'averaging' behavior of the q-value tends to yield inflated probabilities for a protein to be a false positive.

The local false discovery rate on the other hand gives an estimate of the false

discovery rate attached to each protein. The estimated local false discovery rate for a given protein provides a measure of belief in the j^{th} protein's significance that depends only on the value of z_j, and not on its inclusion in a larger set of possible values, $Z \leq z_j$. Therefore the $locfdr$ is much preferable in situations where the primary interest is in identifying proteins that show some evidence of differential expression for further biological study. We refer to Aubert *et al.*, (2004)[8] for a more substantial discussion on the need for information at the individual observation level for a given observation to be considered a false positive.

3.2.5 Proposed Two-Groups Models

Many approaches have been proposed for *locfdr* estimation by fitting the two-groups model. These include fully parametric, nonparametric, Bayesian and empirical Bayes, and semi-parametric approaches. With any of these approaches, fitting the model requires knowledge of p_0, $f_0(z)$, and of either $f_1(z)$ or $f(z)$.

The marginal distribution of all z_j's, $f(z)$, is estimated using the data for all the proteins in the experiment. The sub-density $f_0(z)$ is typically estimated using only the central part of the distribution of z_j's in the neighborhood of the zero point. The rationale being that this central part consists mainly of null proteins. In microarray gene expression analyses; Allison *et al.*, (2002)[3] estimates $f(z)$ by fitting a mixture of beta distributions, when the two-group model is specified using p-values. Efron (2002)[37] estimates $f(z)$ by maximum likelihood estimates of high-order polynomials and natural spline basis with 7 degrees of freedom.

In Allison *et al.*, (2002)[3], the distribution of the null genes, $f_0(z)$, is simply the beta(1,1) component of the mixture of beta distributions, since under the null hypothesis for a well defined test statistic; the p-values follow a uniform distribution on [0,1], which is equivalent to a beta(1,1). Efron (2002)[37] estimates $f_0(z)$ using an empirical null distribution using both central matching and maximum likelihood

estimates, where the estimation is done using a fixed-sized window around the peak corresponding to the 0 point of the empirical distribution of the z_j's. Pan *et al.*, (2003)[94] used a normal mixture model with the number of mixture components estimated by a likelihood ratio test based procedure for the estimation of both $f(z)$ and $f_0(z)$.

We propose to investigate the performance of several distributional choices for both $f_0(z)$ and $f(z)$, including normal, skew-normal (sN) and skew-t (sT), and finite mixtures of them. Finite mixtures of distributions have found wide recognition in modeling heterogeneous data and as approximations to complicated probability densities, presenting multimodality, skewness and heavy tails. Comprehensive surveys of the application of mixture models are available in Böhning (2000) [21], McLachlan and Peel (2000) [86], and from a Bayesian perspective, in Frühwirth-Schnatter (2006) [47]. Furthermore, we propose to investigate the utility of the Generalized Hyperbolic (GH) distribution, as a means of dealing with the excess kurtosis that is sometimes observed in the central peak region of the distribution of z, while at the same time effectively capturing heavier tailed behavior.

The above choices are motivated by our experience in our previous research, where we used mixtures of nonparametric (kernel) densities as an approximation to the distribution of $f(z)$, in combination with mixtures of normals as an approximation to $f_0(z)$. However, the local false discovery strategy using these model setups did not produce robustly reproducible results, mainly due to issues of skewness, multi-modality, outliers, and sparsity of data points beyond the central region of the distribution of the test statistic z and over fitting. We say an estimated distribution is over fitted if it assigns high likelihood to the data, but does not generalize to new samples similarly drawn. However, over fit is not synonymous with inconsistency. The use of simple models (for instance models involving the normal distribution) is convenient and computationally cheaper. But this simplicity comes at the expense of relatively

low likelihood (unless the true distribution is of the assumed form), i.e., under fitting complex data distributions. On the other hand, even when the need for accurate nonparametric fits (consistency) is recognized, the use of such techniques as kernel density estimates leads to over fitting (though consistent), through the use of a distribution that is significantly more nuanced than is required to accurately represent the data. In this sense, kernel estimators provide little understanding of the data or practical usefulness since all of the data is retained rather than summarized, leading to non generalizable results.

Our approach is to seek a middle ground between these two extremes by employing sufficient insight and computational resources to consider a rich variety of distributions and find a distributional setting which best explains the data, without sacrificing generalizability. We achieve this goal by considering generalized forms of normal, Student's t, and hyperbolic distributions, and where necessary finite mixtures of them. These distributions allow the fitting of skewed and kurtotic distributions while providing a heavier or lighter tailed fit as compared to the normal.
We consider three pairs of distributional choices,

$$\begin{aligned}
\text{Nmix-Tmix} &\equiv f_0(z) \sim \text{Mixture of } N_{DT}(\mu, \sigma^2); f(z) \sim \text{Mixture of } t(\mu, \sigma^2) \\
\text{sN-sTmix} &\equiv f_0(z) \sim sN_{DT}(\mu, \sigma^2, \lambda); f(z) \sim \text{Mixture of } sT(\mu, \sigma^2, \lambda, \tau) \\
\text{sNmix-GH} &\equiv f_0(z) \sim \text{Mixture of } sN_{DT}(\mu, \sigma^2, \lambda); f(z) \sim GH,
\end{aligned}$$

where the subscript 'DT' refers to *Doubly Truncated*, indicating the fact that the support for the mixture components considered for $f_0(z)$ is truncated to the left and right over a pre-specified interval.

The motivation for considering a mixture model for $f_0(z)$ is that it allows us to accommodate multi-modality and asymmetry of the distribution around the zero point that results from the fact that experimental sources of error include not only those

that are symmetric and continuously distributed (e.g., pipette and other sampling handling perturbances) but also errors that are asymmetric and/or discontinuously distributed (e.g., peptides matched to the wrong protein, co-eluting peptides that are suppressed). The use of a mixture model to fit $f(z)$ allow the left tail (corresponding to under expressed proteins) and the right tail (corresponding to over expressed proteins) to be modeled separately. This helps in capturing atypical observations that usually would have been considered outliers and removed from further analyses. As noted previously, atypical observations that are detected far away from the central peak region is a common phenomenon in proteomics data.

Although these three model setups are considered separately, in practice any of the above choices for $f_0(z)$ can be used in conjunction with any of the choices for $f(z)$. Typically the only restriction is the availability of a sufficiently large number of data points to reliably estimate all model parameters. We also note that, to our knowledge, this work is the first that attempts to model proteomics data within the framework of skew-Normal, Skew-t, or GH distributions.

3.2.6 Fitting a Two-Groups Model

Fitting any two-groups model involves, at a minimum, the following four steps:

1. Selecting a subset of the data as belonging to the null distribution of the test statistic

2. Estimating the proportion of null cases, $\hat{p}_0$

3. Fitting f_0 and f distributions to the data, and

4. Evaluating the goodness of fit of each fitted distribution

The above four steps yield $(\hat{p}_0, \widehat{f_0(z)}, \widehat{f(z)})$, which are the required components needed to calculate the local false discovery rate of each protein using (3.11). The

analytical methods that we employ in steps 1 through 4 are described in detail in the following sections.

3.2.6.1 Identifying the Null Region

As mentioned earlier, a majority of proteins will not be differentially expressed in any given SILAC experiment. These null proteins have an expected log-2 relative expression ratio of zero and are tightly packed around the center of the distribution of z. A simplistic method of identifying this central 'null' region is to use a cutoff such as $\pm 1\,\mathrm{SD}$ or $|z| \leq 2$. Turnbull (2007)[118], developed a method of optimally calculating this center region that produces consistent FDR estimates under the *weak zero assumption*. This assumption states that there exists an optimal cutoff t such that $f_1(z) = 0$ for $z \in [-t, t]$. We adopt the same strategy in our work since this approach minimizes the MSE of the estimator $\widehat{f_0^+(z)}$ and is asymptotically unbiased, when $f_0(z)$ is assumed to be $N(\mu_0, \sigma_0^2)$. Under this approach, the optimal cutoff rule, as a function of the total number of proteins, m, is

$$t^* \approx \hat{\mu_0} + b\hat{\sigma_0} \tag{3.13}$$

where $\hat{\mu}_0 = \mathrm{median}(z)$, $\hat{\sigma}_0 = \mathrm{IQR}(z)/1.349$, and $b = \max(1\,,\, 4.3\, m^{-0.112966})$.

In the following sections, let, $\boldsymbol{z_0} = \{z : z \in [-t^*, t^*]\}$, $l_0 = 2b\hat{\sigma_0}$, and $m_0 = \#\,(\text{proteins in } \boldsymbol{z_0})$, denote the assumed null protein set, the length of the null region of z, and the number of assumed null proteins, respectively.

3.2.6.2 Proportion of null proteins

The *strong zero assumption* underlying the two-groups model states that: $f_1(z) = 0\ \forall\ z \in [-t^*, t^*]$ for some fixed t^*. Turnbull (2007)[118] and Efron [38, 39]), estimate p_0 using the following simple construction.

Let, $f_0 \sim N(\mu_0, \sigma_0)$ and $H_0(\mu_0, \sigma_0) = \int_{-t^*}^{t^*} f_0 = \Phi\left(\frac{t^* - \mu_0}{\sigma_0}\right) - \Phi\left(\frac{-t^* - \mu_0}{\sigma_0}\right)$, and define the truncated normal distribution with density proportional to the normal distribution over $[-t^*, t^*]$ and zero else where as,

$$f_0^{t^*}(z) = \frac{f_0(z)}{H_0(\mu_0, \sigma_0)},\ z \in [-t^*, t^*] \tag{3.14}$$

Define,

$$\theta_{t^*} \equiv \int_{-t^*}^{t^*} p_0 f_0(z) = p_0 H_0(\mu_0, \sigma_0). \tag{3.15}$$

Then under the strong zero assumption,

$$m_0 \sim Bin(m, \theta_{t^*})$$

and the elements of $\boldsymbol{z_0}$ are *i.i.d.* with distribution given in (3.14), giving the likelihood

$$\mathcal{L}_{t^*}(m, \boldsymbol{z_0} \mid \mu_0, \sigma_0, \theta_{t^*}) = \binom{m}{m_0} \theta_{t^*}^{m_0} (1 - \theta_{t^*})^{m - m_0} \prod_{z \in \boldsymbol{z_0}} f_0^{t^*}(z). \tag{3.16}$$

Maximizing $\mathcal{L}_{t^*}$ yields, $\hat{\theta}_{t^*} = \frac{m_0}{m}$. Then by definition (3.15),

$$\hat{p}_0 = \frac{\hat{\theta}_{t^*}}{H_0(\hat{\mu}_0, \hat{\sigma}_0)}. \tag{3.17}$$

In this work, we extend the work of Turnbull and Efron by allowing for more robust and flexible distributions other than the truncated normal. I.e., in the construction of (3.14 - 3.17), we consider f_0 to be either a mixture of truncated normals, a truncated skew-Normal or a truncated general logistic distribution. Then $H_0(\Psi_0) = \int_{-t^*}^{t^*} f_0 = F\left(t^*; \Psi_0\right) - F\left(-t^*; \Psi_0\right)$, where Ψ_0 denotes the set of parameters

needed to uniquely define f_0, and $\hat{p}_0 = \dfrac{\hat{\theta}_{t^*}}{H_0(\hat{\Psi}_0)}$.

3.2.6.3 Evaluating the goodness of fit of fitted distributions

We evaluate the goodness-of-fit of the fitted mixture distributions by performing the well known χ^2 goodness-of-fit test that tests the null hypothesis that the relative frequencies of occurrence of observed events follow a specified frequency distribution. The test is applied to binned data with the bin width selected to ensure the expected frequency within each bin is sufficiently large. For k bins, the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.18}$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i. The expected frequency is calculated by

$$E_i = m\left(\tilde{F}(Y_{up}) - \tilde{F}(Y_{lo})\right)$$

where where $\tilde{F}$ is the cumulative distribution function for the mixture distribution being tested, Y_{up} is the upper limit for bin i, Y_{lo} is the lower limit for bin i, and m is the total number of proteins in the sample. The test statistic follows, approximately, a chi-square distribution with $k - v - 1$ degrees of freedom, where v is the number of estimated parameters corresponding to a particular fitted mixture. A non-significant p-value for the test indicates a reasonable fit between the data and the fitted distribution.

Plotting the Empirical Cumulative Distribution Function (ECDF) against the CDF of the hypothesized mixture can also provide a rough indication of the adequacy of model fit.

Where appropriate and practical, we will also make use of the Anderson-Darling test to measure the fit of the model. The A-D test statistic pays more attention to the tails of the distribution (Hurst *et al.*, (1995)[60]), and therefore offers a means of measuring the fit of the model to extreme events in the data. In our work, this assessment is quite informative, especially with respect to the fitting of $f(z)$. The A-D test statistic is given by

$$\text{A-D} = -m - \sum_{i=1}^{m} \frac{(2i-1)}{m} \left[ln\tilde{F}(Z_{(i)}) + ln(1 - \tilde{F}(Z_{(m+1-i)})) \right], \tag{3.19}$$

where the $Z_{(i)}$ are the ordered data. The p-value of the test is derived using a recursive algorithm that was developed by Marsaglia *et al.*, (2004)[80]).

3.2.6.4 Selecting the number of mixture components

There are a multitude of methods that have been suggested for the selection of the number of mixture components, g. Among them, those based on the information criterion: AIC of Akaike (1974)[2], and BIC of Schwarz (1978)[110] are the most widely used. However, McLachlan and Peel (2000)[86], showed that the justification for both AIC and BIC, other than as an informal guide, do not hold in this context, since regularity conditions are not satisfied. A test based on the likelihood ratio test statistic, λ, is also not applicable since regularity conditions do not hold for $-2\,log\,\lambda$ to asymptotically follow a null distribution of chi-squared (McLachlan and Basford, 1988[83]). An alternative approach that is based on re-sampling and provides a formal test of the number of components is that of McLachlan (1987), where parametric bootstrapping of the likelihood ratio test statistic is done for sequentially testing the

hypothesis

$$H_0 : k = k_0 \qquad (3.20)$$
$$H_1 : k = k_0 + 1$$

for $k_0 = 1, 2, \ldots,$ terminating after the bootstrapped P-value for one of these tests exceeds a specified significance level. Briefly, this approach proceeds as follows.

1. A bootstrap sample is generated from the mixture density under H_0, with unknown parameter vector, Ψ_{k_0}. Let $\hat{\Psi}_{k_0}$ be the estimated maximum likelihood under this model.

2. The value of the likelihood ratio test statistic, $-2\,log\,\lambda$, is calculated for the bootstrap sample in step 1, after fitting a mixture model for $k = k_0$ and $k = k_0 + 1$ in turn to it.

3. Steps 1, 2 are then independently repeated B times, yielding B estimates of $-2\,log\,\lambda$, the distribution of which gives an approximation to the true null distribution of $-2\,log\,\lambda$.

4. The achieved significance of testing the hypothesis in (3.20) is then obtained by referencing the $-2\,log\,\lambda$ estimate from the original sample against this null distribution.

In our analyses, we make use of the above procedure in making a decision about the correct number of mixture components for $f(z)$. It has been our experience that the number of mixture components required to adequately model $f_0(z)$ or $f(z)$ is never more than two. Therefore we restrict ourselves to testing between $k = 1$ or $k = 2$. For comparison purposes, we also report the BIC criterion for selecting between $k = 1$ or $k = 2$. For the most part, we expect the results based on these two approaches to be the similar since (McLachlan (1987)[82]) had observed that when

the number of mixture components is relatively small, the results of the BIC criterion in selecting the optimal number of components is in general identical to that of the above re-sampling procedure.

3.2.6.5 EM algorithms for finite mixtures

Let Ψ be the vector of unknown parameters consisting of the mixing proportions and the other unknown component density specific parameters. We use the Expectation-Maximization (EM) algorithm of Dempster *et al.*, (1977)[32] to estimate Ψ for both $f_0(z)$ and $f(z)$. The EM algorithm is applied in the framework where each observed value of z is thought to have come from one of the mixture components, but the indicator variable denoting this component membership is missing. The E- and M-steps are alternated repeatedly until the likelihood changes by an arbitrarily small amount in the case of convergence.

An important consideration in using the EM algorithm is the specification of starting values for the algorithm. The EM algorithm is known to have reliable global convergence in the sense that regardless of initial values, the likelihood function of Ψ is increased after each EM iteration. However, we need to guard against the algorithm converging to 'spikes' in the likelihood function that may be far from the actual global maximum. For Gaussian and skew-t mixture distributions, we initialize the EM algorithm for a given number of mixture components (identified using the procedure described in section (3.2.6.4), with empirical means and standard deviations of clusters of z values that are identified by a normal mixture model based cluster analyses. If the number of clusters is g, the initial estimates of mixing proportions are taken to be $1/g$, i.e., we assume equal mixing. In the case of fitting the doubly truncated normal or the doubly truncated skew normal mixture models, starting values are derived using procedures that will be described in Section (3.3.1) and Section (3.4.1), respectively.

The main difficulty with fitting mixture models is finding the global maximizer of Ψ. For instance, the likelihood function $L(\Psi|\boldsymbol{z})$ might be unbounded or relatively flat in certain situations. It is also well known that EM-type procedures tend to gravitate towards local modes. A convenient way to address these issues is to try several EM iterations under a variety of starting values. The resulting EM estimates with different starting values can then be used to assess the stability of the final parameter estimates. If multiple modes do exist, the global mode can be found by comparing their respective log-likelihood values.

3.2.6.6 Identifiability of Mixture Distributions

Mixture models can present particular difficulties with *identifiability*. Let

$$\mathfrak{F} = \left\{ f(x,\theta) | \theta \in \Omega, x \in \mathbb{R}^d \right\},$$

be the class of distribution functions from which mixtures are to be formed. Then if $\mathfrak{M} = \{M, M'\}$, where $M = \sum_{j=1}^{c} \pi_j f_j$ and $M' = \sum_{j=1}^{c'} \pi_{j'} f_{j'}$, and $f_j(\cdot), f_{j'}(\cdot) \in \mathfrak{F}$, then we say $\mathfrak{M}$ is *identifiable* if $M \equiv M' \Leftrightarrow c = c'$, and we can order the summations such that $\pi_j = \pi_{j'}$, $f_j = f_{j'}$, for $j = 1, \ldots, c = c'$.

Everitt and Hand (1981)[45], Titterington, Smith and Makov (1985)[116] discuss a number of sufficient conditions for identifiability with regards to univariate Gaussian, Exponential and Poisson mixtures. In particular, they note that Gaussian mixtures are *identifiable* provided the θ_js are all different. However, to our knowledge, the issue of *identifiability* has not been addressed rigorously for mixtures of more complicated distributions such as the skew normal, skew t, and truncated versions thereof. Therefore, we conduct our analyses under the assumption that for mixtures of no more than two components, *identifiability* criteria are satisfied, as long as the θ_j's are well separated from each other.

3.2.6.7 Estimating the local false discovery rate

Once $f_0(z)$ and $f(z)$ have been estimated from the data, and their goodness of fit verified, then together with the estimate of p_0, the local false discovery rate corresponding to each protein j can be calculated using (3.11) as

$$\begin{aligned} locfdr_j(z^*) &= Pr(\text{protein } j \text{ is null} \mid z_j = z^*) \\ &= \hat{f_0^+}(z^*)/\hat{f}(z^*) \end{aligned} \tag{3.21}$$

Choosing a cutoff value. Typically the appropriate cutoff value for the local false discovery rate is set *a priori*, based on the experimenter's expert knowledge. Standard cutoff values are 0.2, 0.1 or 0.05. The cutoff can also be chosen based on the objective evaluation of a cost function. The estimated cutoff would be the value at which the cost function optimally balances the cost of false positive protein validations and the cost of not discovering a differentially expressed protein.

Alternatively, the cutoff can be chosen based on mathematical considerations. For example, the cutoff can be set to the value that corresponds to the maximum second derivative of the monotone curve of the ordered $locfdr$ values. The maximum second derivative corresponds to the point on the curve at which the instantaneous change in the rate of change of the local false discovery rate is highest.

In the following sections, we describe the details of fitting each of the pairs of distributional choices, discussed in Section 3.2.5.

3.3 Nmix - Tmix Model

In the Nmix-Tmix two-groups model setup, we estimate the distribution of the z values associated with the null proteins using a finite mixture of univariate normal densities, and the full distribution of all z values using a finite mixture of univariate

Student's t densities. The primary advantage of using the t distribution to model $f(z)$ is that it provides a longer and heavier tailed alternative to that provided by a normal distribution. Another potential advantage of using the t-distribution instead of the normal to form the mixture $f(z)$ is that due to its greater flexibility, one may require fewer terms (mixing densities) to achieve a given accuracy of approximation to the true density. In essence, this particular combination of mixture models for $f_0(z)$ and $f(z)$ provide a way to flexibly model the central region of the distribution of z values, while providing more stable estimates of tail probabilities.

3.3.1 Estimating $f_0(z)$ and $f(z)$

The empirical distribution of the true null proteins, $f_0(z)$, can be estimated by fitting a mixture of normal densities to the null protein set, $\boldsymbol{z_0}$, each with it's own mean and variance. Assuming that we use a maximum of 2 components, we can represent $f_0(z)$ as

$$f_0(\boldsymbol{z}; \boldsymbol{\Theta_0}) \doteq \sum_{k=1}^{2} \pi_{0k} \phi^*(\boldsymbol{z_0}; \mu_{0k}, \sigma_{0k}^2) \tag{3.22}$$

where $\phi^*(.)$ denotes the doubly truncated normal distribution with support $[-t^*, t^*]$, $\boldsymbol{\Theta_0} = (\pi_{01}, \mu_{01}, \sigma_{01}^2, \mu_{02}, \sigma_{02}^2)$ denotes the unknown parameters corresponding to the two mixture components, and $\sum_{k=1}^{2} \pi_{0k} = 1$. Following the work of Shah *et al.*, (1966)[111], we can obtain the method of moments estimates of a doubly truncated normal distribution as follows.

Let $\mathfrak{Z} = \dfrac{z - \mu}{\sigma}$ be the standardized form of random variable Z with support $-\infty \leq -t < t \leq +\infty$. The value of $t \in \mathbb{R}$ is given by (3.22), using the fact that $f_0(z)$ is now $\phi(\mathfrak{z})$. If $m_r(-t, t) = \mathbb{E}\left(\mathfrak{Z}^r | -t < \mathfrak{Z} \leq t\right)$ denotes the r^{th} moment of the doubly

truncated $\mathfrak{Z}$, then,

$$m_{2k}(-t,t) = (2k-1)!! \left(1 - \sum_{i=1}^{k} \frac{1}{(2i-1)!!} \frac{[\mathfrak{z}^{2i-1}\phi(\mathfrak{z})]_{-t}^{t}}{[\Phi(\mathfrak{z})]_{-t}^{t}}\right), \quad for \;\; k = 1, 2, \ldots, \tag{3.23}$$

$$m_{2k+1}(-t,t) = -\sum_{i=0}^{k} \frac{(2k)!!}{(2i)!!} \frac{[\mathfrak{z}^{2i}\phi(\mathfrak{z})]_{-t}^{t}}{[\Phi(\mathfrak{z})]_{-t}^{t}}, \quad for \;\; k = 0, 1, \ldots, \tag{3.24}$$

where $n!!$ denotes the double factorial defined as

$$n!! = \begin{cases} 1, & \text{if } n = -1, 0, 1; \\ n \times (n-2)!!, & \text{if } n \geq 2. \end{cases}$$

Now, assuming a total of g mixture components, we can represent $f(z)$ as a mixture of standard t-distributions as

$$f(\boldsymbol{z}; \boldsymbol{\Theta}) \doteq \sum_{k=1}^{g} \pi_k \, t(\boldsymbol{z}; \mu_k, \sigma_k^2, \nu_k) \tag{3.25}$$

where $\boldsymbol{\Theta} = (\theta_{\mathbf{1}}, \ldots, \theta_{\mathbf{g}})$ with $\theta_{\mathbf{k}} = (\pi_k, \mu_k, \sigma_k^2, \nu_k)$, denoting the unknown parameters corresponding to mixture component k, with $\sum_{k=1}^{g} \pi_k = 1$.

We fit the mixture model in (3.22) using the EM algorithm by adopting the methodological developments of McLachlan and Jones (1988)[85] for fitting mixture models to truncated data. Starting values for the EM algorithm are derived using (3.23) and (3.24), separately for the two doubly truncated distributions supported over $[-t, \text{median}(\mathfrak{z})]$ and $[\text{median}(\mathfrak{z}), t]$.

Mixture model (3.25) is fitted using a variant of the EM algorithm due to Hoogerheide *et al.*, (2007)[57]; and Hoogerheide and van Dijk (2008)[58].

3.4 sN - sTmix Model

The distribution of z values can sometimes exhibit skewness. As mentioned in Section 3.2.5, this asymmetry is mostly due to errors that are asymmetric and/or discontinuously distributed. For example, Ramos-Fernández *et al.*, (2007)[98] found that the distribution of ^{18}O labeled log-2 ratios was not correctly centered on zero but was significantly biased toward the non-labeled sample, displaying a significant asymmetry comprising of an extended right tail. The authors found that this asymmetry was due to the presence of a small number of peptides having a low but significant proportion of non-labeled species. These types of distortions in the symmetry of the distribution is often not effectively captured by mixture models that use symmetric components, unless allowance is made for the asymmetry introduced by the skewness. In the sN-sTmix model setup, we extend the Nmix-Tmix model by allowing the component normal and t distributions to have an additional shape parameter that determines skewness. This provides a more flexible approach to the fitting of asymmetric subclasses that exhibit varying degrees of skewness. For the most part, the additional flexibility provided by the skew-Normal distribution allows us to fit $f_0(z)$ accurately with only a single component model. Similarly, the additional flexibility afforded by the skew-t distribution usually results in fewer mixture components needed in the fitting of $f(z)$.

3.4.1 The Skew-Normal (sN) Distribution

The sN distribution, developed by Azzalini [9, 10] is a class of density functions dependent on an additional shape parameter, and includes the normal density as a special case. With the additional skewness parameter, $\lambda \in \mathbb{R}$, a skew normally

distributed random variable Z has the density

$$sN(z \mid \mu, \sigma^2, \lambda) = \frac{2}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right)\Phi\left(\lambda\frac{z-\mu}{\sigma}\right), \tag{3.26}$$

where $\phi(.)$ and $\Phi()$ denote the standard normal density function and cumulative distribution function, respectively. Note that if $\lambda = 0$, the density of Z reduces to the $N(\mu, \sigma^2)$ density.

For a sample of size n, Arnold *et al.*, (1993)[6], derived the following method of moments estimators for the parameters (μ, σ^2, λ) of the skew normal distribution:

$$\tilde{\mu} = m_1 - a_1\left(\frac{m_3}{b_1}\right)^{\frac{1}{3}}, \tag{3.27}$$

$$\tilde{\sigma^2} = m_2 + a_1^2\left(\frac{m_3}{b_1}\right)^{\frac{2}{3}}, \tag{3.28}$$

$$\tilde{\delta}(\lambda) = \left[a_1^2 + m_2\left(\frac{b_1}{m_3}\right)^{\frac{2}{3}}\right]^{-\frac{1}{2}}, \tag{3.29}$$

where $a_1 = \sqrt{2/\pi}$, $b_1 = (4/\pi - 1)a_1$, $m_1 = n^{-1}\sum_{i=1}^{n} Z_i$, $m_2 = (n-1)^{-1}\sum_{i=1}^{n}(Z_i - m_1)^2$, $m_3 = (n-1)^{-1}\sum_{i=1}^{n}(Z_i - m_1)^3$, and $\delta(\lambda) = \lambda/\sqrt{(1+\lambda^2)}$.

3.4.2 The Doubly Truncated Skew-Normal (DTsN) Distribution

Let, $f_{\mu,\sigma,\lambda}(z)$ and $F_{\mu,\sigma,\lambda}(z)$ represent the *p.d.f.* and *c.d.f.* of a random variable Z, following (3.26). Then,

$$f_{\mu,\sigma,\lambda}\left(z \mid -t^* < Z \leq t^*\right) = \begin{cases} \frac{1}{\left[F_{\mu,\sigma,\lambda}(z)\right]_{-t^*}^{t^*}} f_{\mu,\sigma,\lambda}(z), & \text{if } -t^* < z \leq t^*; \\ 0, & \text{elsewhere,} \end{cases} \tag{3.30}$$

gives the *p.d.f.* of a doubly truncated skew-Normal (DTsN) distribution, truncated

at $-\infty \leq -t^* < t^* \leq +\infty$. The moments of the DTsN distribution can be derived following the works of Martinez *et al.*, (2008)[81], Genton *et al.*, (2001)[48], and Flecher *et al.*, (2009)[46], as follows. Let $Z \sim DTsN(\mu, \sigma, \lambda)$. Then we have,

$$\mathbb{E}\left[Z^m | -t^* < Z \leq t^*\right] = \sum_{r=0}^{m} \binom{r}{m} \mu^{m-r} \sigma^r s_{\lambda,r}(u, v), \tag{3.31}$$

where $u = (-t^* - \mu)/\sigma$, $v = (t^* - \mu)/\sigma$ and $s_{\lambda,r}(u, v) = \mathbb{E}\left[X^r | u < X \leq v\right]$ is the r^{th} moment of a random variable X, distributed as $DTsN(0, 1, \lambda)$, and $s_{\lambda,0}(u, v) = 1$. From Flecher *et al.*, (2009)[46], we get

$$s_{\lambda,2p}(u, v) = (2p-1)!! + \sum_{k=1}^{p} \frac{(2p-1)!!}{(2k-1)!!} r_{\lambda,2k}(u, v), \quad with \;\; p = 1, 2, \ldots, \tag{3.32}$$

$$s_{\lambda,2p+1}(u, v) = \sum_{k=0}^{p} \frac{(2p)!!}{(2k)!!} r_{\lambda,2k+1}(u, v), \quad with \;\; p = 0, 1, \ldots, \tag{3.33}$$

where with $\lambda_* = (1 + \lambda^2)^{1/2}$ and $r = 1, 2, \ldots$; $r_{\lambda,r}(u, v)$ is given by

$$r_{\lambda,r}(u, v) = -\frac{\left[x^{r-1} f_{0,1,\lambda}(x)\right]_u^v}{\left[F_{0,1,\lambda}(x)\right]_u^v} + \frac{2}{\sqrt{2\pi}} \frac{\lambda}{\lambda_*^r} \frac{\left[\Phi(\lambda_* x)\right]_u^v}{\left[F_{0,1,\lambda}(x)\right]_u^v} m_{r-1}(\lambda_* u, \lambda_* v) \tag{3.34}$$

3.4.3 The Skew-t (sT) Distribution

The sT distribution, introduced by Azzalini and Capitaino (2003)[11], simultaneously allows for both heavier tails and skewness. This allows for more robust fitting in the presence of outlying observations, while accounting for asymmetry in each component distribution. A random variable T is said to follow the sT distribution $sT(\mu, \sigma^2, \lambda, \nu)$, with skewness parameter $\lambda \in \mathbb{R}$ and degrees of freedom $\nu \in (0, \infty)$, if it has the following representation

$$T = \mu + \sigma \frac{Z}{\sqrt{\tau}}, \quad Z \sim sN(\mu, \sigma^2, \lambda), \quad \tau \sim \Gamma(\nu/2, \nu/2), \quad Z \perp \tau, \tag{3.35}$$

However, it is computationally more convenient to use the following representation of the sT model, due to Azzalini (1986)[10] and Henze (1986)[56]. Adopting the stochastically equivalent representation of $Z \sim sN(\mu, \sigma^2, \lambda)$ given by $Z = \delta_\lambda |U_1| + \sqrt{1-\delta_\lambda^2}\, U_2$, where $\delta_\lambda = \lambda/\sqrt{1+\lambda^2}$, and U_1 and U_2 are independent $N(0,1)$ random variables, we get the following hierarchical representation of $sT(\mu, \sigma^2, \lambda, \nu)$ given by

$$\begin{aligned} T|\gamma, \tau &\sim N\left(\mu + \delta_\lambda \gamma, \frac{1-\delta_\lambda^2}{\tau}\sigma^2\right), \\ \gamma|\tau &\sim TN\left(0, \frac{\sigma^2}{\tau}; (0, \infty)\right), \\ \tau &\sim \Gamma(\nu/2, \nu/2), \end{aligned} \tag{3.36}$$

where, $TN\left(\mu, \sigma^2; (a, b)\right)$ represents the the truncated normal distribution with $N(\mu, \sigma^2)$ lying within the truncated interval (a, b).

3.4.4 Estimating $f_0(z)$ and $f(z)$

Adopting the representation in (3.30) for $f_0(z)$, the method of moments estimates of (μ, σ, λ) can be obtained by equating the sample moments to the first three moments of the DTsN distribution and solving (3.31) using the L-BFGS-B algorithm (Byrd *et al.*, (1995))[23], a non-linear optimization procedure with bound constraints on the variables. A key feature of this nonlinear solver is that it does not require second derivatives or knowledge of the structure of the objective function. I.e., knowledge of the Hessian matrix is not required. The solver computes search directions by keeping track of a quadratic model of the objective function with a limited-memory BFGS (Broyden-Fletcher-Goldfarb-Shanno) approximation to the Hessian. The algorithm allows lower and upper bounds to be set for each variable. In our work, we make use of the implementation of the L-BFGS-B algorithm provided in the optim function in R. Initial values for the parameters to be optimized over are obtained using (3.27) - (3.29), ensuring that they satisfy the boundary constraints.

Given the representation in (3.35), and a total of g mixture components, we can represent $f(z)$ as

$$f(\boldsymbol{z};\boldsymbol{\Theta}) \doteq \sum_{k=1}^{g} \pi_k \, sT(\boldsymbol{z};\mu_k,\sigma_k^2,\lambda_k,\nu_k), \tag{3.37}$$

where $\boldsymbol{\Theta} = (\theta_{\mathbf{1}}, \ldots, \theta_{\mathbf{g}})$; $\theta_{\mathbf{k}} = (\pi_k, \mu_k, \sigma_k^2, \lambda_k, \nu_k)$, denotes the unknown parameters of mixture component k; and $\sum_{k=1}^{g} \pi_k = 1$.

Let, $\mathbf{I}_j = (I_{1j}, \ldots, I_{gj})$ be a multinomial random vector with 1 trial and cell probabilities $\pi_1, \ldots, \pi_g$, corresponding to the j^{th} observation from a sample of size m. Then using the hierarchical representation given in (**??**), we can re-express (3.37) as

$$\begin{aligned} T_j|\gamma_j,\tau_j,I_{kj}=1 &\sim N\left(\mu_k+\delta_{\lambda_k}\gamma_j, \frac{1-\delta_{\lambda_k}^2}{\tau_j}\sigma_k^2\right), \\ \gamma_j|\tau_j,I_{kj}=1 &\sim TN\left(0,\frac{\sigma_k^2}{\tau_j};(0,\infty)\right), \\ \tau_j|I_{kj}=1 &\sim \Gamma(\nu_k/2,\nu_k/2), \end{aligned} \tag{3.38}$$

where, $\mathbf{I}_j \sim Multinomial\,(1;\pi_1,\pi_2,\ldots,\pi_g)$, and $j = 1, \ldots, m$.

It follows from (3.38) that the complete data log-likelihood of $\boldsymbol{\Theta}$, ignoring constants, is given by

$$\begin{aligned} \ell(\boldsymbol{\Theta}) = \sum_{j=1}^{n}\sum_{k=1}^{g} I_{kj}\Bigg\{ &\log\pi_k - \frac{\nu_k\tau_j}{2} - \frac{\tau_j\eta_{kj}^2}{2(1-\delta_{\lambda_k}^2)} + \frac{\delta_{\lambda_k}\eta_{kj}\gamma_j\tau_j}{(1-\delta_{\lambda_k}^2)\sigma_k} - \frac{\gamma_j^2\tau_j}{2(1-\delta_{\lambda_k}^2)\sigma_k^2} \\ &- \frac{1}{2}\log(1-\delta_{\lambda_k}^2) - \log\sigma_k^2 + \frac{\nu_k}{2}\log\frac{\nu_k}{2} - \log\Gamma\left(\frac{\nu_k}{2}\right) + \frac{\nu_k}{2}\log\tau_j\Bigg\}, \end{aligned} \tag{3.39}$$

where, $\eta_{kj} = (z_j - \mu_k)/\sigma_k$ and $\delta_{\lambda_k} = \lambda_k/\sqrt{1+\lambda_k^2}$.

3.5 sNmix-GH Model

In the sNmix-GH two-groups model, we estimate $f_0(z)$ using a mixture of doubly truncated skew normal distributions and $f(z)$ using a generalized hyperbolic distributions. The GH distribution provides a flexible framework for modeling data that exhibit skew and/or leptokurtic properties in the neighborhood of the central peak of the z distribution, and tail behavior that decays slower than the normal. Leptokurtic distributions have higher peaks around the mean compared to normal distributions, which leads to thick tails on both sides. These peaks result from the data being highly concentrated around the mean, due to lower variations within observations. This behavior is not unexpected, since under the 'strong zero assumption' underlying the two groups model, we expect roughly over 90% of the data to to be tightly packed around the central region of the support of the z distribution. The main motivation behind using a mixture of skew normals for modeling $f_0(z)$ is to demonstrate its ability to capture leptokurtic behavior more accurately compared to a single component skew-Normal or a mixture of Gaussians.

The GH distribution can be thought of as a normal mean-variance mixture distribution with the Generalized Inverse Gaussian (GIG) as the mixing density. The various distributions that can be derived from the GH, differ in the behavior of the central peak of the density and in the type of decay at the tail ends. This flexibility also means that a single component GH distribution usually suffices for fitting both $f_0(z)$ and $f(z)$, as opposed to mixtures of standard statistical distributions such as the normal or Student's t.

3.5.1 The Generalized Hyperbolic (GH) Distribution

The Generalized Hyperbolic (GH) distribution, first introduced by Barndorff-Nielsen (1977)[14], is a class of distributions that have found a wide audience among econometrists

in the areas of modeling the log return distributions of financial assets and the pricing of derivatives. The one-dimensional generalized hyperbolic distribution is defined by the following density

$$\begin{aligned} gh(z;\lambda,\alpha,\beta,\delta,\mu) = & a(\lambda,\alpha,\beta,\delta)\left(\delta^2+(z-\mu)^2\right)^{(\lambda-\frac{1}{2})/2} \\ & \times K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+(z-\mu)^2}\right)\exp\left(\beta(z-\mu)\right), \end{aligned} \tag{3.40}$$

where,

$$a(\lambda,\alpha,\beta,\delta) = \frac{(\alpha^2-\beta^2)^{\lambda/2}}{\sqrt{2\pi}\,\alpha^{\lambda-\frac{1}{2}}\,\delta^{\lambda}K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}\ , \text{ and} \tag{3.41}$$

$K_\lambda(x)$ is a modified Bessel function of the second kind, which gives the solutions to the modified Bessel's equation

$$x^2\,\frac{d^2\lambda}{dx^2} + x\,\frac{d\lambda}{dx} - (x^2+\lambda^2)\lambda = 0, \text{ and} \tag{3.42}$$

There are many possible parameterizations of the GH distribution. Under the parameterization we have adopted for (3.40), the domain of variation of the parameters is $\mu \in \mathbb{R}$ and

$$\begin{aligned} &\delta \geq 0\,,\ |\beta| < \alpha \quad if\ \lambda > 0 \\ &\delta > 0\,,\ |\beta| < \alpha \quad if\ \lambda = 0 \\ &\delta > 0\,,\ |\beta| \leq \alpha \quad if\ \lambda < 0 \end{aligned}$$

Under this parameterization, The parameters μ, δ describe the location and the scale, and β describes the skewness. Decreasing $\bar{\alpha}\,(=\alpha\delta)$ reflects an increase in the kurtosis. Normal distributions are characterized only by the scale and location parameters. The additional parameters of GH distributions allow us to specify in

particular the tail of the distribution more exactly. Another important aspect of GH distributions is that they embrace many special cases. When $\beta = 0$, the distributions in (3.45) become symmetric. The normal distribution is obtained as a limiting case when $\delta \to \infty$ and $\delta/\alpha \to \sigma^2$ (Barndorff-Nielsen, 1977)[14]; and for $\lambda = -\upsilon/2$, $\alpha = \beta = \mu = 0$, $\delta = \sqrt{\upsilon}$, we get the Student's t distribution (Eberlein and Hammerstein, 2003)[36].

All moments of (3.40) exist (Gut, 1995)[52]. In particular, the mean and variance of the GH distribution are given by

$$\mathbb{E}(Z) = \mu + \frac{\beta\delta}{\sqrt{\alpha^2 - \beta^2}} \frac{K_{\lambda+1}(\zeta)}{K_\lambda(\zeta)} \tag{3.43}$$

$$\mathbb{V}(Z) = \delta^2 \left(\frac{K_{\lambda+1}(\zeta)}{\zeta K_\lambda(\zeta)} + \frac{\beta^2}{\alpha^2 - \beta^2} \left[\frac{K_{\lambda+2}(\zeta)}{K_\lambda(\zeta)} - \left(\frac{K_{\lambda+1}(\zeta)}{K_\lambda(\zeta)} \right)^2 \right] \right), \tag{3.44}$$

where, $\zeta = \delta\sqrt{\alpha^2 - \beta^2}$.

3.5.2 Estimating $f_0(z)$ and $f(z)$

A mixture of DTsN's can be fitted using the computational techniques developed by Lin *et al.*, (2007)[67] and McLachlan and Jones (1988)[85]. Initial values for the parameters to be optimized over are obtained using the method of moments estimates (3.27) - (3.29), separately for each identified cluster.

The log-likelihood function for the GH distribution for observations $z_1, \ldots, z_m$ is

$$\begin{aligned} L = &\log a(\lambda, \alpha, \beta, \delta) + \left(\frac{\lambda}{2} - \frac{1}{4} \right) \sum_{j=1}^{m} \log \left(\delta^2 + (z_j - \mu)^2 \right) \\ &+ \sum_{j=1}^{m} \left[\log K_{\lambda - \frac{1}{2}} \left(\alpha \sqrt{\delta^2 + (z_j - \mu)^2} \right) + \beta(z_j - \mu) \right] \end{aligned} \tag{3.45}$$

An important point concerning the speed of the ML estimation of (3.45) is the selection of starting values. Let, $(\lambda_0, \alpha_0, \beta_0, \delta_0, \mu_0)$ be the vector of starting values. We

choose these starting values by re-scaling a symmetric GH distribution (i.e, $\beta = 0$), with a reasonable kurtosis, e.g. the kurtosis of the observed data under symmetry, $\hat{\kappa}_{emp}$, such that the empirical variance $\hat{\sigma}^2_{emp}$ and the variance of the GH distribution given in (3.44) are equal. That is, when $\beta = 0$, we have $\hat{\bar{\alpha}} = \alpha_0\delta_0 = \hat{\kappa}_{emp}$. Then from (3.43) and (3.44) we get,

$$\begin{aligned}
\beta_0 &= 0 \\
\mu_0 &= \hat{\mu}_{emp} \\
\hat{\sigma}^2_{emp} &= \delta_0^2 \left(\frac{K_{\lambda+1}(\hat{\bar{\alpha}})}{\hat{\bar{\alpha}} K_\lambda(\hat{\bar{\alpha}})} \right) \\
\alpha_0 &= \hat{\bar{\alpha}}/\delta_0
\end{aligned}$$

We set $\lambda_0 = 1$. This setting corresponds to a hyperbolic distribution (Eberlein and Keller, 1995)[35], and is a reasonable starting point for estimating the GH distribution (Barndorff-Nielsen and Shephard, 1998[15]).

A second important factor for the speed of the estimation is the number of modified Bessel functions to compute. We estimate the modified Bessel functions using the besselk routine in MATLAB, which implements a numerical approximation method, described in Press, Teukolsky, Vetterling, and Flannery (1992)[97].

3.6 Results

The typical choices for fitting the empirical distribution of f include the normal, a mixture of normals, a kernel density estimate, or a polynomial or natural spline fit. Our choices for f are (a) Student's t mixture, (b) a skew-t mixture, and (c) the generalized hyperbolic distribution.

We begin by applying our methods to the three motivating SILAC data sets. The two control data sets consist of light and heavy isotope labeled yeast samples that

were mixed in a 1:1 ratio. All yeast protein expression ratios derived from these control data sets should be equal to one, with allowance made for random noise contamination. The first control data set consists of 614 identified proteins. The second sample is essentially a technical replicate of the same 1:1 protein mixture and consists of 588 identified proteins. The mammalian cellular proteome - Hela cell line data set consists 1536 proteins. Hereafter, we will refer to these three data sets as *Sample A*, *Sample B*, and *HCL*, respectively. We conduct separate analysis on all three data sets, by first estimating the true relative expression level of each protein using the MINQUE methodology, and then fitting different two-groups models to the empirical distribution of the MINQUE statistic, z.

3.6.1 Fitting the Null Distribution, $f_0(z)$

The standard practice when using the local false discovery approach is to fit a truncated normal distribution to the subset of proteins that are deemed non-differentially expressed. We considered four choices for f_0, namely: (a) a truncated normal distribution, (b) a mixture of truncated normals, (c) a truncated skew-normal distribution, and (d) a mixture of truncated skew normals.

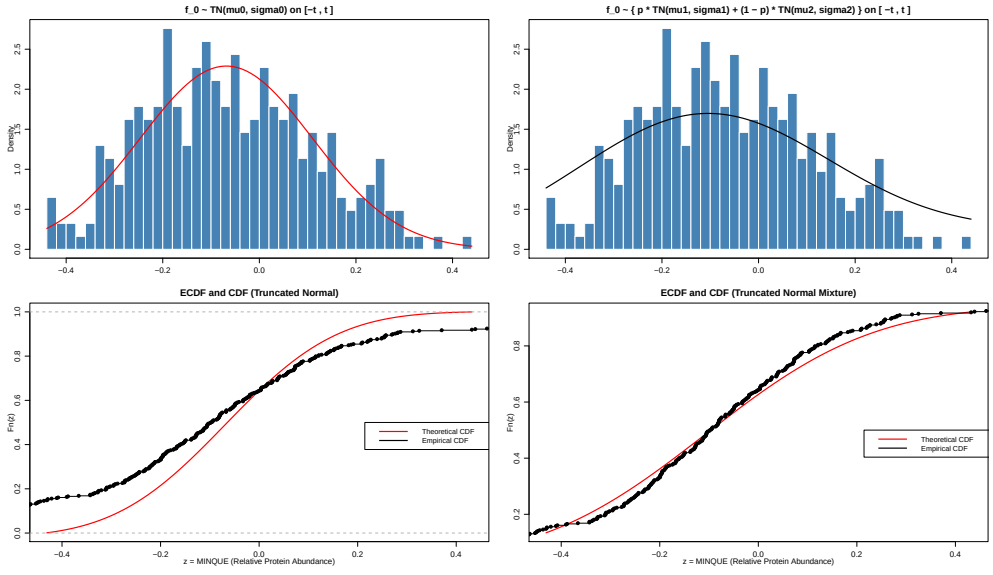
Sample A. The empirical mean, median, and variance of the data are -0.191, -0.172, and 0.738, respectively (as opposed to an expected value of zero for both the mean and the median). The sample estimate of excess skewness compared to a normal distribution is 0.59, indicating that the distribution has a longer right tail. The fit of these four distributional choices (Figure 3.1) to the central region of z, clearly indicate the inadequacy of single component distributions in fully capturing the observed asymmetry in the data.

On the other hand, mixture distributions do a much better job of approximating the empirical distribution compared to their single component alternatives. Of the four fits, the skew normal mixture is the best, followed closely by the mixture of

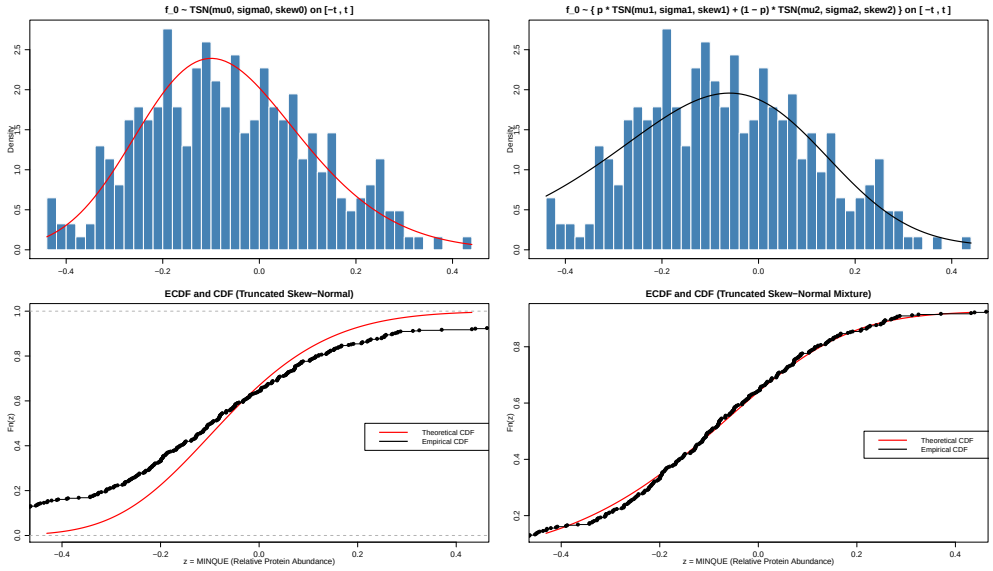
normals. The parameter estimates and the estimated proportion of non-differentially expressed proteins corresponding to the fitted truncated mixtures are given in Table 3.1. Both mixtures consist of components with estimated means on either side of zero. The truncated skew normal mixture is able to capture the longer right tail in the data through its skewness parameter, λ, for the first fitted component. The estimated proportion of non-differentially expressed proteins are 92.1% and 87.1% for the normal and skew normal mixtures, respectively.

Sample B. Data analysis of Sample B revealed similar findings to that of Sample A. The empirical mean, median, and the variance of the data are -0.135, -0.047, and 0.667, respectively. The sample estimate of excess skewness is 1.08, again indicating the longer right tail of the data. The fits of single and two component mixtures and the estimated parameters of the two component mixtures are given in Figure 3.2 and Table 3.2, respectively.

HCL. The empirical mean, median, and variance of the HCL data are -0.288, -0.099, and 1.89, respectively. The higher variability of the HCL data compared to the control data sets is to be expected. The sample estimate of excess skewness is -1.59. This suggests that the data are negatively skewed, i.e., the data exhibit a longer left tail compared to symmetrically distributed data. The fits of single and two component mixtures (Figure 3.3) and the estimated parameters of the two component mixtures (Table 3.3) are again illustrative of the superiority of the two component mixtures in fitting the f_0 distribution and of the superiority of the skew normal mixture in capturing asymmetry.

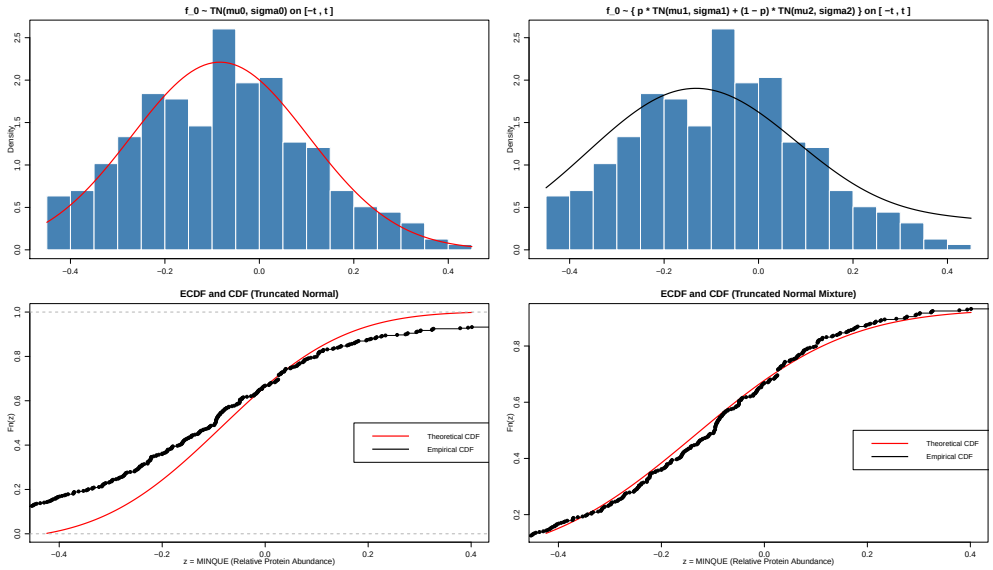


(a) The fit of a truncated normal and a mixture of truncated normals to $f_0(z)$

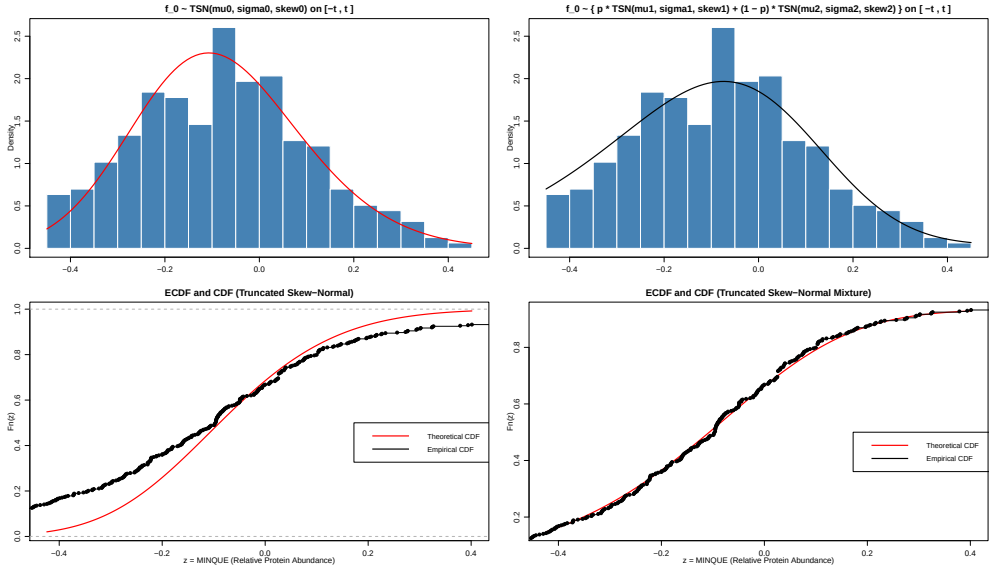


(b) The fit of a truncated skew normal and a mixture of truncated skew normals to $f_0(z)$

Figure 3.1: **Sample A** : The fit of a truncated normals and skew normals as approximations to the distribution of non-differentially expressed proteins, f_0, having support $[-0.579, 0.579]$. A visual comparison of the ECDF to the CDF indicate: the lack-of-fit of single component models and the excellent fit of the skew normal mixture.

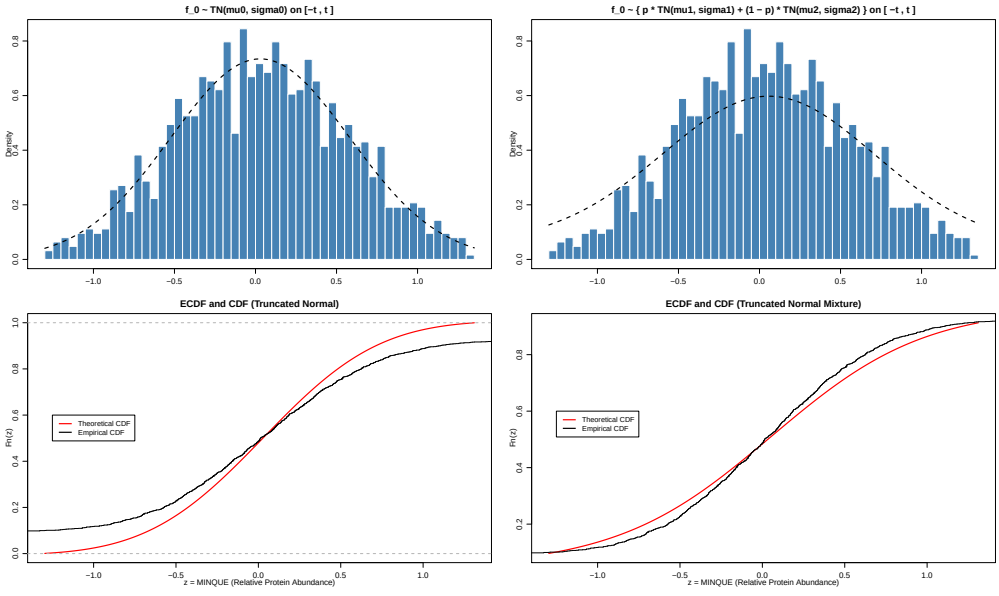


(a) The fit of a truncated normal and a mixture of truncated normals to $f_0(z)$

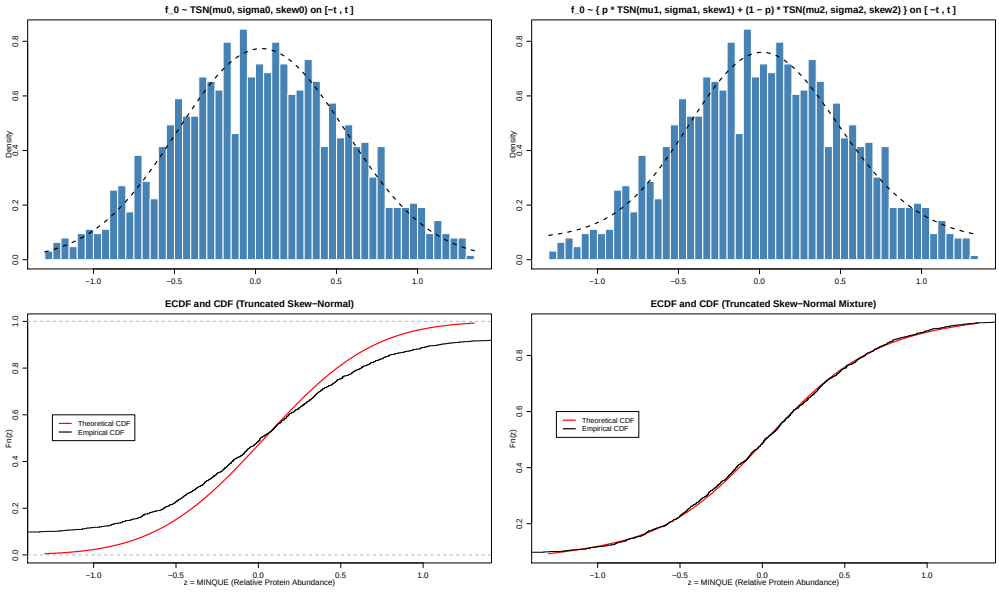


(b) The fit of a truncated skew normal and a mixture of truncated skew normals to $f_0(z)$

Figure 3.2: **Sample B** : The fit of truncated normals and skew normals as approximations to the distribution of non-differentially expressed proteins, f_0, having support $[-0.429, 0.429]$.



(a) The fit of a truncated normal and a mixture of truncated normals to $f_0(z)$



(b) The fit of a truncated skew normal and a mixture of truncated skew normals to $f_0(z)$

Figure 3.3: **HCL** : The fit of truncated normals and skew normals as approximations to the distribution of non-differentially expressed proteins, f_0, having support $[-1.143, 1.143]$. A visual comparison of the ECDF to the CDF again indicate: the lack-of-fit of single component models and the excellent fit of the skew normal mixture.

f_0	$\hat{p}_0$	$\hat{p}$	Parameter Estimates
$p * N_t(\mu_1, \sigma_1) + (1-p) * N_t(\mu_2, \sigma_2)$	0.871	0.139	$\hat{\mu}_1 = 0.498$ $\hat{\mu}_2 = -0.135$ $\hat{\sigma}_1 = 2.021$ $\hat{\sigma}_2 = 0.228$
$p * SN_t(\mu_1, \sigma_1, \lambda_1) + (1-p) * SN_t(\mu_2, \sigma_2, \lambda_2)$	0.921	0.892	$\hat{\mu}_1 = 0.153$ $\hat{\mu}_2 = -0.105$ $\hat{\sigma}_1 = 0.534$ $\hat{\sigma}_2 = 0.613$ $\hat{\lambda}_1 = -1.765$ $\hat{\lambda}_2 = -0.081$

Table 3.1: **Sample A** : Parameter estimates under each of the fitted mixture models for f_0. $\hat{p}_0$ is the estimated proportion of non-differentially expressed proteins under each model.

f_0	$\hat{p}_0$	$\hat{p}$	Parameter Estimates
$p * N_t(\mu_1, \sigma_1) + (1-p) * N_t(\mu_2, \sigma_2)$	0.899	0.117	$\hat{\mu}_1 = 0.311$ $\hat{\mu}_2 = -0.128$ $\hat{\sigma}_1 = 1.821$ $\hat{\sigma}_2 = 0.428$
$p * SN_t(\mu_1, \sigma_1, \lambda_1) + (1-p) * SN_t(\mu_2, \sigma_2, \lambda_2)$	0.951	0.829	$\hat{\mu}_1 = 0.079$ $\hat{\mu}_2 = -0.705$ $\hat{\sigma}_1 = 0.301$ $\hat{\sigma}_2 = 2.858$ $\hat{\lambda}_1 = -1.961$ $\hat{\lambda}_2 = 6.868$

Table 3.2: **Sample B** : Parameter estimates under each of the fitted mixture models for f_0.

f_0	$\hat{p}_0$	$\hat{p}$	Parameter Estimates
$p * N_t(\mu_1, \sigma_1) + (1-p) * N_t(\mu_2, \sigma_2)$	0.921	0.335	$\hat{\mu}_1 = -0.120$ $\hat{\mu}_2 = 0.061$ $\hat{\sigma}_1 = 1.775$ $\hat{\sigma}_2 = 0.639$
$p * SN_t(\mu_1, \sigma_1, \lambda_1) + (1-p) * SN_t(\mu_2, \sigma_2, \lambda_2)$	0.960	0.382	$\hat{\mu}_1 = 0.018$ $\hat{\mu}_2 = -0.256$ $\hat{\sigma}_1 = 1.751$ $\hat{\sigma}_2 = 0.550$ $\hat{\lambda}_1 = -0.064$ $\hat{\lambda}_2 = 0.947$

Table 3.3: **Hela Cell Line** : Parameter estimates under each of the fitted mixture models for f_0.

3.6.2 Fitting the Full Distribution, $f(z)$

Sample A and Sample B. The fit of a mixture of Student's t, a mixture of skew-t and the generalized hyperbolic distribution as approximations to the full empirical distribution of the z values is shown in Figure 3.4. For comparison, the fit of a two component normal is also shown. the estimated parameters for each model setup is given in Table 3.4. The two component Student's t mixture, whilst capturing the extreme values in the data fairly well, fails to approximate the remaining 'normal' data. The two component normal mixture is comparably much better. However, it performs poorly in the vicinity of the central peak of the data. The skew-t mixture and the generalized hyperbolic both fit the full range of the data quite well, as illustrated by the overlapping curves of the ECDF and CDF for these two models.

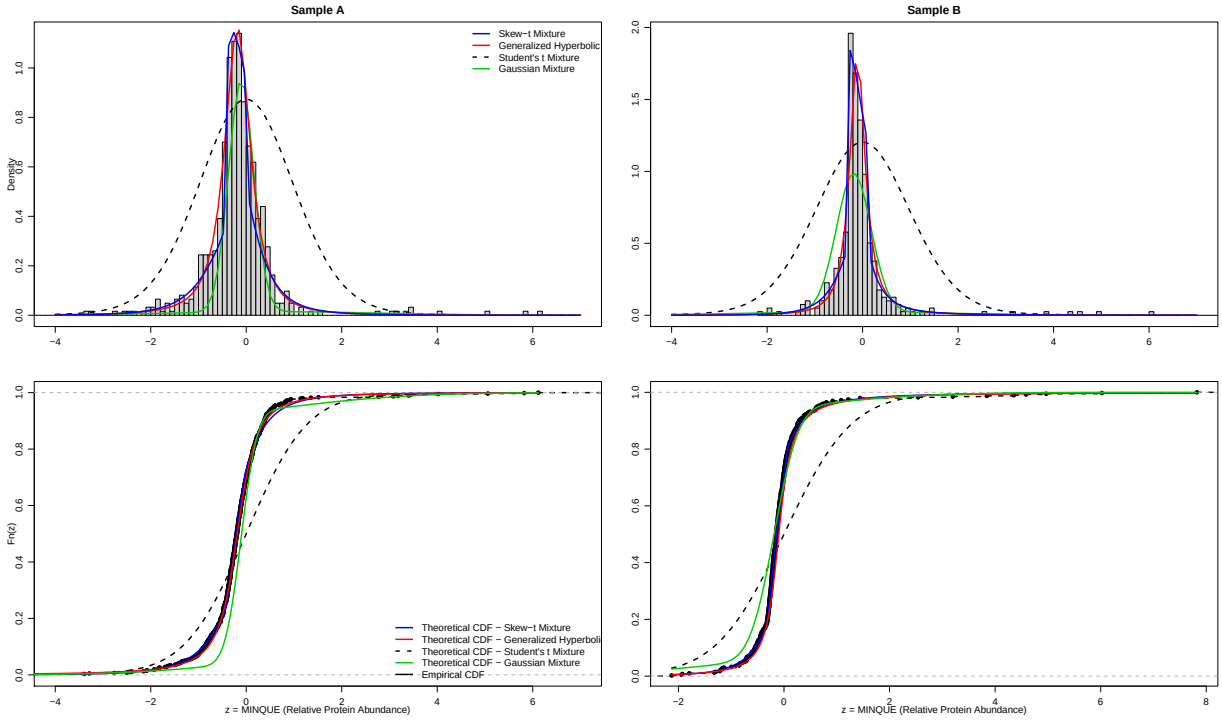


Figure 3.4: Fit of distributions to f, the full distribution of the z values (Sample A and Sample B). The skew-t mixture and the generalized hyperbolic fit the full distribution of z much better than mixtures of Student's t's or Gaussians.

Model	Sample A		Sample B	
t - mixture	$\hat{p} = 0.908$		$\hat{p} = 0.901$	
	$\hat{v}_1 = 29.953$		$\hat{v}_1 = 35.969$	
	$\hat{v}_2 = 1.787$		$\hat{v}_2 = 1.561$	
Skew-t mixture	$\hat{p} = 0.532$		$\hat{p} = 0.413$	
	$\hat{\mu}_1 = 0.171$	$\hat{\mu}_2 = -0.421$	$\hat{\mu}_1 = 0.082$	$\hat{\mu}_2 = -0.322$
	$\hat{\sigma}_1 = 0.635$	$\hat{\sigma}_2 = 0.581$	$\hat{\sigma}_1 = 0.444$	$\hat{\sigma}_2 = 0.299$
	$\hat{\lambda}_1 = -2.298$	$\hat{\lambda}_2 = 8.024$	$\hat{\lambda}_1 = -5.113$	$\hat{\lambda}_2 = 1.312$
	$\hat{\tau}_1 = 3.040$	$\hat{\tau}_2 = 3.207$	$\hat{\tau}_1 = 3.479$	$\hat{\tau}_2 = 1.706$
Gaussian mixture	$\hat{p} = 0.886$		$\hat{p} = 0.877$	
	$\hat{\mu}_1 = -0.107$	$\hat{\mu}_2 = 0.608$	$\hat{\mu}_1 = -0.185$	$\hat{\mu}_2 = -0.289$
	$\hat{\sigma}_1 = 0.264$	$\hat{\sigma}_2 = 2.204$	$\hat{\sigma}_1 = 0.363$	$\hat{\sigma}_2 = 2.220$
Generalized Hyperbolic	$\hat{\alpha}_1 = 0.134$		$\hat{\alpha}_2 = 0.256$	
	$\hat{\beta}_1 = -0.033$		$\hat{\beta}_2 = 0.161$	
	$\hat{\delta}_1 = 0.392$		$\hat{\delta}_2 = 0.213$	
	$\hat{\mu}_1 = -0.181$		$\hat{\mu}_2 = -0.113$	
	$\hat{\lambda}_1 = -0.853$		$\hat{\lambda}_2 = -0.655$	

Table 3.4: Parameter estimates for the fitted models for f - Sample A and Sample B

HCL. The results of fitting the same four distributions to the f distribution of the HCL data is shown in Figure 3.5. The estimated parameters for each model setup is given in Table 3.5. Here again the two component Student's t mixture fails to capture the middle part of the density histogram of the z values. The two component normal mixture estimates the central region and the left tail reasonably well, but the density estimate drops to zero two quickly at the right tail. The skew-t mixture and the generalized hyperbolic both approximate the data quite well. However, compared to the skew-t mixture, the fit of the generalized hyperbolic is smoother. This is because the generalized hyperbolic, since it is a single component model, does not follow the peaks of the histogram too closely.

3.6.3 Number of Mixture Components and Goodness of Fit

We evaluated the goodness-of-fit of the fitted mixture distributions by performing χ^2 goodness-of-fit tests. The appropriate number of mixture components was selected using two methods: the BIC, and by parametric bootstrapping of the likelihood ratio test statistic (λ) for testing between one and two component mixtures. The results

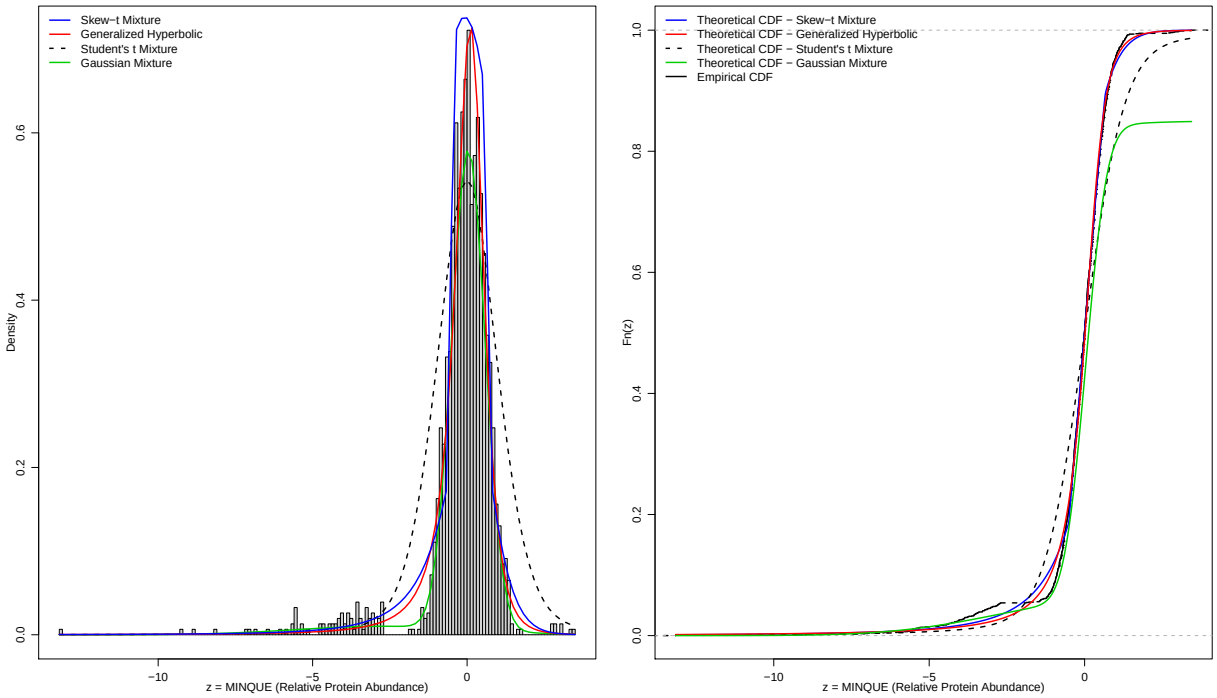


Figure 3.5: Fit of distributions to f (HCL).

Model	**Parameter Estimates**
t - mixture	$\hat{p} = 0.908$ $\hat{v}_1 = 29.953$ $\hat{v}_2 = 1.787$
Skew-t mixture	$\hat{p} = 0.532$ $\hat{\mu}_1 = 0.171$ | $\hat{\mu}_2 = -0.421$ $\hat{\sigma}_1 = 0.635$ | $\hat{\sigma}_2 = 0.581$ $\hat{\lambda}_1 = -2.298$ | $\hat{\lambda}_2 = 8.024$ $\hat{\tau}_1 = 3.040$ | $\hat{\tau}_2 = 3.207$
Gaussian mixture	$\hat{p} = 0.886$ $\hat{\mu}_1 = -0.107$ | $\hat{\mu}_2 = 0.608$ $\hat{\sigma}_1 = 0.264$ | $\hat{\sigma}_2 = 2.204$
Generalized Hyperbolic	$\hat{\alpha}_1 = 0.134$ $\hat{\beta}_1 = -0.033$ $\hat{\delta}_1 = 0.392$ $\hat{\mu}_1 = -0.181$ $\hat{\lambda}_1 = -0.853$

Table 3.5: Parameter estimates for the fitted models for f - HCL.

of these tests for fitting a one/two component normal/skew normal distribution to $f_0(z)$ are shown in Table 3.6, and provide significant evidence to support the decision to fit a two component mixture of skew normals to all three data sets.

	Sample-A $m_0 = 548$, $K = 18$		**Sample-B** $m_0 = 465$, $K = 18$		**HCL** $m_0 = 1367$, $K = 25$	
	Truncated Normal					
Number of Components	1	2	1	2	1	2
χ^2(value , df)	22.714 , 15 ^{b}	16.317 , 12 ^{c}	20.341 , 15 ^{c}	14.303 , 12	28.449 , 22 ^{c}	13.468 , 19
BIC	246.881	242.156	191.962	185.3466	1700.762	1659.888
-2 log(λ), p*	23.644 , 0.03		25.041 , 0.02		62.535 , < 0.01	
	Truncated Skew Normal					
Number of Components	1	2	1	2	1	2
χ^2(value , df)	17.940 , 14	6.730 , 10	17.512 , 14	6.007 , 10	27.915 , 21 ^{c}	11.086 , 17
BIC	246.109	241.024	189.537	184.449	1641.291	1633.689
-2 log(λ), p*	30.310 , 0.01		29.656, 0.04		36.483 , < 0.01	

Table 3.6: Number of mixture components and goodness of fit of fitted models for $f_0(z)$. (b, c indicate significance at the 0.1, 0.2 level, respectively; p^* indicates the attained significance level of the likelihood ratio test based on 250 bootstrap replicates; K = number of bins used in the χ^2 test; and m_0 = number of proteins used to fit $f_0(z)$).

The results of the χ^2, BIC, and likelihood ratio tests for fitting a one/two component normal/sdutent's t/skew-t distribution to $f(z)$ are shown in Table 3.7. In addition to these three criteria, the fit of the generalized hyperbolic was evaluated using the Anderson-Darling (A-D) test. The Student's t distribution does not provide an adequate fit except in the case of fitting a two component mixture to the HCL data. In fact, all three two-component mixtures provide a good fit to the HCL data. Based on the χ^2 criterion alone, a single component skew-t model appears adequate for fitting f. However, both the likelihood ratio test and the BIC clearly favor a two component model. Both the χ^2 and the A-D test results confirm the fit of the generalized hyperbolic distribution to all three data sets. Additionally, the A-D test provides assurance that the generalized hyperbolic distribution adequately fits the extreme values in the data.

	Sample-A m = 614, K = 21		**Sample-B m = 588, K = 21**		**HCL m = 1536, K = 30**	
	Normal					
Number of Components	1	2	1	2	1	2
χ^2(value, df)	35.120, 18 [a]	23.300, 15 [c]	34.707, 18 [a]	21.509, 15	48.430, 27 [a]	28.736, 24
BIC	1148.774	1104.570	1161.070	1142.845	1753.637	1730.844
-2 log(λ), p*	63.464 , < 0.01		37.355 , < 0.01		44.804 , < 0.01	
	Student's t					
Number of Components	1	2	1	2	1	2
χ^2(value, df)	39.114, 19 [a]	30.020, 17 [b]	37.581, 19 [a]	28.661, 17 [b]	41.362, 28 [b]	23.728, 26
BIC	1162.525	1130.974	1202.376	1180.768	1767.284	1733.900
-2 log(λ), p*	44.391 , < 0.01		34.361, 0.03		48.058 , < 0.01	
	Skew t					
Number of Components	1	2	1	2	1	2
χ^2(value, df)	18.943, 16	10.115, 11	19.217, 16	9.849, 11	31.475, 25	16.977, 20
BIC	1088.393	1072.056	1101.442	1096.575	1710.321	1707.761
-2 log(λ), p*	48.437 , < 0.01		36.751, 0.01		39.245 , < 0.01	
	Generalized Hyperbolic					
χ^2(value, df)	12.443, 15		12.559, 15		18.842, 24	
A-D (value, p)	1.301, 0.237		0.821, 0.399		0.699, 0.559	

Table 3.7: Number of mixture components and goodness of fit of fitted models for $f(z)$. (a indicates significance at the 0.05 level; and m = number of proteins used to fit $f(z)$).

3.6.4 Local False Discovery Rate

We calculate a false discovery rate for all the proteins in each of the three data sets, and identify the set of proteins that are deemed significant using two cut-off points. The first cut-off point chooses as significant all proteins with local fdr $\leq$ 0.1. The second cut-off is taken to be the value that corresponds to the maximum second derivative of the smoothed-monotonic local fdr curve. In these calculations we omitted the two groups model, truncated normal mixture - Student's t mixture, since this particular combination of distributional components produced significantly worse results compared to the other combinations. Table 3.8 gives the number and proportion of proteins declared significant under each of the cut-offs.

3.6.5 False Positive and False Negative Rates

The two control samples, Sample A and Sample B, allows us to evaluate the false positive and false negative rates associated with each of the fitted two-groups models.

	$f_0 \sim$ Truncated Normal Mixture, $f \sim$ Normal Mixture	
	Local fdr < 0.1	Local fdr < 0.051
Sample-A	56 (9.12%)	49 (7.98%)
Sample-B	47 (7.99%)	42 (7.14%)
HCL	226 (14.71%)	201 (13.09%)
	$f_0 \sim$ Truncated Skew-Normal Mixture, $f \sim$ Skew-t Mixture	
	Local fdr < 0.1	Local fdr < 0.043
Sample-A	39 (6.35%)	37 (6.03%)
Sample-B	31 (5.27%)	27 (4.59%)
HCL	105 (6.84%)	91 (5.92%)
	$f_0 \sim$ Truncated Skew-Normal Mixture, $f \sim$ GH	
	Local fdr < 0.1	Local fdr < 0.089
Sample-A	21 (3.42%)	20 (3.26%)
Sample-B	16 (2.72%)	16 (2.72%)
HCL	87 (5.66%)	84 (5.47%)

Table 3.8: Number and proportion of significant proteins under each of the fitted two-groups models.

Since we know that all proteins in these two samples were mixed in a 1:1 ratio, all proteins that are declared as significant are by default, false positives. However, not all the identified proteins in these two data sets are yeast proteins. There are 9 identified contaminants in Sample A and 7 identified contaminants in Sample B. These contaminants are highly unlikely to be present in equal amounts in the light and heavy labeled samples. In fact, the log-2 relative expression ratios of these contaminants are all significantly different than zero. In this sense, the presence of these contaminants provides us a way to also calculate a false negative rate for our methods. I.e., we declare as a false negative any of the contaminant proteins that our methods fail to identify as being differentially expressed. However, since the number of false negatives are quite small, the estimates of the false negative rate are likely to be less reliable compared to the false positive rates. The results given in Table 3.9 demonstrates the improved performance of the two two-groups models fitted with asymmetric components: the skew normal - skew-t, and the skew normal - generalized hyperbolic combinations. With our data, the performance of the skew normal - generalized hyperbolic two-groups model, at least with respect to the false positive rate, is clearly superior to the other model setups considered.

	Sample A, m = 614			**Sample B, m = 588**		
	TNM, NM	TSNM, STM	TSNM, GH	TNM, NM	TSNM, STM	TSNM, GH
Proteins with locfdr < 0.1	56	39	21	47	31	16
False Positives	50	31	13	42	26	11
False Negatives	3	1	1	2	2	2
False Positive Rate	8.26 %	5.12 %	2.15 %	7.23 %	4.48 %	1.89 %
False Negative Rate	33.33 %	11.11 %	11.11 %	28.57 %	28.57 %	28.57 %

Table 3.9: False positive and false negative rates for the three two-groups models. TNM-NM = Truncated Normal Mixture - Normal Mixture, TSNM-STM = Truncated Skew Normal Mixture - Skew-t Mixture, TSNM-GH = Truncated Skew Normal Mixture - Generalized Hyperbolic.

3.6.6 Robustness of Results

Since Sample B is a technical replicate of Sample A, we can assess the degree to which results of data analysis on Sample A agrees with those of Sample B. In other words we can assess the degree to which our methods produce reproducible results. Here we use the term *agreement* in the general sense of consistency of the set of proteins declared significant or not significant across the two repeat samples. Since the proteins identified in Sample B are a subset of the proteins in Sample A, we define the percentage of agreement as

$$\text{Agreement (\%)} = \frac{\text{Number significant in both samples + Number not significant in both samples}}{\text{Number of proteins in common}} * 100$$

The statistics shown in Table 3.10 indicate that *agreement* of results is generally good for all three model setups: TNM - NM, STNM - STM, and STNM - GH. This is to be expected since both Sample A and Sample B are in fact control samples. However there is a clear separation of the level of agreement between the model that used symmetric mixture components (TNM - NM) and the two models that used asymmetric mixture components (TSNM-STM, STNM-GH).

		Sample B		
TNM - NM		Significant	Not significant	Agreement (%)
Sample A	Significant	30	19	93.9
	Not significant	17	522	
TSNM - STM		Significant	Not significant	
Sample A	Significant	24	10	97.1
	Not significant	7	547	
TSNM - GH		Significant	Not significant	
Sample A	Significant	13	4	98.8
	Not significant	3	568	

Table 3.10: Reproducibility of results under each of the fitted two-groups models.

3.7 Discussion

In this work, we undertook an extensive exploration of the application of empirical Bayes methods for both estimating the relative protein expressions and controlling the false discovery rate. A large majority of proteomics experiments are not run in replicate. Therefore strong parametric assumptions are often the only recourse. However, since the typical sample sizes are quite large, there is no need to restrict data analyses to strictly parametric approaches. An alternative to fully parametric methods is to base analysis on empirical Bayes methods. However, current empirical Bayes methods lack flexibility and robustness when the data contain non-Gaussian tails, regions of data sparsity, excess kurtosis, or are asymmetrically distributed. Our work focuses on developing methodologies that make maximum use of the available data and maximizing the flexibility in using empirical Bayes models by fitting a richer variety of distributional components to the full shape of the class-conditional probability distributions.

Our empirical Bayes approaches are founded on the distribution of a summary statistic that accurately represents the true relative expression ratio of the proteins. For this purpose, we formulate a random effects model that allows between and within peptide heterogeneity, and propose an estimation scheme based on a variant of the minimum norm quadratic unbiased estimation (MINQUE) method. This estimation scheme should be particularly appealing to proteomics practitioners since parameter estimates can be derived without resorting to iterative procedures or specialized sta-

tistical software. It is sometimes the practice to apply a skewness or kurtosis reduction transformation on the data, and then fit the resulting symmetric distribution using a normal or normal mixture model. However, we believe that this practice should be discouraged, at least in the field of proteomics. If present, the observed skewness and kurtosis of the data provide valuable information regarding the data generating mechanisms. In proteomics, leptokurtic data are the norm, and skewness can point out unusual grouping of proteins, which for example could be the result of an imbalance in the labeling efficiency between the *light* and *heavy* labeled samples. Therefore, a more reasonable and accurate estimation procedure should consider modeling both skewness and kurtosis, as is. In our work, we propose to model the distribution of the MINQUE based estimates of relative expression ratios by considering generalized forms of normal, Student's t, and hyperbolic distributions, and where necessary finite mixtures of them. These distributions allow the fitting of skewed and kurtotic distributions, while also providing a heavier or lighter tailed fit as compared to the normal. This modeling is done under the framework of a 'two-groups model' that assumes a two component mixture model for the distribution of the estimated protein ratios.

We make use of the local false discovery strategy to identify a list of significant proteins at a pre-specified false discovery rate cut-off or at a cut-off estimated from the data. We believe that the local false discovery rate is preferable in situations where the primary interest is in identifying proteins that show some evidence of differential expression for further biological study. This is primarily because the local false discovery rate attaches an individual false discovery rate to each protein which does not depend on the significance level of other proteins. This individuality is particularly useful for calculating a combined false discovery rate for a group of proteins. For example, the false discovery rate associated with a group of proteins along the same network pathway is simply the sum of their individual local false discovery rates. This type of aggregation is not possible with the classical formulations

of a false discovery rate.

We demonstrate our methods on three protein data sets. Two of these data sets are control data sets derived from yeast. The third data set consists of proteins derived from the Hela cell line. We apply our methods to each data set separately and evaluate the performance of several two-groups models on each of them. At each stage of model fitting, we test for the appropriate number of mixture components and the goodness of fit of fitted models. '

Data analysis demonstrated the improved performance of the two-groups models fitted with asymmetric components. The combination of a mixture of truncated skew normal distributions and the generalized hyperbolic distribution was found to performs particularly well. The generalized hyperbolic distribution fits the data just as well or better than a two component mixture of skew t distributions. Since the generalized hyperbolic only requires five parameters to be estimated (as opposed to nine parameters for a two component skew t mixture), it should be preferred in situations where adequate sample size requirements are not met. None of the mixture models considered in the analysis required the fitting of more than two components.

An assessment of the number of false positives, false negatives, and the degree of agreement of results for the two control data sets also show the improved ability of asymmetric mixtures to capture important features in the data while ignoring spurious artifacts. In essence, our methods offer a compromise between over-fitting (as is the case with kernel density estimates and spline or polynomial fits), and under-fitting (as is the case with a normal or normal mixture) the data.

In summary, we developed flexible empirical Bayes methods for accurately and robustly estimating the number of significant proteins by controlling the false discovery rate locally. These methods do not require the calculation of a p-value for the estimated protein abundances or making assumptions about the distribution of the protein abundances under the alternative hypothesis of differential expression.

Furthermore, since the local false discovery strategy is only based on the marginal distribution of the estimated protein expression levels, independence of the protein expression estimates is not a strong requirement. Both skew normal, skew t and generalized hyperbolic distributions reduce to their symmetric versions, when the data are symmetric and mesokurtic. Therefore it is advisable to start a model fitting exercise with these asymmetric components, and let the data itself determine if symmetric components will suffice.

3.8 Future Work

In this chapter, we developed methodologies for the quantification and significance assessment of protein relative expression ratios, when data come from non-replicated proteomics experiments. In our future research, we propose to investigate incorporating data across replicate samples using a Bayesian hierarchical modeling approach. In this section, we present a detailed outline of this planned research.

3.8.1 Bayesian Hierarchical Modeling of Replicated SILAC Data

When a SILAC experiment is conducted in replicate, we have additional layers of data that we can use to estimate the true relative expression ratio of each protein. Let the true relative expression ratio of the j^{th} protein in the s^{th} replicated sample be denoted by μ_{js} and the sample estimate of this quantity by y_{js}; $j = 1, \ldots, m^*$, $s = 1, \ldots, S$, where S is the number of replicate samples and m^* is the number of proteins identified in common in all replicate samples. These additional layers of data can be brought into the framework discussed in Section 3.2.2 by augmenting the two level hierarchical model in (3.4) as follows

Let,

$$y_{js} \sim N(\mu_{js}, \sigma^2_{\mu_{js}}) \tag{3.46}$$

$$\mu_{js} \sim t_v(\theta_j, \varrho_j^2) \tag{3.47}$$

$$\theta_j \sim \sum_{c=1}^{C} \pi_c \, N(\eta_c, \delta_c^2) \tag{3.48}$$

where t_v is the Student's t-distribution with v degrees of freedom and θ_j is assumed to come from a Gaussian mixture distribution with C components, each having mixing probability π_c. The hierarchical setup above can be explained as follows, The replicate level estimates of relative protein expression for protein j, y_{js}'s, are thought to come from a normally distributed population with expected value equal to the true population relative expression μ_{js}, with sampling error $\sigma^2_{\mu_{js}}$. We assume that, for each replicate sample, μ_{js} and $\sigma^2_{\mu_{js}}$ are given by the estimated mean and variance given in (3.7) and (3.8), respectively. We further assume that the μ_{js}'s are in turn sampled from a t-distribution with location parameter θ_j and v degrees of freedom, where v is taken to be $S - 1$. Here we use the t distribution as a heavier tailed and more robust alternative to the normal, since S is typically small in practice. At the final stage, we assume a that the θ_j's are sampled from a normal mixture prior with mixing proportion π_c. This assumption is again based on the "two-groups model" hypothesis discussed in Section 3.2.4.1, which states that differentially expressed proteins follow a separate distribution to that of non differentially expressed proteins.

Ideally, this mixture would only contain two components, one corresponding to *null* proteins and the other corresponding to *non-null*. However, as was the case with data from non-replicated SILAC experiments, we often find that one or both of these components can in turn be better approximated using a mixture of distributions. Note that our primary interest is in the θ_j parameter, which represents the expression level of the j^{th} protein, averaged over all available data. In (5.3), we have assumed

that all component distributions are Gaussian and that the number of component distributions in the mixture, C, is not known *a priori*.

The posterior densities of the unknown parameters specified in the above hierarchical model setup are not analytically tractable. Here we have chosen to estimate them using the computationally convenient MCMC algorithm of Jung *et al.*, (2006)[61], which they used in the context of meta-analyses of microarray data. Implementation details of the Gibbs sampler based MCMC method are presented in Section 3.8.2. Let, for $j = 1, \ldots, m^*$ and $c = 1, \ldots, C$,

$$\begin{aligned} \varrho_j^2 &\sim \mathcal{IG}(\alpha, \beta) \\ (\pi_1, \ldots, \pi_c) &\sim \mathcal{D}(\alpha_1, \ldots . \alpha_C) \\ \eta_c &\sim N(a_0, b_c^2) \\ \delta_c^2 &\sim \mathcal{IG}(c_0, d_0) \end{aligned}$$

where $\mathcal{IG}$ and $\mathcal{D}$ denote the inverse gamma, and Dirichlet distributions, which are the conjugate priors for normal variances ϱ_j^2, δ_c^2; and the Multinomial weights, π_c; $\sum_{c=1}^{C} \pi_c = 1$, respectively.

3.8.2 Estimation of Model Parameters

Adopting the same convenient presentations of the t-distribution for μ_{sj}; and the mixture prior of θ_j using latent variables, as Jung *et al.*, (2006); and assuming independence for all the prior distributions of the unknown parameters, we can derive the the full conditional posterior distribution for each unknown parameter as described below.

Let,

$$\begin{aligned}
\mu_{js} \mid \xi &\sim N(\theta_j, v\xi\varrho_j^2/2) \\
\xi &\sim \mathcal{IG}(v/2, 1) \\
V_j &\sim Multinomial(1, \pi_c, \ldots, \pi_c) \\
I_{jc} &= \begin{cases} 1, & \text{if } V_j = c \\ 0, & \text{if } V_j \neq c \end{cases} \\
n_c &= \sum_{j=1}^{m^*} I_{jc} \\
\theta_j \mid I_{jc} = 1 &\sim N(\eta_c, \delta_c^2)
\end{aligned}$$

Then the joint posterior density function, $f(\boldsymbol{\Omega} \mid \boldsymbol{y})$, of all the unknown parameters,

$$\boldsymbol{\Omega} = (\{\mu_{js}\}, \xi, \{\varrho_j^2\}, \{\mu_j\}, \{V_j\}, \boldsymbol{\Lambda} = \{\pi_c, \eta_c, \delta_c^2\})$$

given the data $\boldsymbol{y} = \{y_{js}\}_{j=1,\ldots,m^*;\, s=1,\ldots,S}$, is given as

$$f(\boldsymbol{\Omega} \mid \boldsymbol{y}) \propto \prod_{j=1}^{m^*}\prod_{s=1}^{S} f(\hat{\mu}_{js} \mid \mu_{js}) \cdot \prod_{j=1}^{m^*}\prod_{s=1}^{S} f(\mu_{js} \mid \xi, \theta_j, \varrho_j^2) \cdot \pi(\xi) \cdot \prod_{j=1}^{m^*} \pi(\varrho_j^2) \cdot \prod_{j=1}^{m^*} \pi(\theta_j, V_j \mid \boldsymbol{\Lambda}) \cdot \pi(\boldsymbol{\Lambda})$$

The full conditional posterior distributions are:

$$
\begin{aligned}
[\mu_{js} \mid \mathbf{\Omega}^-] &\sim N\left(\frac{y_{js}/\sigma^2_{\mu_{js}} + (2/v\xi\varrho_j^2)\theta_j}{1/\sigma^2_{\mu_{js}} + 2/v\xi\varrho_j^2}\,,\ \left(\frac{1}{\sigma^2_{\mu_{js}}} + \frac{2}{v\xi\varrho_j^2}\right)^{-1}\right) \\
[\xi \mid \mathbf{\Omega}^-] &\sim \mathcal{IG}\left(\frac{m^*S+v}{2}, 1 + \sum_{j=1}^{m^*}\sum_{s=1}^{S}\frac{(\mu_{js}-\theta_j)^2}{v\varrho_j^2}\right) \\
[\varrho_j^2 \mid \mathbf{\Omega}^-] &\sim \mathcal{IG}\left(\frac{S}{2} + \alpha\,,\ \sum_{s=1}^{S}(\mu_{js}-\theta_j)^2/v\xi + \beta\right) \\
[\theta_j \mid V_j = c, \mathbf{\Omega}^-] &\sim N\left(\frac{(2S/v\xi\varrho_j^2)\frac{\sum_{s=1}^{S}\mu_{js}}{S} + \eta_c/\delta_c^2}{2S/v\xi\varrho_j^2 + 1/\delta_c^2}\,,\ \left(\frac{2S}{v\xi\varrho_j^2} + \frac{1}{\delta_c^2}\right)^{-1}\right) \\
[\{\pi_c\}_{c=1,\ldots,C} \mid \mathbf{\Omega}^-] &\sim \mathcal{D}\left(n_1 + \alpha_1,\ \ldots,\ n_C + \alpha_C\right) \\
[\eta_c \mid \mathbf{\Omega}^-] &\sim N\left(\frac{\sum_{j=1}^{m^*}\theta_j I_{jc}/\delta_c^2 + a_0/b_c^2}{n_c/\delta_c^2 + 1/b_c^2}\,,\ \left(\frac{n_c}{\delta_c^2} + \frac{1}{b_c^2}\right)^{-1}\right) \\
[\delta_c^2 \mid \mathbf{\Omega}^-] &\sim \mathcal{IG}\left(\frac{n_c}{2} + c\,,\ \sum_{j=1}^{m^*}(\theta_j - \eta_c)^2 I_{jc}/2 + d_0\right) \\
V_j &\sim Multinomial(1, p_{j1},\ \ldots,\ p_{jC}) \\
p_{jc} = P(V_j = c \mid \mathbf{\Omega}^-) &= \frac{\pi_c\phi(\theta_j;\eta_c,\delta_c^2)}{\sum_{t=1}^{C}\pi_t\phi(\theta_j;\eta_t,\delta_t^2)}
\end{aligned}
$$

where $\mathbf{\Omega}^-$ represents all unknown parameters in Ω minus the parameter(s) being conditioned upon, and $\phi(.;\eta,\delta^2)$ is the normal density function of $N(\eta,\delta^2)$.

3.8.3 Estimating the local fdr

If we restrict the the number of mixture components, C, in (3.48) to be just two, i.e.,

$$\theta_j \sim \pi_0 N(\eta_0, \delta_0^2) + \pi_1 N(\eta_1, \delta_1^2),$$

then this extended hierarchical setup falls conveniently into the local fdr framework given in Section 3.2.4.2. Then, from the model in (3.12) and by plugging in the posterior estimates, the local false discovery rate associated with protein j, given

that $\theta_j = \theta^*$ is given as

$$locfdr_j(\theta^*) = \frac{\hat{\pi}_0\, \phi(\theta^*; \hat{\eta}_0, \hat{\delta}_0^2)}{\left(\hat{\pi}_0\, \phi(\theta^*; \hat{\eta}_0, \hat{\delta}_0^2) + \hat{\pi}_1\, \phi(\theta^*; \hat{\eta}_1, \hat{\delta}_1^2)\right)}. \tag{3.49}$$

Chapter 4

Resampling Based Methods for Identifying Differentially Expressed Proteins using XIC Area

4.1 Introduction

The robustness of the results of proteomics data analyses is greatly affected by the inherent variability in LC/MS-MS based proteomics strategies. Even under a controlled setting, the observed protein profiles from two replicates of the same protein sample will differ, both in terms of the set of proteins identified and in their observed signal intensity levels. Typically, low-abundance proteins may or may not be detected in a given MS run, and higher abundance proteins are detected at varying levels in different runs. For instance, a reproducibility study by Durr *et al.*, (2004)[34] performed on rat lung endothelial cell plasma membranes concluded that ten replicate runs would be necessary to reach completeness of protein detections with 95% confidence. As discussed in Section 1.5, this variability is in part due to the nature of proteins themselves (biological variability) and also in part to the variability of MS as

a high throughput technique (technical variability). Furthermore, the processes and interactions between processes responsible for both biological and technical variability are not well understood. Although the need for replicate samples in proteomics have been widely noted, the increased costs and time required for stable isotope labeling of samples has meant that most labeling-based proteomics data come from non replicated experiments. Standard statistical analyses of these data sets, often do not account for the cumulative effects of the underlying biological and technical variability in the data, yielding results that are hard to reproduce. From a statistical point of view, attempting to quantify this error, for example through a error propagation model, has two major road blocks: availability of data from a properly designed experiment that consists of both technical and biological replicate samples, and the fact that any such estimated model would be very hard to generalize since the error processes are specific to the lab, protein sample, and the MS platform that produced the data. Much of this variability can be reduced through careful quality control at each stage of the experiment; by removing 'unreliable' data points; and by using a more stable method to quantify the peptide level relative expression ratios.

A drawback of the type of statistical analysis of relative protein abundances discussed in Chapter 1 is that the analysis is based only on 'complete' peptide and protein level data. For example, in any given gel fraction, if a subset of constituent peptides of a particular protein only have either the light or the heavy signal observed, or if both signals are not observed, then these peptides would be removed from further analysis. This seems inefficient considering that these same set of peptides are likely to be fully observed in a number of other gel fractions.

In the above context, we propose in this chapter a set of nonparametric statistical methodologies that are based on capturing the inherent variability in labeling based proteomics data through class preserving resampling. In these methods, our interest is in estimating an 'error - adjusted' expression level for each identified protein, through

resampling of all available peptide level data. Prior to the resampling analysis, we pre-process the data to remove 'less reliable' data points, through a bivariate mixture model based cluster analysis. We note here that this reliability analysis step can precede any type of analysis involving labeling based proteomics data. We then proceed to introduce a method for quantification of relative expression ratios, through a 'area under the curve' approach that for a given peptide, down-weights data points that are observed farther away from the highest observed peak in its MS/MS elution profile. Finally, we develop two resampling based strategies for estimating the overall relative expression ratio of a protein, a method for calculating a p-value for quantifying the significance of each protein, and a two-groups model based local false discovery rate estimation procedure.

4.2 Reliability Analysis of SILAC Data

Not all the data that results from a proteomics experiment are equally reliable. Low reliability data could be the result of many factors that affected the data starting from sample preparation to preprocessing. However, unreliable data are mostly a by product of low signal intensities. In other words, signals that meet a pre-specified signal-to-noise ratio cutoff, but by only a small margin. The use of less reliable data is a major reason that affects the reproducibility of statistical data analyses results in proteomics. Therefore the use of a subset of the available data that are deemed more reliable could potentially mean that more accurate and robust ratios of protein expression can be obtained.

In this section, we propose a statistical approach based on bivariate normal mixture modeling to separate the available set of data into two groups, namely groups of reliable and unreliable data. This type of reliability analysis was proposed by Asyali et al. (2004)[7], in the context of testing the reproducibility of two dye channel

(red and green) data resulting from cDNA microarray experiments. We adapt this methodology to SILAC data, by recognizing that the (light, heavy) signal streams can be treated the same way as the dye channels in cDNA microarray experiments. We note here that, even though our adaptation of this method is based on SILAC data, it would work equally well for any type of proteomics data generated using a labeling strategy.

Adopting the same nomenclature and data structure that was described in Section 3.2.1, we again denote the peptide level signal intensity pairs corresponding to protein j by $\mathcal{R}^j = \left\{ \frac{L_h^{jkr}}{H_h^{jkr}}; k = 1, \ldots, m_j, r = 1, \ldots, m_{jk}, h = 1, \ldots, m_{jkr} \right\}$. Let, $I_{jk} = \sum_{r=1}^{m_{jk}} m_{jkr}$ denote the total number of intensity pairs available for the jk^{th} peptide across its m_{jk} repeat occurrences, and $I_j = \sum_{k=1}^{m_j} I_{jk}$ denote the total number of intensity pairs available for protein j across the m_j peptides that are derived from protein j. Now let $\mathbf{y}_t = (l_t, h_t), t = 1, \ldots, I_j$ be the 2-dimensional random sample of size I_j. As mentioned in the previous section, not all of the I_j signal intensity pairs are equally reliable. If we assume that the I_j signal intensity pairs come from two groups; one group corresponding to signal pairs that have higher reliability compared to the other, then we can frame this scenario within the framework of a bivariate normal mixture model. Under this approach, each intensity pair is assumed to be a realization of the random 2-dimensional vector $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{I_j})$ with the 2-component bivariate normal mixture probability density function

$$f(\mathbf{y}; \boldsymbol{\Psi}) = \sum_{g=1}^{2} \pi_g \phi(\mathbf{y}; \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}), \tag{4.1}$$

where, $\boldsymbol{\Psi} = (\pi_1, \pi_2; \boldsymbol{\theta}')'$, π_1, π_2 are the mixing weights of the two component densities each representing a reliability group (either high or low); and $\boldsymbol{\theta}$ consists of the elements of $\boldsymbol{\mu_g}$ and the distinct elements of the $\boldsymbol{\Sigma_g}$, for $g = 1, 2$; and $\phi(\mathbf{y}; \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}) = (2\pi)^{-1}|\Sigma_g|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu_g})^T \boldsymbol{\Sigma_g}^{-1}(\mathbf{y} - \boldsymbol{\mu_g})\right\}$.

We estimate (4.1) using the EM algorithm, with initial estimates of the parameters obtained by a simple application of the k-Means clustering algorithm. From the estimated mixture model, the posterior probability of group membership of each observation $\mathbf{y}_t$ can be derived as

$$P\left(\text{observation}\,\mathbf{y}_t \in \text{group}\ g\right) = \hat{\pi_{gt}} = \frac{\hat{\pi}_g \phi(\mathbf{y_t}; \hat{\boldsymbol{\mu}}_{\boldsymbol{g}}, \hat{\mathbf{\Sigma}}_{\mathbf{g}})}{\sum_{g=1}^{2} \hat{\pi}_g \phi(\mathbf{y_t}; \hat{\boldsymbol{\mu}}_{\boldsymbol{g}}, \hat{\mathbf{\Sigma}}_{\mathbf{g}})}, \tag{4.2}$$

where $t = 1, \ldots, I_j$ and $g = 1, 2$.

Additionally, we can obtain a decision boundary for group membership by equating the group posterior probabilities and solving for $\mathbf{y}_t$ for $t = 1, \ldots, I_j$. I.e., the decision boundary is the set of all points $\mathbf{y}$ such that $\hat{\pi}_1 \phi(\mathbf{y}; \hat{\boldsymbol{\mu}}_1, \hat{\mathbf{\Sigma}}_{\mathbf{1}}) = \hat{\pi}_2 \phi(\mathbf{y}; \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{\Sigma}}_{\mathbf{2}})$. If we represent the (l, h) signal intensity pairs on a 2-dimensional Cartesian coordinate system, with the l-signal on the y-axis and the h-signal on the x-axis, then this decision boundary typically takes the form of a hyper-ellipsoid, lying mostly on the 45° line, above the 45° line, or below it, depending on whether or not the protein under consideration is non-differentially expressed, over expressed in the light isotope labeled sample, or under expressed in the light isotope labeled sample, respectively. Now let,

$$\delta(\mathbf{y}, \boldsymbol{\mu}; \mathbf{\Sigma}) = (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \tag{4.3}$$

$$\tilde{\pi} = \max\left(\hat{\pi_1}, \hat{\pi_2}\right), \tag{4.4}$$

where (4.3) denote the Mahalanobis squared distance between $\mathbf{y}$ and $\boldsymbol{\mu}$ (with $\mathbf{\Sigma}$ as the covariance matrix). Then we treat an observation $\mathbf{y_t}$ as unreliable if the quantity $\delta(\mathbf{y_t}, \hat{\boldsymbol{\mu}}_{\tilde{\pi}}; \hat{\mathbf{\Sigma}}_{\tilde{\pi}})$ exceeds the 90^{th} percentile of the chi-squared distribution with 2 degrees of freedom, where $\hat{\boldsymbol{\mu}}_{\tilde{\pi}}$, and $\hat{\mathbf{\Sigma}}_{\tilde{\pi}}$ represent the mean vector and covariance matrix of the more reliable group. Unreliable pairs of intensity signals thus identified, are then removed from the original set of scans, $\mathcal{R}^j$. After repeating the above procedure for

all proteins, we are left with a 'reduced in size, but improved in quality' data set, which we will make use of in all down-stream analyses.

We note here that this approach can be made more robust by using mixtures of bivariate t distributions instead of the bivariate normal. In our data, it is the norm that not all the proteins in any given data set will have sufficiently large number of scans to allow the reliable fitting of a bivariate t mixture. However, when sufficient sample sizes are available, extending this analysis to a t mixture framework is fairly straight forward, with model fitting easily achieved by the augmented EM based methods of either Liu and Rubin (1995)[75], or McLachlan and Peel (2000)[86].

4.3 Evaluation of the Protein Relative Expression Ratio using Extracted Ion Current (XIC) Area

Up to now, we have used individual chromatographic peaks to calculate the relative expression ratios of peptides. I.e., given a heavy and light signal intensity pair corresponding to a particular peptide, an estimate of the relative expression ratio of that peptide is given by the actual ratio of the heavy and light chromatographic peaks. However, the variability of the relative expression ratios estimated using individual peak ratios is still quite high, even when all signal pairs are derived from the same peptide and same gel fraction. This variability can be reduced by the elimination of less reliable signal pairs, as discussed in Section 4.2. We can further reduce this variability by considering ratios of chromatographic peak areas (or ion current areas), instead of ratios of individual peaks. The relative expression ratio of a peptide is then estimated using the ratio of the chromatographic peak areas of heavy and light signals eluting over a range of time in which chromatographic peaks corresponding to that peptide were detected by the mass spectrometer. Before calculating the chromatographic peak area for either the heavy or the light signals, we obtain a smooth

demarcation of the chromatographic peak profile using a second order Savitzky-Golay smoothing filter [97]. This smoothing filter was also used in Li *et al.*, (2003)[66] and Ryu *et al.*, (2008)[107] for estimating the peptide ion current area.

4.3.0.1 The Savitzky-Golay smoothing filter

The premise of data smoothing is that one is measuring a variable that is both slowly varying and also corrupted by random noise. This is exactly the case with the typical peak elution pattern, or the extracted ion chromatogram (XIC; peptide ion signal as a function of elution time), of peptides. As illustrated in Figure 4.1, the XIC starts at around the 36.6 minute mark, peaking at around the 36.8 mark and completely disappears at around the 37.0 minute mark. During this 0.4 minute interval, five (light, heavy) signal intensity pairs are observed. The five individual (h/l) ratios are all estimating the same relative peptide expression ratio, with possible added variation introduced by the slowly varying growth and decay rates and random (experimental) noise present in the XIC. The ratio given by, $\frac{\text{area (red curve)}}{\text{area (black curve)}}$, is expected to provide a much more reliable estimate of the true peptide relative expression ratio, compared to simply picking the highest peak in the elution profile. Simply put, the purpose of using a smoothing filter is to replace the separate peaks shown in Panel A with a smooth curve as in Panel B.

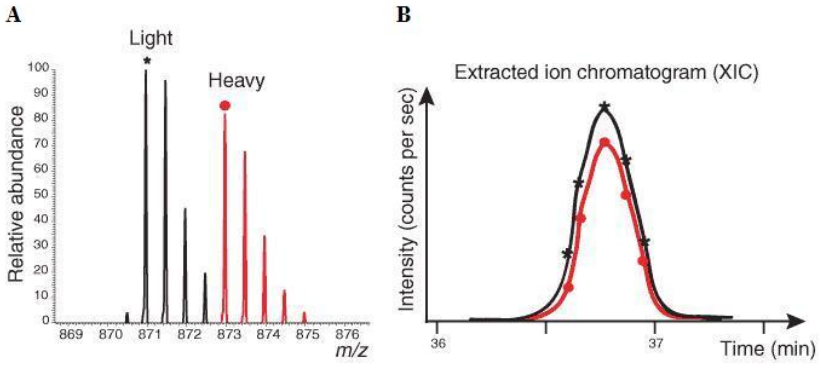


Figure 4.1: Example of an Extracted Ion Chromatogram (XIC).

Let, $f_i \equiv f(t_i)$, represent a set of equally spaced data values, where $t_i \equiv t_0 + i\,\Delta$ for some constant sample spacing Δ and i is represents a sequence that extends symmetrically in opposite directions of a pre-specified point. A smoothing filter basically replaces each data value f_i by a linear combination g_i of itself and some number of it's closest neighbors,

$$g_i = \sum_{n=-n_L}^{n_R} c_n f_{i+n}, \tag{4.5}$$

where, n_L, n_R are the number of points used to the left and right of a given datum i.

The basic idea behind the Savitzky-Golay filtering (Savitzky and Golay, 1964[109]) is to find filter coefficients c_n, by approximating the underlying function by a polynomial of higher order, typically quadratic or quartic, within a moving window of fixed size. For each point f_i, a polynomial is fit using least squares fitting to all $n_L + n_R + 1$ points in the moving window, and g_i is set to be the value of the fitted polynomial at position i. Starting with f_{-n_L}, this procedure is repeated, each time within a shifted window from the previous, until the procedure terminates at the point f_{n_R}. With SILAC data n_L, n_R is usually at least three, yielding at a minimum seven data points for the fitting procedure.

4.3.0.2 Estimating the relative expression ratio using XIC area

For our data, we use a Savitzky-Golay filter with third-order polynomial smoothing and a moving window of width nine for both light and heavy XIC's. This essentially means that we replace the observed intensity peak at Time = t^*, say, with a fitted value obtained from a polynomial fit to nine neighboring points, which includes the peak at Time = t^* and the four closest peaks to the left and right of Time = t^*. For each XIC, we select t_0 to be the point that corresponds the highest observed peak. This allows the Savitzky-Golay filter to capture the central region of the XIC, which

has the desirable properties of: having the most reliable peaks, since they have large signal-to-noise ratios; and having the least variable peaks, since growth/decay rates are typically slower near the center of the elution profile. Now let, $\tilde{f}_l$, $\tilde{f}_h$ denote the smoothed estimates of the light and heavy XIC's, respectively. Then the peptide expression levels for light and heavy samples can be determined by numerical estimation of the chromatographic peak areas of $\tilde{f}_l$, and $\tilde{f}_h$. The relative expression ratio of the peptide is then given by the ratio $\frac{\text{area}\,(\tilde{f}_h)}{\text{area}\,(\tilde{f}_l)}$. Usually these ratios are log 2-base transformed, for reasons mentioned in Section 3.2.2.

4.4 Resampling Based Estimation of Overall Protein Relative Expression using XIC area

Let $\tilde{A}_{jkr}$ be the XIC area ratio derived from the m_{jkr} signal intensity pairs corresponding to Pp^{jkr}, where $\tilde{A}_{jkr} = \log_2\left(\frac{\text{area}\,(\tilde{f}^h_{jkr})}{\text{area}\,(\tilde{f}^l_{jkr})}\right)$, and $\tilde{f}^l_{jkr}$, $\tilde{f}^h_{jkr}$ are the estimated smooth profiles of the light and heavy XIC's, respectively. Now let $\mathfrak{R}_{jk}$ denote the set of all XIC area ratios for Pp^{jk} and $\mathfrak{R}_j$ denote the union of all such sets for protein j. It now remains for us to estimate an overall relative expression ratio for protein j using all peptide level area ratios in the set $\mathfrak{R}_j$.

Let $\hat{w}_{jk} = \frac{|\mathfrak{R}_{jk}|}{|\mathfrak{R}_j|}$, where $|.|$ represents the cardinality of the set and $\sum_{k=1}^{m_j} w_{jk} = 1$. Then $\hat{w}_{jk}$ is indicative of the empirical probability of observing peptide k as being derived from protein j, under the experimental setup that generated this set of XIC area ratios. The higher the value of $\hat{w}_{jk}$, the higher the chance that peptide k is more 'proteotypic'. I.e, that this peptide has certain desirable properties, either physicochemical and/or experimental, that makes peptide k more likely to be observed for protein j, compared to a peptide that has a smaller $\hat{w}$. This concept of the existence of 'proteotypic' peptides was first espoused by Mallick *et al.*, (2007)[77].

In subsequent discussions, we make use of the following notations and definitions.

Let, for some protein j,

$$\mathfrak{R}_{\mathfrak{j}} = \{\tilde{A}_{jkr} : k = 1, \ldots, m_j; r = 1, \ldots, m_{jk}\}, \tag{4.6}$$

$$\mathfrak{w}_{\mathfrak{j}} = \{\hat{w}_{jk} : k = 1, \ldots, m_j\}, \tag{4.7}$$

where, $\mathfrak{R}_{\mathfrak{j}}$ denotes the set of all available peptide level XIC based relative expression ratios for protein j, and $\mathfrak{w}_{\mathfrak{j}}$ denotes the set of estimated weights corresponding to each unique peptide in set $\mathfrak{R}_{\mathfrak{j}}$.

If only a single peptide was found for a particular protein, i.e, $m_u = 1$ for some protein u and if that peptide was observed only once across all gel fractions, then we simply pass on the calculated XIC area ratio for that peptide as the final estimate of the overall relative expression ratio for protein u for all down-stream analyses. When there is more than one peptide available for a protein j, i.e., when $m_j > 1$), and each of these peptides is observed multiple times, then we take the overall average of all available peptide level relative expression ratios as the overall estimate, $\hat{\theta}_j$, of the true relative expression ratio, θ_j, of protein j,

$$\hat{\theta}_j = \frac{\sum_{k=1}^{m_j} \sum_{r=1}^{m_{jk}} \tilde{A}_{jkr}}{|\mathfrak{R}_j|} \tag{4.8}$$

4.4.1 Estimation of Relative Protein Expression using a Bootstrap Partial Maximum Likelihood Estimator (BPMLE)

The bootstrap partial likelihood (BPL) method (Davison *et al.*, (1992)[31], affords a way for us to seek an approximate likelihood function for θ_j, $p\left(\hat{\theta}_j|\theta_j\right)$, using a nested bootstrap procedure. In other words, we seek an estimate of the sampling distribution of $\hat{\theta}_j$, when the true parameter is θ_j. The BPMLE method proceeds as follows. First, we draw B_1 first stage bootstrap samples $\mathfrak{R}_{\mathfrak{j}}^{*1}$, $\ldots$, $\mathfrak{R}_{\mathfrak{j}}^{*B_1}$, from $\mathfrak{R}_{\mathfrak{j}}$.

Each bootstrap sample is drawn in two steps. First, we draw m_j peptide blocks with replacement, where the k^{th} peptide block consists of the m_{jk} ion current ratios corresponding to Pp^{jk}, and has probability $\hat{w}_{jk}$ of being drawn. If the k^{th} peptide block is selected in the peptide block sampling step, then from that peptide block, we again draw with replacement a random sample of size m_{jk}. At the end of this process, we obtain the first stage bootstrap replications $\hat{\theta}_j^{*1}, \ldots, \hat{\theta}_j^{*B_1}$. Next, from each of the first stage bootstrap samples, $\mathfrak{R}_\mathfrak{j}^{*b}, b = 1, \ldots, B_1$, we generate B_2 second stage bootstrap samples, again following the same two step drawing procedure. These second stage bootstrap replicates are denoted as $\hat{\theta}_{bj}^{**1}, \ldots, \hat{\theta}_{bj}^{**B_2}$. Note that our analysis here differs from a typical BPL construction in the sense that we are drawing weighted samples at both the first and second stages, with drawing probabilities taken to be given by $\mathfrak{w}_\mathfrak{j}$. Consequently, these weighted bootstrap samples are more representative of the actual phenomena governing the size and characteristics of the peptide complement that is detected in any given MS run. Finally, we form the kernel density estimates

$$\hat{p}\left(\theta_j|\hat{\theta}_j^{*b}\right) = \frac{1}{B_2 s}\sum_{t=1}^{B_2} k\left(\frac{\theta_j - \theta_{bj}^{**t}}{s}\right), \text{ for } b = 1, \ldots, B_1, \tag{4.9}$$

where, $k(.)$ is any kernel function with window width s. In this work, we use a standard normal density kernel, and estimate the optimal bandwidth s, by the version of *solve-the-equation plug-in* method proposed by Sheather and Jones (1991)[112]. Next, we evaluate $\hat{p}(\theta_j|\hat{\theta}_j^{*b})$ for $\theta_j = \hat{\theta}_j$, $b = 1, \ldots, B_1$. Each $\hat{p}(\hat{\theta}_j|\hat{\theta}_j^{*b})$ provides an estimate of the likelihood of θ_j for parameter value $\theta_j = \hat{\theta}_j^{*b}$. We then apply a *loess* smoother to the pairs $[\hat{\theta}_j^{*b}, \hat{p}(\hat{\theta}_j|\hat{\theta}_j^{*b})], b = 1, \ldots, B_1$, to get a smooth estimate of the likelihood, $\mathcal{L}$. The BPL-based estimate of the overall relative expression ratio of protein j is then given by $\hat{\theta}_j^{\text{BPLME}} = \text{argmax}_{\theta_j}\, \mathcal{L}$.

4.4.2 Estimation of Relative Protein Expression using a Model-based Bootstrap

In this section, we make use of covariate information, viz. four physicochemical/experimental properties of peptides, to generate bootstrap samples that are 'adjusted' for these covariates. For the k^{th} peptide derived from the j^{th} protein, we denote these four properties as: CH_{jk} - the charge associated with each peptide , TP_{jk} - trypticity (i.e., the extent to which the peptide is digested in the mixture), LN_{jk} - the length of the its amino acid sequence, and XC_{jk} - the cross correlation score for the peptide outputted by The SEQUEST algorithm.

Let, $\varepsilon_{jk} = \tilde{A}_{jk} - \theta_j$ represent the error in estimating the overall relative protein expression ratio, θ_j, using the peptide level estimator $\tilde{A}_{jk} = \frac{\sum_{r=1}^{m_{jk}} \tilde{A}_{jkr}}{|\mathfrak{R}_{jk}|}$. Ideally, this error should be a white noise process. However, as discussed previously, a major part of this variation is due to physicochemical/experimental properties of peptides themselves and the inherent variability in mass spectrometry based methods. Let $\tilde{\varepsilon}_{jk} = \tilde{A}_{jk} - \hat{\theta}_j$ be the dependent variable in the following linear model

$$\tilde{\varepsilon}_{jk} = \beta_0 + \beta_1 \,\mathrm{CH}_{jk} + \beta_2 \,\mathrm{TP}_{jk} + \beta_3 \,\mathrm{LN}_{jk} + \beta_4 \,\mathrm{XC}_{jk} + \epsilon_{jk} \;\; ; \;\; k = 1, \ldots, m_j \tag{4.10}$$

$$= \mathbf{X}'_{jk}\boldsymbol{\beta} + \epsilon_{jk} \tag{4.11}$$

and let the fitted model be

$$\tilde{\varepsilon}_{jk} = b_0 + b_1 \,\mathrm{CH}_{jk} + b_2 \,\mathrm{TP}_{jk} + b_3 \,\mathrm{LN}_{jk} + b_4 \,\mathrm{XC}_{jk} + e_{jk} \;\; ; \;\; k = 1, \ldots, m_j \tag{4.12}$$

$$= \mathbf{X}'_{jk}\mathbf{b} + e_{jk} \tag{4.13}$$

The accuracy of the least squares estimates, $\mathbf{b}$, depend on the distribution of ϵ_{jk} being normal. When this is not the case, in particular if the error distribution is heavier tailed compared to the normal, we need to consider a fitting criterion that is

more robust to unusual data.

4.4.2.1 Robust regression using M-estimation

A convenient method of robust regression is M-estimation, introduced by Huber (1964)[59]. The general M-estimator is calculated by minimizing a pre-specified objective function

$$\sum_{k=1}^{m_j} \delta\left(e_{jk}\right) = \sum_{k=1}^{m_j} \delta\left(\tilde{\varepsilon}_{jk} - \mathbf{X}'_{jk}\mathbf{b}\right) \tag{4.14}$$

where the function δ represents the contribution of each residual to the objective function. Note that the standard least squares estimation corresponds to, $\delta\left(e_{jk}\right) = e_{jk}^2$.

Now let ψ be the derivative of δ, the weight function $\omega(e) = \psi(e)/e$, and $\omega_{jk} = \omega(e_{jk})$. Then for the coefficients, $\mathbf{b}$, in (4.21), we can form a system of five estimating equations

$$\sum_{k=1}^{m_j} \omega_{jk}\left(\tilde{\varepsilon}_{jk} - \mathbf{X}'_{jk}\mathbf{b}\right)\mathbf{X}'_{jk} = 0 \tag{4.15}$$

Solving the estimating equations is a weighted least-squares problem involving the minimizing of $\sum_k \omega_{jk}^2 e_{jk}^2$, and can be achieved by the iteratively reweighted least-squares (IRLS) method.

The M-estimator we consider for our data is the Tukey bi-square (or biweight)

estimator. The objective and weight functions corresponding to this estimator are

$$\delta(e) = \begin{cases} \frac{\lambda^2}{6}\left\{1 - \left[1 - \left(\frac{e}{\lambda}\right)^2\right]^3\right\}, & \text{for } |e| \leq \lambda \\ \frac{\lambda^2}{6}, & \text{for } |e| > \lambda, \end{cases} \tag{4.16}$$

$$\omega(e) = \begin{cases} \left[1 - \left(\frac{e}{\lambda}\right)^2\right]^2, & \text{for } |e| \leq \lambda \\ 0, & \text{for } |e| > \lambda, \end{cases} \tag{4.17}$$

where λ is a tuning constant governing the level of resistance to outliers. Smaller values of λ produce more resistance to outliers, but at the expense of lower efficiency when the errors are normally distributed. In practice, λ is set at $4.685\hat{\sigma}$, where $\hat{\sigma}$ is the standard deviation of the errors. This value of λ yields 95-percent efficiency when the errors are normal, while still offering reasonable protection against outliers. Note that for the bi-square estimator the objective function eventually levels off (for $|e| > \lambda$) and the weights decline as soon as e diverges from 0, and are 0 for $|e| > \lambda$.

Let the set of predicted values for protein j from (4.20) be $\hat{\tilde{\boldsymbol{\varepsilon}}}_j = \{\hat{\tilde{\varepsilon}}_{j1}, \ldots, \hat{\tilde{\varepsilon}}_{jm_j}\}$. We now use the same procedure described in Section 4.4.1, but now, instead of drawing first stage weighted samples from $\mathfrak{R}_j$, we draw random samples from $\hat{\tilde{\boldsymbol{\varepsilon}}}_j$. After the b^{th} first stage draw, we re-construct the b^{th} first stage sample ${\mathfrak{R}_j}^{*b}$ as $\hat{\theta}_j + \hat{\tilde{\varepsilon}}_{jk}$, $k = 1, \ldots, m_j$. Upon completion of this process for B_1 first stage samples and B_2 second stage samples, we obtain the first stage bootstrap replications $\hat{\theta}_j^{*1}, \ldots, \hat{\theta}_j^{*B_1}$, and second stage bootstrap replicates $\hat{\theta}_{bj}^{**1}, \ldots, \hat{\theta}_{bj}^{**B_2}$, for $b = 1, \ldots, B_1$. Subsequent steps needed for the estimation of the BPL estimate, $\hat{\theta}_j^{\text{BPLME}}$, of the overall relative expression ratio of protein j, are identical to those that were presented under Section 4.4.1.

4.4.2.2 Influence of covariates on protein expression estimation

As an interesting aside, we also investigate the influence of the covariates CH, TP, LN, and XC on the overall protein expression ratio estimates. This is done by a standard bootstrap M-estimation of the peptide level data available for all the proteins on the covariates. I.e., we extend the regression model (4.10) to include all m identified proteins. This extended regression model is

$$\tilde{\varepsilon}_{jk} = \beta_0 + \beta_1 \,\mathrm{CH}_{jk} + \beta_2 \,\mathrm{TP}_{jk} + \beta_3 \,\mathrm{LN}_{jk} + \beta_4 \,\mathrm{PC}_{jk} + \epsilon_{jk} \;\; ; \;\; j = 1, \ldots, m \,; \, k = 1, \ldots, m_j \tag{4.18}$$

This regression model is particularly useful as a baseline reference for the behavior of null proteins. We can compare the results of a regression done on the data from a single protein (using model **??**) against this reference. A protein that is behaving in a manner that is inconsistent with the reference can be considered for further investigation.

4.4.3 p-value Estimation and FDR

If we let $z_j = \hat{\theta}_j$ for $j = 1, \ldots, m$, then we could adopt any of the methods of Chapter 3, to fit a two-groups model to the z values, leading to the estimation of a local false discovery rate for each protein. Alternatively, we can construct a bootstrap p-value for each statistic $\hat{\theta}_j$, and use the estimated p-values to control for multiple testing by any of the standard false discovery rate methods. For example, we can use the classical FDR method of Benjamini and Hochberg (1995)[16], or the q - value method of Storey (2003)[113]. In this work, we have chosen to remain within the two-groups model framework, whereby we estimate local false discovery rates using the nested-bootstrap p-values of the $\hat{\theta}_j$'s.

4.4.3.1 A p-value based on the nested-bootstrap samples

We can construct a bootstrap p-value for testing the two sided hypothesis $H_0 : \theta_j = 0$, by using the B_1 first stage bootstrap replications $\hat{\theta}_j^{*1}, \ldots, \hat{\theta}_j^{*B_1}$ as

$$\hat{p}_j^* = 2 \min \left(\frac{1}{B_1} \sum_{b=1}^{B_1} I(\hat{\theta}_j^{*b} < 0) \, , \, \frac{1}{B_1} \sum_{b=1}^{B_1} I(\hat{\theta}_j^{*b} > 0) \right) \tag{4.19}$$

In (4.19), we calculate the p-values for one-tailed tests in each tail and reject H_0 if either of these p-values is less than $\alpha/2$, where α is the level of the test. Note that this formulation of the bootstrap p-value does not assume θ_j is symmetric around zero in finite samples. The p-value in (4.19) can be improved upon, by making use of the information afforded by the B_2 second stage samples. This approach of using the second stage samples to estimate a p-value is called the nested-bootstrap or the double-bootstrap p-value (Beran, 1988)[17]. The method proceeds as follows. For each first-level bootstrap sample indexed by b, we can compute the second-stage bootstrap p-value

$$\hat{p}_{bj}^{**} = 2 \min \left(\frac{1}{B_2} \sum_{l=1}^{B_2} I(\hat{\theta}_{bj}^{**l} < \hat{\theta}_j^{*b}) \, , \, \frac{1}{B_2} \sum_{l=1}^{B_2} I(\hat{\theta}_{bj}^{**l} > \hat{\theta}_j^{*b}) \right) \tag{4.20}$$

The p-value in (4.20) corresponds to the bootstrap replicate $\hat{\theta}_j^{*b}$, based on the empirical distribution of the second level replicates $\hat{\theta}_{bj}^{**l}$; $l = 1, \ldots, B_2$. We then use the $\hat{p}_{bj}^{**}$ to calculate the nested-bootstrap p-value as

$$\hat{p}_j^{**} = \frac{1}{B_1} \sum_{b=1}^{B_1} I(\hat{p}_{bj}^{**} \leq \hat{p}_j^*). \tag{4.21}$$

Therefore the nested-bootstrap p-value, $\hat{p}_j^{**}$, is equal to the proportion of the second-level bootstrap p-values that are smaller (and hence more extreme) than the first-level bootstrap p-value.

The nested-bootstrap p-value is an improvement on the standard bootstrap p-value in the following heuristic sense. If the bootstrapping process causes the distribution of the $\hat{\theta}_j^{*b}$ to contain fewer extreme values than the distribution of θ_j itself, then the p-values associated with moderately extreme values of $\hat{\theta}_j$ would tend to be too small. However, we can still reasonably expect that the distributions of the $\hat{\theta}_{bj}^{**l}$ would contain even fewer extreme values than the distribution of the $\hat{\theta}_j^{*b}$. Therefore, the $\hat{p}_{bj}^{**}$ should tend to be too small, at least for small values of $\hat{p}_j^{*}$. The implication being that the nested-bootstrap p-value $\hat{p}_j^{**}$ will be larger than $\hat{p}_j^{*}$, which is what we want. Similarly, $\hat{p}_j^{**}$ will tend to be smaller than $\hat{p}_j^{*}$ when the distribution of the $\hat{\theta}_j^{*b}$ contains more extreme values than the distribution of θ_j.

4.4.3.2 Local False Discovery Rate Estimation

The two-groups model that we use is the same as that of Allison *et al.*, 2002[3]; the only difference being that instead of traditional p-values derived from a t-test like statistic, we use nested-bootstrap derived p-values to construct our mixture model. Under this approach, we model the distribution of the nested-bootstrap p-values, $\hat{p}_j^{**};\ j = 1, \ldots, m$, as a mixture of a uniform distribution on [0,1], with weight π_0, for proteins that are non differentially expressed and a Beta (a, b) distribution, with weight $1 - \pi_0$, for proteins that are differentially expressed. This particular mixture model has been shown to perform exceptionably well with p-values in previous studies (Pounds and Morris (2003)[96], Allison *et al.*, (2002)[3], Pan (2002)[93]). For simplicity of notation, let $p = \hat{p}^{**}$. Then, using (3.10), the density function for all nested-bootstrap p-values can be given as

$$f(p; a, b) = \pi_0 + (1 - \pi_0)\frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1 - p)^{b-1} \tag{4.22}$$

Method of moments estimates of a, b, which serve as starting values, are given by

$$a_{mom} = \bar{p}\left(\frac{\bar{p}(1-\bar{p})}{v} - 1\right) \tag{4.23}$$

$$b_{mom} = (1-\bar{p})\left(\frac{\bar{p}(1-\bar{p})}{v} - 1\right), \tag{4.24}$$

where $\bar{p}$ gives the mean p-value and $v = \frac{m-1}{m}\text{Var}(p)$. Now using EM estimates of (π_0, a, b) and definition (3.12), the local false discovery rate corresponding to protein j can be calculated as

$$locfdr_j(p*) = \frac{\hat{\pi}_0}{\hat{\pi}_0 + (1-\hat{\pi}_0)\text{Beta}(p^*; \hat{a}, \hat{b})} \tag{4.25}$$

Since, by definition, $locfdr_j(p^*)$ gives the posterior probability of being 'null' given that we observed $p_j = p^*$, we can choose a probability cut-off (say, 0.05), and choose all proteins with a local false discovery rate below this cut-off to select a list of proteins that are significantly differentially expressed between the light and heavy isotope labeled samples.

4.5 Results

We begin by selecting a more reliable set of data points for each protein through the bivariate mixture model based clustering method presented in Section 4.2. The resulting cluster plot for a randomly chosen protein from Sample A, *YPL240C*, is shown in Figure 4.2. As expected for a non-differentially expressed protein, the hyper-ellipsoid demarcating the class membership decision boundary lies primarily on the 45° line. Note that if we divide the scatter plot of the data into four quadrants based on the average noise level, then we expect that most of the reliable data points would lie within the first quadrant (top-right). This quadrant corresponds to signal pairs that have the highest signal-to-noise ratio. Conversely, most of the less reliable

data points should lie within the third quadrant (bottom-left). We select a data point as unreliable if its Mahalanobis squared distance from the mean of the more reliable cluster exceeds the 90^{th} percentile of the chi-squared distribution with 2 degrees of freedom. For *YPL240C*, this criterion identified 19 data points (out of 271 total) as less reliable. The plot of cluster membership scores (figure 4.3) shows that there are only a few data points with nearly equal scores.

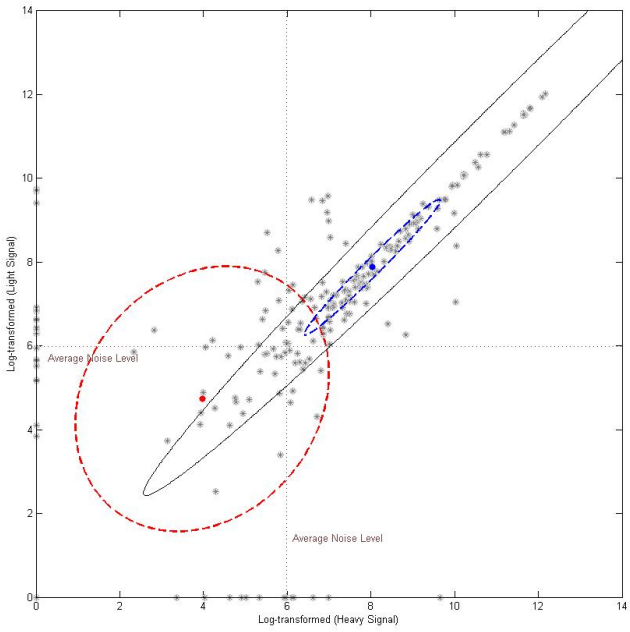


Figure 4.2: Choosing a reliable subset of the data. Red and blue ellipses indicate the two identified clusters; solid black contour indicates the hyper-ellipsoid demarcating the class membership decision boundary; vertical and horizontal lines represent the average noise level in the *heavy* and *light* signals, respectively.

After removing the unreliable data points, we smooth the elution profiles of the constituent peptides of each of the proteins in Sample A and Sample B by using the Savitzky-Golay filter with third-order polynomial smoothing and a moving window of maximum width nine, on both the *light* and *heavy* signals separately. As an illustrative case, the observed and filtered elution profiles corresponding to peptide *DFELEETDEEK* of protein *YPL240C* are shown in Figure 4.4. The ratio of

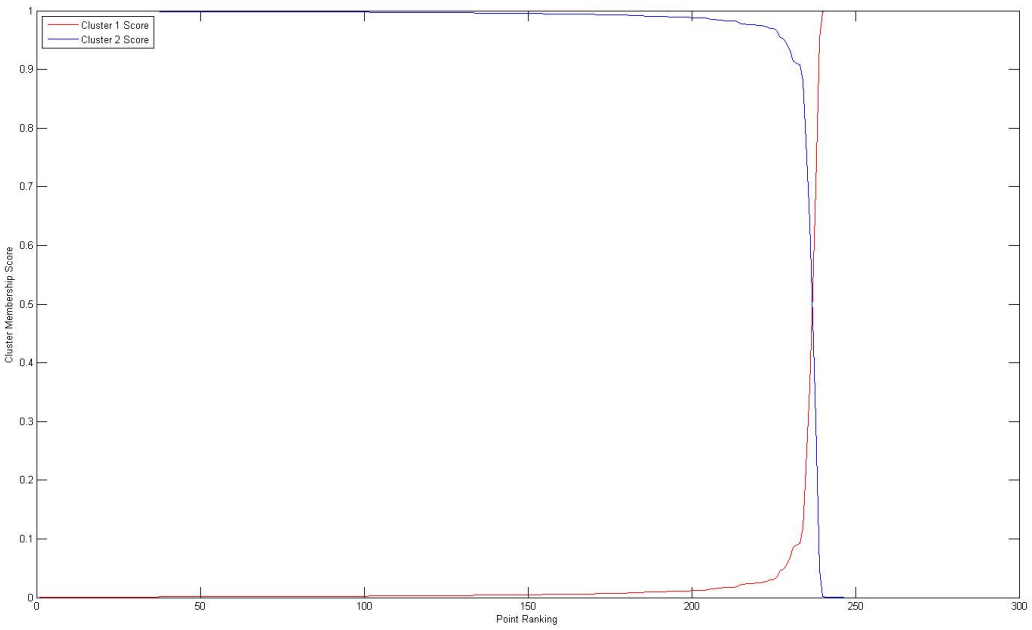


Figure 4.3: Cluster membership scores. Membership scores clearly separate the data into two classes, with relatively few data points having nearly equal scores.

the AUCs of the filtered profiles is 1.015. The matching ratio for *DFELEETDEEK* obtained from Sample B is 1.044. The two ratios are fairly consistent in both magnitude and direction. On the other hand, the ratio of the *light* and *heavy* signals of the highest observed peak is 1.108 in Sample A, and 0.893 in Sample B. In this case, the two estimates reverse direction, going from an estimate of up-regulation in Sample A to an estimate indicating down-regulation in Sample B. In general, the estimates obtained using AUCs of the filtered profiles are far more robust compared to the highest peak ratio based estimates. This behavior is expected since we set up our filtering algorithm to capture the region of the elution profile surrounding its highest peak, thereby ensuring that we work with only the most stable and reliable intensity signals.

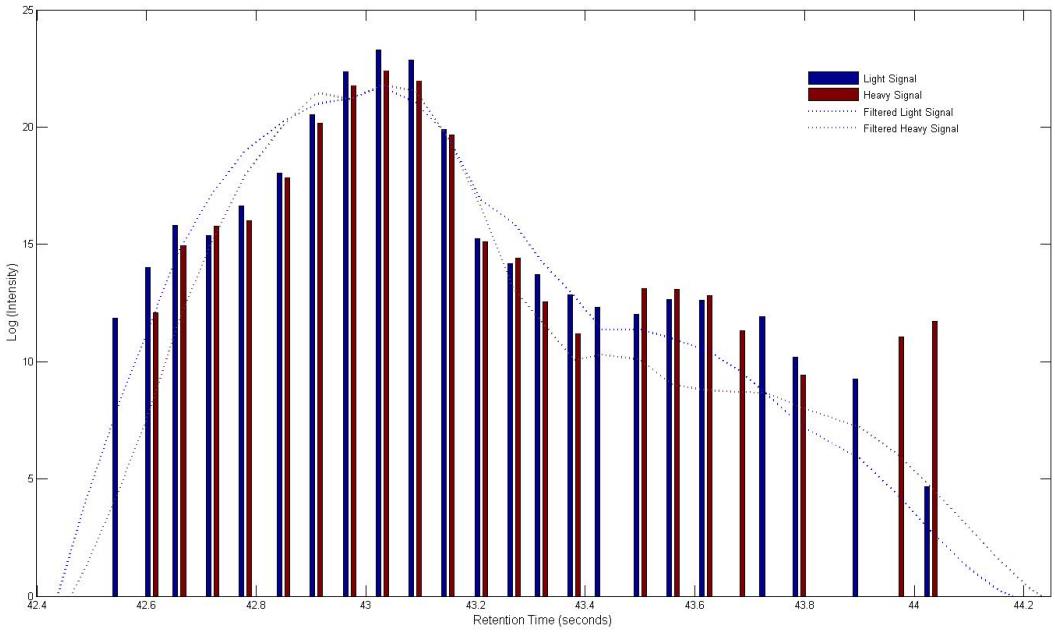


Figure 4.4: Savitzky-Golay filtered ion-current profile of *DFELEETDEEK*.

4.5.1 Estimation of Relative Protein Expression using a Bootstrap Partial Maximum Likelihood Estimator (BPMLE)

After applying the soothing filter to the data, we make use of the weighted double bootstrap methodology presented in Section 4.4.1 to draw $B_1 = 200$ first stage bootstrap samples and $B_2 = 50$ second stage samples, and estimate the bootstrap partial likelihood based estimator of overall relative protein expression ratio for each protein. We calculate a significance value for each expression ratio using the nested bootstrap p-value estimation method discussed in Section 4.4.3.1. Finally, we obtain a list of significant proteins by applying the local false discovery strategy on the estimated p-values using a Beta-Uniform mixture model.

The fitted Beta-Uniform mixture, posterior probability of non-differential expression, and the Quantile-Quantile plot assessing the fit of the mixture to the data are shown in Figure 4.5.

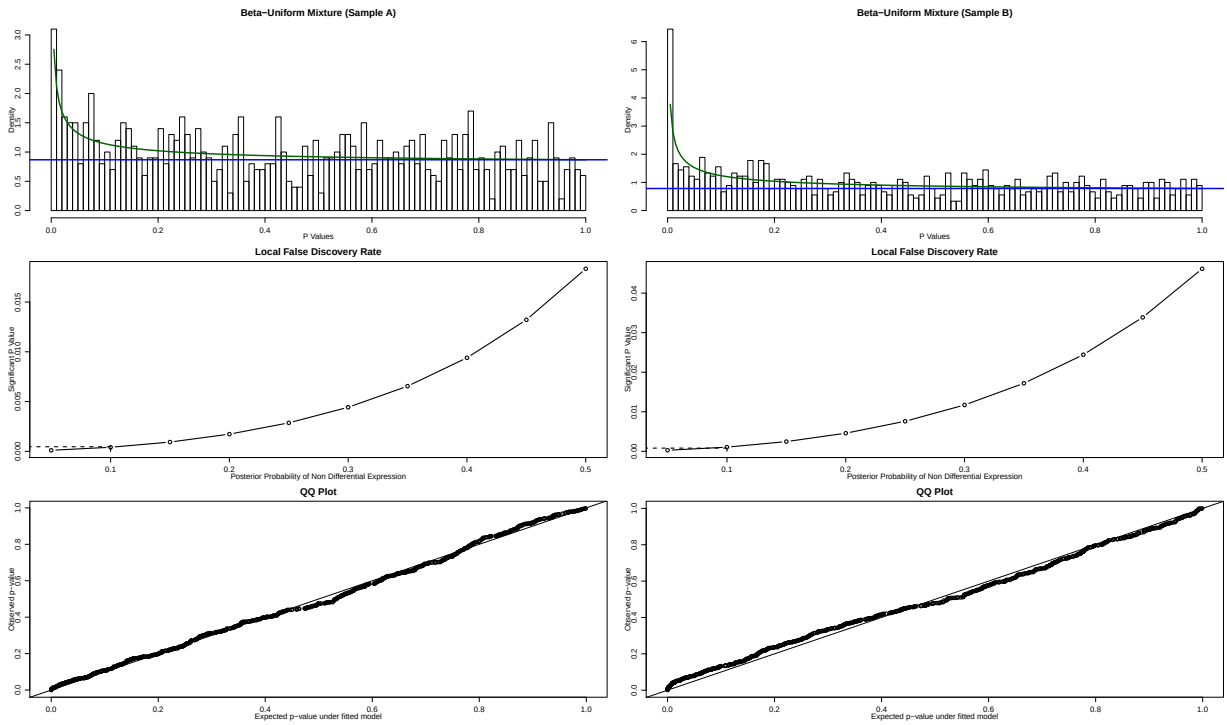


Figure 4.5: A Beta-Uniform mixture for bootstrap partial likelihood based p-values. The QQ-plots show the fit of the to the estimated Beta-Uniform mixture; the dotted horizontal line traces the p-value corresponding to a local false discovery rate cut-off of 0.1.

Parameter estimates of the mixture model, the number of significant proteins identified, and the nested bootstrap p-value corresponding to a local false discovery rate cut-off of 0.1 are shown in Table 4.1. The BPMLE - local fdr based significance analysis identified 7 proteins from Sample A and 5 proteins from Sample B as being significantly up or down regulated. Of the 7 proteins identified from Sample A, 3 are known contaminants. Of the 5 proteins identified from Sample B, only 1 is a known contaminant.

	Sample A (m = 614)	**Sample B (m = 588)**
Estimation Method	*Bootstrap Partial Maximum Likelihood*	
$\hat{a}$	0.490	0.404
$\hat{b}$	1.020	0.979
$\hat{\pi}_0$	0.933	0.952
p-value cutoff (locfdr < 0.1)	0.00061	0.00098
Number Significant	7	5

Table 4.1: Bootstrap partial likelihood based assessment of significant differential expression

4.5.2 Estimation of Relative Protein Expression using a Model-based Bootstrap

We start by separately fitting the reference regression model (4.18) using the Tukey bisquare M-estimator to all available peptide level data in Sample A and Sample B. Scatter plots of bootstrap replications corresponding to all possible two-way pairings of regression coefficients are shown in Figure 4.6 and Figure 4.7. The over-laid concentration ellipses correspond to 50, 90, and 99-percent levels and are drawn using a robust estimate of the covariance matrix of the coefficients. A concentration ellipse with a nearly horizontal major axis indicates that the two regression coefficients are *unaffected* by each other. Of note, are the inverse relationships observed between the coefficients for peptide length and charge, and peptide length and Xcorr, in predicting the distance of a given peptide's relative expression ratio from the overall mean

estimate. In general, we observe the same patterns of interaction between two-way pairings of coefficients in Sample A and sample B.

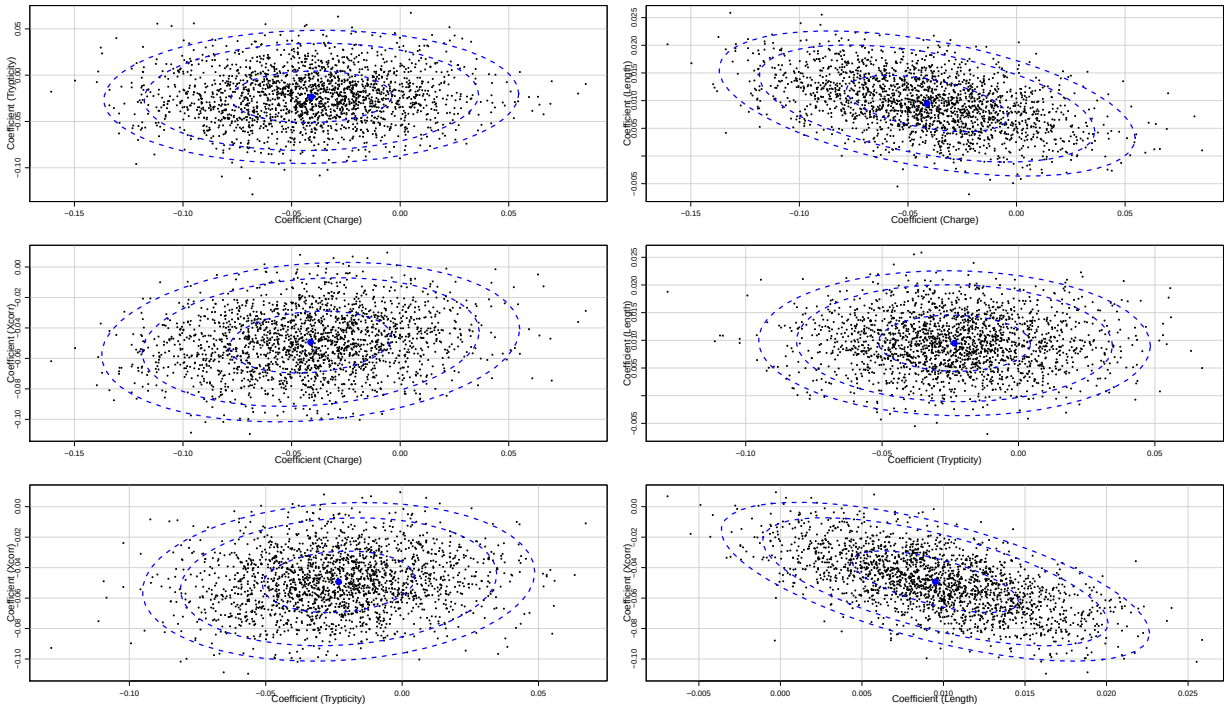


Figure 4.6: Scatterplot of bootstrap replications of the different combinations of regression coefficients from the Huber-Tukey bisquare regression for Sample A data. The concentration ellipses are drawn at the 50, 90, and 99-percent levels using a robust estimate of the covariance matrix of the coefficients.

Next, we fit the regression model (4.10) to each of the proteins in Sample A and Sample B. The results of this regression are shown for a selected yeast protein and a contaminant protein from Sample B: *YAL005C*, and *keratin_2.a*, respectively. The scatter plots of regression coefficients for these two proteins (Figures 4.8, 4.9) illustrate the consistent behavior of a protein that is homogenous with a majority of the proteins in the sample, and the deviant behavior a protein which is not.

The bootstrap statistics for the two fitted reference models and proteins *YAL005C*, *keratin_2.a* are given in Table 4.2. The estimated intercept term in each regression model represents the overall relative expression ratio. For the null yeast protein *YAL005C* and the contaminant *keratin_2.a*, the overall estimated relative protein

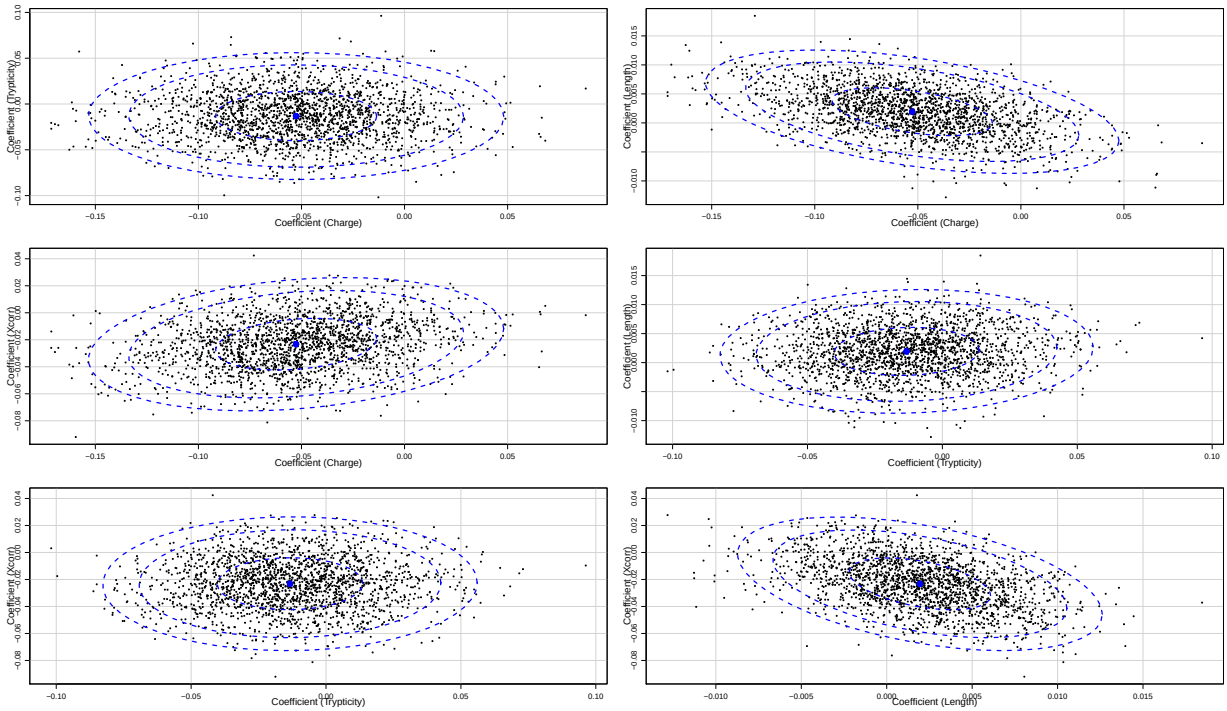


Figure 4.7: Scatterplot of bootstrap replications of the different combinations of regression coefficients from the Huber-Tukey bisquare regression for Sample B data.

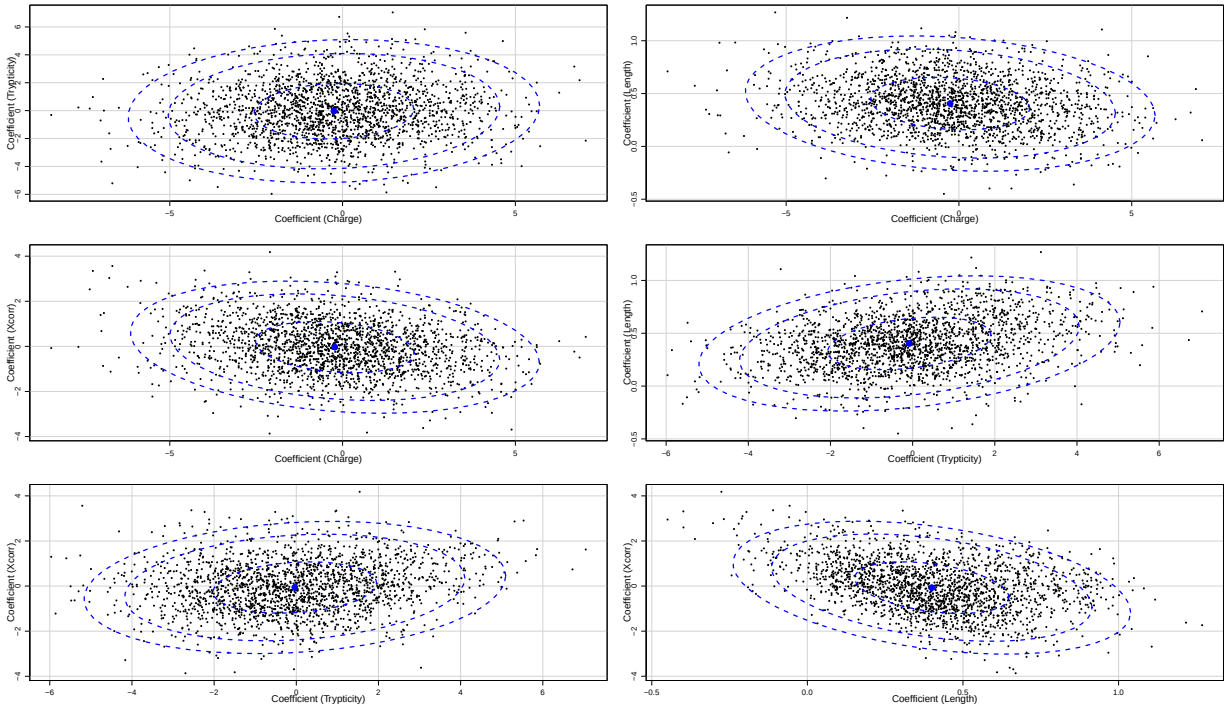


Figure 4.8: Behavior of Huber regression coefficients for *YAL005C*.The behavior of this non-differentially expressed protein is quite similar to the reference of its parent sample, Sample B, shown in Figure 4.7..

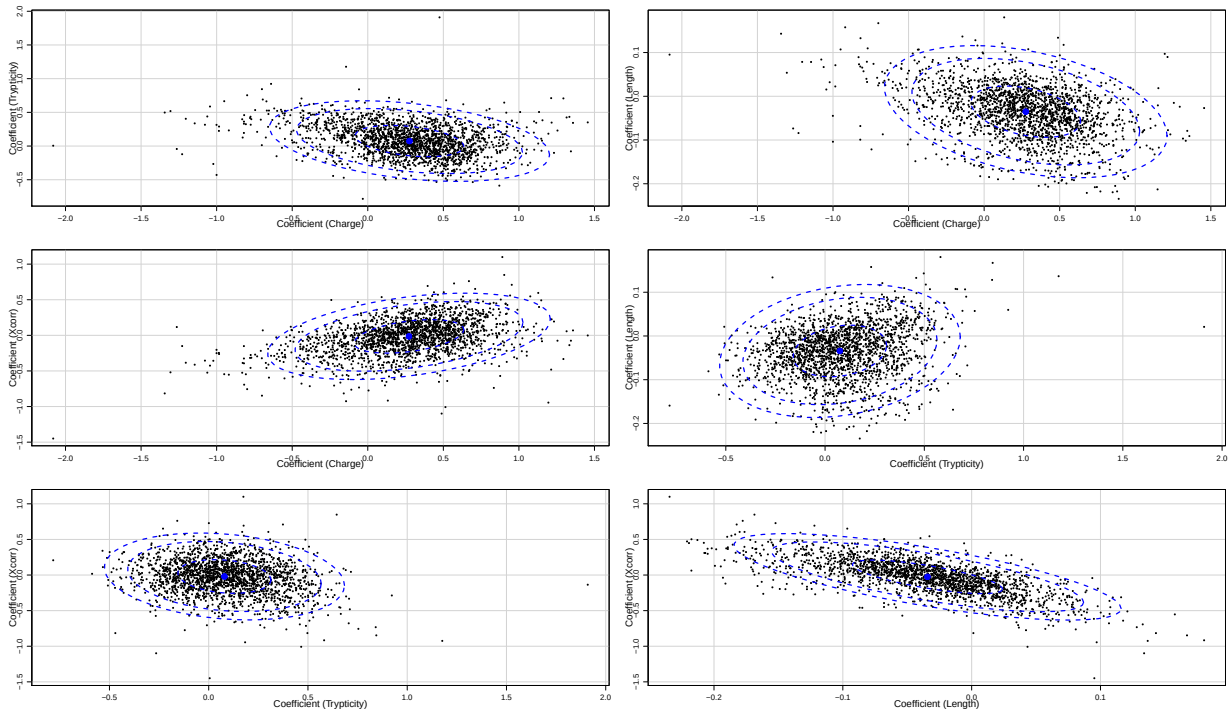


Figure 4.9: Behavior of Huber regression coefficients for *keratin_2.a*. The behavior of the known contaminant *keratin_2.a* is markedly different from the reference of its parent sample, Sample B (see Figure 4.7)

expression ratios are $2^{(-0.0613)} = 0.958$, and $2^{(3.5967)} = 12.098$, respectively. For both proteins, the *Charge*, *Trypticity*, and *Xcorr* coefficients are negative. This indicates that the relative expression ratio of a peptide with a higher charge state, better trypsin digestion, and a more reliable SEQUEST identification, will tend to be closer to the overall relative expression ratio of its parent protein. The peptide length coefficient seems to indicate that smaller peptides are preferable for obtaining a more reliable protein level estimate.

	Bootstrap Statistics (R = 2000)			
	Original	**Bias**	**Std. Error**	**95% C.I.**
Sample A – Reference (N = 6074)				
Intercept	0.0937	0.0003	0.0962	(-0.0952, 0.2818)
Charge	-0.0415	0.0001	0.0361	(-0.1124, 0.0292)
Trypticity	-0.0234	0.0002	0.0273	(-0.0772, 0.0300)
Length	0.0096	-0.0001	0.0049	(0.0001, 0.0194)
Xcorr	-0.0495	0.0002	0.0195	(-0.0880, -0.0114)
Sample B – Reference (N = 3790)				
Intercept	0.0596	-0.0010	0.1009	(-0.1373, 0.2584)
Charge	-0.0542	0.0013	0.0384	(-0.1307, 0.0197)
Trypticity	-0.0125	-0.0006	0.0262	(-0.0634, 0.0394)
Length	0.0020	-0.0001	0.0041	(-0.0058, 0.0101)
Xcorr	-0.0233	< 0.0001	0.0186	(-0.0597, 0.0131)
YAL005C (N = 85)	$B_1 = 200$			
Intercept	-0.0613	0.0683	0.0833	(-0.2249, 0.1023)
Charge	-0.0898	-0.0329	0.0367	(-0.2404, -0.0608)
Trypticity	-0.0317	0.0543	0.0224	(-0.0419, -0.0095)
Length	0.0364	< 0.0000	0.0294	(0.0103, 0.1531)
Xcorr	-0.0026	-0.0316	0.0242	(-0.0501, 0.0449)
Keratin_2.a (N = 82)				
Intercept	3.5967	-0.7224	6.6806	(-8.775, 17.413)
Charge	-0.0746	-0.1794	2.2630	(-4.3308, 4.5402)
Trypticity	-0.2250	0.1984	1.9424	(-4.2304, 3.3836)
Length	0.3711	0.0350	0.2434	(-0.1069, 0.8491)
Xcorr	-0.1449	0.1186	1.1181	(-2.4549, 1.9278)

Table 4.2: Bootstrap statistics for Huber regression coefficients. R = number of bootstrap replications used for the reference models; B_1 is the number of first stage samples drawn for each protein; 95% C.I. are based on the 2.5^{th} to 97.5^{th} percentiles of the distribution of the bootstrap replicates.

After carrying out the Huber bootstrap regression on each protein, we again make use of the double bootstrap sampling methodology presented in Section 4.4.2.1 to draw $B_1 = 200$ first stage re-constructed bootstrap samples and $B_2 = 50$ second stage samples, and estimate the bootstrap partial likelihood estimator of overall relative protein expression ratio for all proteins. Associated p-values are obtained by the bootstrap p-value estimation method presented in Section 4.4.3.1. Finally, we carry

out the local false discovery rate estimation using a Beta-Uniform mixture model on the estimated p-values. The fitted Beta-Uniform mixture, posterior probability of non-differential expression, and the Quantile-Quantile plot assessing the fit of the mixture to the data are shown in Figure 4.10.

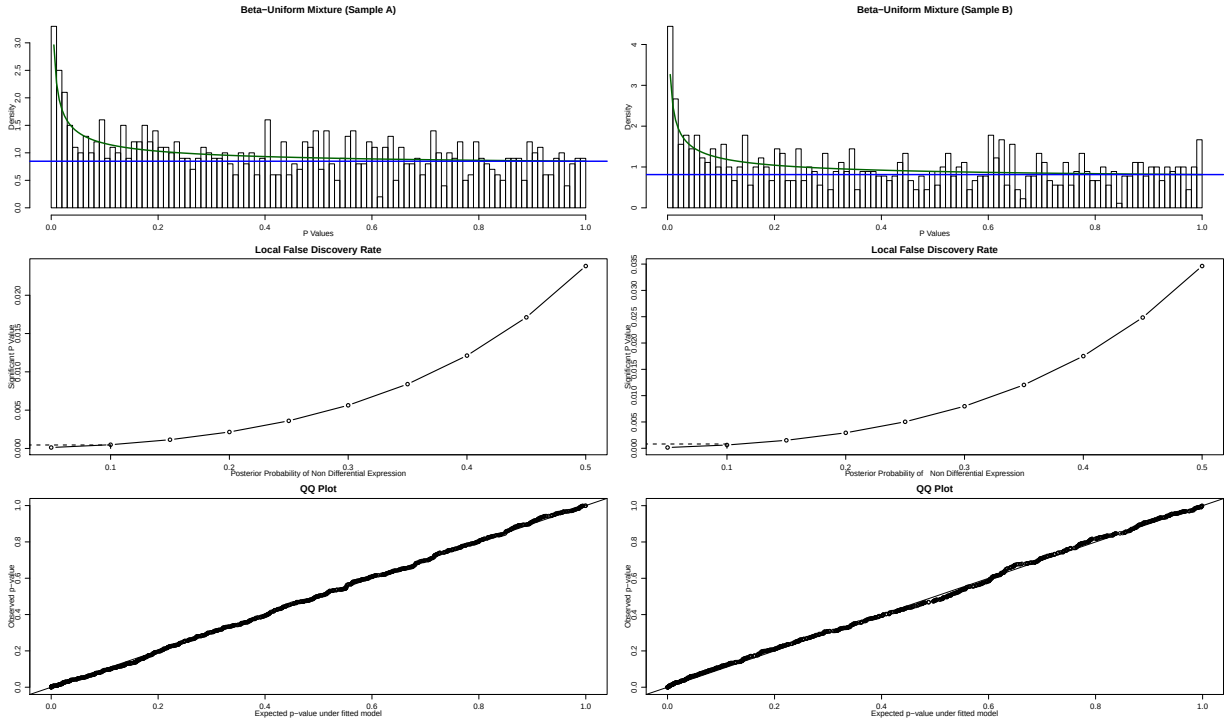


Figure 4.10: A Beta-Uniform mixture for Huber regression based bootstrap p-values. The QQ-plots show the fit of the to the estimated Beta-Uniform mixture; the dotted horizontal line traces the p-value corresponding to a local false discovery rate cut-off of 0.1.

Parameter estimates and the number of significant proteins identified are given in Table 4.3. The data analysis identified 15 proteins from Sample A and 12 proteins from Sample B as being significantly up or down regulated. Seven out of fifteen, and six out of the twelve of the identified proteins from Sample A and Sample B are known contaminants.

	Sample A (m = 614)	**Sample B (m = 588)**
Estimation Method	*Robust Regression based Bootstrap*	
$\hat{a}$	0.276	0.237
$\hat{b}$	1.119	1.513
$\hat{\pi}_0$	0.931	0.948
p-value cutoff (locfdr < 0.1)	0.0023	0.0028
Number Significant	15	12

Table 4.3: Bootstrap regression based assessment of significant differential expression.

4.6 Discussion

The validity and robustness of the results of proteomics data analyses are affected by the inherent variabilities in LC/MS-MS based proteomics strategies. While there have been some studies that looked into the processes and interactions between processes responsible for both biological and technical variability, there is a dearth of research in to the actual impact of this variability on peptide and protein expression estimates. Some of this variability can be reduced through careful quality control at each stage of a proteomics experiment. At the data analysis stage, we propose two additional steps that can be used to further reduce variability: removal of 'unreliable' data points through a bivariate mixture model based cluster analysis; and quantifying a peptide level relative expression ratio based on the ratio of the area under the filtered ion current profiles. These two steps together improve the robustness of the peptide level estimates that are then used to estimate a protein level overall relative expression ratio.

We propose two resampling based methods for estimating the relative protein expression from peptide level expression estimates. The first method is an extension of the *bootstrap partial likelihood* approach. The standard *bootstrap partial likelihood* estimates the likelihood for a parameter of interest θ based on the sampling density $p\left(\hat{\theta} \mid \theta\right)$, which is estimated directly from the data using a nested bootstrap computation. We propose to improve the efficiency of this standard construction by introducing a weighted resampling mechanism, at each level of nesting. The second method combines the *bootstrap partial likelihood* approach with a robust regression

of the error in estimating the overall relative protein expression ratio on covariate information available at the peptide level. We implement robust regression through the use of the Tukey bi-square M-estimator. For each peptide, considered covariates are: charge state, length of the amino acid sequence, level of trypsin digestion, and the peptide cross correlation score generated by the SEQUEST algorithm. We assign a p-value to each protein based on a strategy that makes use of the same nested bootstrap samples that are used to derive its *bootstrap partial likelihood* estimate of relative expression. We also propose to identify the set of significant proteins by locally controlling the false discovery rate using a Beta-Uniform mixture model.

With our proposed methods, our primary interest is in estimating an 'error - adjusted' expression level for each protein, through weighted resampling of all available peptide level data for that protein. From a different point of view, this process can be thought of as a missing value imputation problem since we are essentially imputing missing peptide values based on information gathered across all gel fractions in which that peptide is fully observed.

We demonstrate the use of these methods on two control data sets derived from the yeast proteome. The fact that all proteins in these data sets are mixed in a 1:1 ratio, allows us the ability to gauge the efficacy of the proposed methods, since the expected number of significant proteins in either data set (other than contaminants) is zero. Our analysis demonstrated that both resampling based approaches cut down on the number of false positives to a remarkable degree. This is not surprising since both resampling methods, over the long run, *average out* the highest and lowest peptide level estimates for each protein which are typically responsible for generating false positives.

A limitation of our methods is that they can be computer resource intensive, when using standard statistical software such as R or SAS and/or when run on a 32-bit system. We used the x86 Open64 C++ compiler on a 64 bit operating system to run

our programs. On this platform, average program run time from start of identifying and removal of less reliable data points to fitting of the Beta-Uniform mixture model, is about 40 minutes. Average run time on a 32-bit system running R is about 380 minutes. Another limitation is the limited number of covariates considered in the Huber regressions. The four covariates we use are merely the ones that are currently available to us, and may not be the best covariates that are predictive of peptide level estimation error in estimating protein expression.

In summary, our proposed resampling approaches provide an appealing alternative to traditional parametric approaches. Our methods do not require any assumptions about the distribution of the peptide level data. In fact, the only assumption we make is that a statistical model that accounts for the underlying error processes governing much of the variability in proteomics data, either through imposing weights on the peptide level data or through modeling the actual error as a function of peptide level covariate information, can reduce the number of false positives and false negatives observed from the data. This assumption seems reasonable given the small number of false positives found from each of the data sets.

Chapter 5

Estimating Relative Protein Expression Levels from Incomplete Data

5.1 Introduction

Current statistical analyses of proteomics data do not adequately consider the issue of dealing with non-expressed or undetected observations. For any given protein, we typically only observe a subset of the peptides that are theoretically predicted for that protein. The peptides that are not observed are *missing* in the sense that an observation that carries useful information towards estimating our outcome of interest, i.e., the true relative expression ratio of the protein, is not observed due to reasons outside of the experimenter's control.

To recap, in relative protein quantification methods such as SILAC, a relative expression estimate for a protein is typically constructed based on the calculated ratios between the *light* and *heavy* signals corresponding to peptides that are indicative of that protein. Typically, we observe four patterns of data for each peptide: (a) both the

light and the matching *heavy* signal is quantified; (b) only the *light* signal is quantified; (c) only the *heavy* signal is quantified; and (d) both signals are not quantified. Note here the distinction we are making between *quantifiability* and *detectability* of peptide signals. The set of quantifiable peptide signals is a sub set of the detected signals. For example, signals that are detected but are below a threshold set by the pre-processing algorithms may not be quantified, or a detected peptide signal may be designated as missing because it was incorrectly assigned to a different protein.

The issue of missing values may be addressed through the incorporation of suitably substituted values for the missing observations and/or by accounting for the mechanism responsible for generating missing values in data analyses. Typical approaches for handling missing values broadly fall under three categories: (1) complete-case only ; (2) imputation based; and (3) model based. For data from a SILAC experiment, a complete-case only analysis would use data corresponding to pattern (a) only. This is generally the most expedient means of dealing with missing values, since standard statistical analyses can be applied without modifications. The biggest disadvantage of this approach is the potential loss of information in discarding all incomplete cases. As Little and Rubin (2002)[73] have noted, this loss of information has two aspects: loss of precision; and bias that is introduced when the missing data mechanism is not MCAR (Missing Completely At Random). Imputation based procedures fill in the missing values before analyzing the resultant complete data by standard methods. However for valid inferences to be made, additional modifications are required that account for imputation uncertainty. The last category, and the one that we will pursue in this chapter, is the model based procedures for handling missing values. These procedures are characterized by defining a model for the observed data and the missingness mechanism and basing inferences on the likelihood under that model.

In this chapter, we look at the issue of missing values in proteomics experiments in several contexts. First in Section (5.2), we look at robustly estimating the true rel-

ative expression ratio of a protein based on incompletely observed peptide level data. Secondly, in Section (5.3) we look at the same estimation problem in situations where only one peptide is available to uniquely identify a protein. In both situations, we do not assume that our data is MCAR. This alone signifies a major advancement since most existing methods are either based on only the complete-cases or use imputation strategies that are only valid when the data are truly MCAR.

5.1.1 Setup of the data

Suppose in a proteomics experiment, we are interested in quantifying the relative expression of observed proteins, where the i^{th} protein's relative expression is estimated using the relative expression of p_i constituent peptides. This subset of p_i peptides would be determined separately for each protein based on the expert judgement of the experimenter as to which set of peptides are considered proteotypic and/or needed to make a high precision identification of the protein.

More formally, let n_i denote the number of times protein i is quantified in the experiment. For example, if data for protein i are observed in g_i gel fractions and r_{g_i} scans within each gel fraction, then $n_i = g_i \times r_{g_i}$. Hereafter, we will refer to each of these n_i instances as a separate *case* of that protein. Now let, $\mathbf{y}_k = (y_1, \ldots, y_{p_i})$ be a $(1 \times p_i)$ vector of values of the continuous variables $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{p_i})$, representing the p_i peptide level relative expression ratios for protein i. Typically, $\mathbf{y}_k$; $k = 1, \ldots, n_i$, is not observed fully due to the presence of data patterns (b), (c), and (d) as described in Section (5.1). We can represent this missingness by constructing a missing-data indicator matrix, $M = (m_{kj})$, such that $m_{kj} = 1$ if y_{kj} is missing and $m_{kj} = 0$ if y_{kj} is available; $k = 1, \ldots, n_i, j = 1, \ldots, p_i$. In addition, we define $\mathbf{Y}_k = \mathbf{Y}_k^{obs} \bigcup \mathbf{Y}_k^{mis}$ and $\mathbf{Y} = \left\{\mathbf{Y}^{obs} \bigcup \mathbf{Y}^{mis}\right\}$, where $\mathbf{Y}^{obs} = \left\{\mathbf{Y}_k^{obs} : k = 1, \ldots, n_i\right\}$ denotes the set of all available peptide level estimates for protein i across the n_i cases, and $\mathbf{Y}^{mis} = \left\{\mathbf{Y}_k^{mis} : k = 1, \ldots, n_i\right\}$ denotes the missing estimates.

5.1.2 Types of Missing Data Patterns and Mechanisms

The missing-data indicator matrix, $\mathbf{M}$, makes it easier to identify patterns in the missing data. Some of the more well known missing data patterns include *univariate non-response* where missingness is confined to a single variable, *monotone* where the variables can be arranged such that $y_{k,j+1}, \ldots, y_{k,p_i}$ are missing for cases where y_{kj} is missing, for all $j = 1, \ldots, p_i - 1$, and *general* where there is no discernible pattern to the missingness in the data.

Another important aspect to consider with missing data is the mechanism that lead to the missingness. In particular, it is important to know whether or not the reasons for missingness in variables is dependent upon the underlying values of the variables in the data set.

If $\mathbf{Y}$ is observed completely, the data setup described in Section (5.1.1) falls within the likelihood-based methods that assume a model for the distribution $f(\mathbf{Y}, \boldsymbol{\lambda})$ with unknown parameters $\boldsymbol{\lambda}$. If the n_i cases behave independently, then we can write:

$$f(\mathbf{Y}, \boldsymbol{\lambda}) = \prod_{k=1}^{n_i} f(\mathbf{Y}_k, \boldsymbol{\lambda}),$$

and the full likelihood of the unknown parameters given the data is

$$L(\boldsymbol{\lambda} \mid \mathbf{Y}) = c \prod_{k=1}^{n_i} f(\mathbf{Y}_k \mid \boldsymbol{\lambda}) \tag{5.1}$$

where c is an arbitrary factor that does not depend on $\boldsymbol{\lambda}$.

When the data are incomplete, our interest is in estimating $\boldsymbol{\lambda}$ based on the incomplete data set $\left(\mathbf{M}, \mathbf{Y}^{mis}\right) = \left\{\left(\mathbf{M}_k, \mathbf{Y}_k^{mis}\right) : k = 1, \ldots, n_i\right\}$. Under independence, the joint distribution of the incomplete data can be written as $f(\mathbf{M}, \mathbf{Y} \mid \boldsymbol{\theta}) = \prod_{k=1}^{n_i} f(\mathbf{M}_k, \mathbf{Y}_k \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\phi})$: $\boldsymbol{\lambda}$ characterizes the model for data $\mathbf{Y}$ and $\boldsymbol{\phi}$ characterizes the model for the missing data indicators $\mathbf{M}$. Likelihood inferences in

the presence of missing data are based on the *observed-data likelihood*, which is obtained by integrating out the missing data component out of the density of $(\mathbf{M}_k, \mathbf{Y}_k)$:

$$L\left(\boldsymbol{\theta} \mid \mathbf{M}, \mathbf{Y}^{obs}\right) = c \prod_{k=1}^{n_i} \int f\left(\mathbf{M}_k, \mathbf{Y}_k \mid \boldsymbol{\theta}\right) d\, \mathbf{Y}_k^{mis} \tag{5.2}$$

Similar to the complete data model, large-sample maximum likelihood inferences can be made under the normal approximation

$$\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right),$$

where $\boldsymbol{\Sigma}_\theta$ is now given by

$$\{-\partial^2 \log L\ \left(\boldsymbol{\theta} \mid \mathbf{M}, \mathbf{Y}^{obs}\right) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'\}^{-1}.$$

Rubin (1976)[106]; Little and Rubin (2002)[73], discuss a number of complications that arise with respect to the missing data likelihood in (5.2) compared to the complete data likelihood in (5.1). Clearly, specification of the joint distribution in (5.2) requires knowledge of the mechanism leading to missing values. If the data are *missing at random* (MAR), in the sense that missingness only depends on the data through the observed values $\mathbf{Y}^{obs}$, and the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$ are distinct, then the missingness mechanism is called *ignorable*.

Ignorability implies

$$f\left(\mathbf{M}_k \mid \mathbf{Y}_k, \boldsymbol{\phi}\right) = f\left(\mathbf{M}_k, \mid \mathbf{Y}_k^{obs}, \boldsymbol{\phi}\right) \ \ \forall \ \ \mathbf{Y}_k^{mis}. \tag{5.3}$$

On the other hand *non-ignorable* missingness mechanisms require the specification of a missing data mechanism and the maximization of the full likelihood $L\left(\boldsymbol{\theta} \mid \mathbf{M}, \mathbf{Y}^{obs}\right)$. In particular, in the context of *pattern-mixture* models (Glynn,

Laird, and Rubin (1986)[50]; Little (1993)[72]), the joint distribution of $\mathbf{M}_k$ and $\mathbf{Y}_k$ is factored into the marginal distribution of $\mathbf{M}_k$ and the conditional distribution of $\mathbf{Y}_k$ given $\mathbf{M}_k$ as

$$f\left(\mathbf{M}_k, \mathbf{Y}_k \mid \boldsymbol{\nu}, \boldsymbol{\delta}\right) = f_M\left(\mathbf{M}_k, \mid \boldsymbol{\delta}\right) f_{Y|M}\left(\mathbf{Y}_k \mid \mathbf{M}_k, \boldsymbol{\nu}\right). \tag{5.4}$$

In (5.37), the first factor models the incidence of the different patterns and the second factor characterizes the the distribution of $\mathbf{Y}_k$ in the strata defined by different patterns of missing data, $\mathbf{M}_k$, and $\boldsymbol{\nu}$, $\boldsymbol{\delta}$ are assumed distinct.

5.2 Estimating Relative Protein Expression Levels from Incomplete Peptide Data

Most proteomics analyses to date have relied on a complete data model. This approach leads to valid inferences only if missing data are missing completely at random (MCAR). I.e., a peptide measurement being missing is not dependent on other observed or unobserved peptide measurements. This is an unrealistic assumption since in reality the missing set of peptide measurements cannot be considered a random subset of the hypothetically complete data. While it is quite difficult to articulate the exact mechanism that is responsible for missingness in the peptide data, there is no reason to limit ourselves to the restrictive MCAR assumption. In fact we believe that a stronger argument can be made in favor of proteomics analyses that are based on the less restrictive MAR assumption. For the data setup described in Section (5.1.1), the MAR assumption implies that the probability of an observation being missing does not depend on the missing values $\mathbf{Y}^{mis}$ of $\mathbf{Y}$ but can depend on the observed values $\mathbf{Y}^{obs}$ in the data set. This assumption is justifiable given the recent findings of Mallick et. al [77] with regards to *proteotypic* peptides.

5.2.1 A Test of MCAR for Multivariate Data

Statistically validating the MAR assumption for multivariate data with a *general* pattern of missingness is quite complicated and beyond the scope of our work envisaged under missing data methods in proteomics. However we believe that a formal test of the MCAR assumption can still be very useful. For example, such a test provides guidance as to what type of standard errors are preferable. In particular, standard errors for the parameter estimates based on the expected information matrix are not valid unless the data are MCAR. On the other hand standard errors based on the observed information matrix are valid only when the data are MAR. In addition, we believe that establishing a test of the MCAR assumption as a standard prerequisite will further the field of proteomics data analyses, at least as far as convincing researchers that the peptide data are indeed not missing completely at random.

A simple test of the MCAR assumption can be based on the two sample t tests for differences in means. For each variable $\mathbf{Y}_k$ with missing values, this can be achieved by splitting the data into cases with that variable observed and cases with that variable missing. The means of observed values of the remaining variables in the two groups are then compared using two sample t tests. If the tests show that the groups are significantly different with respect to their means, then the MCAR assumption clearly does not hold. While this approach is simple and intuitive, there are significant issues associated with multiple testing and with the complex correlation structure of the t statistics themselves.

In our data analyses, we make use of an alternative test of the MCAR assumption that was proposed by Little (1988)[71].

5.2.2 A likelihood Ratio Based Test of MCAR

For protein i, let the population mean vector and covariance matrix of $\mathbf{y}_k$ be denoted by $\boldsymbol{\mu}_{(1\times p_i)}$ and $\boldsymbol{\Sigma}_{(p_i\times p_i)}$. Now let $\mathbf{m}_k = (m_{k1}, \ldots, m_{kp_i})\,; k = 1, \ldots, n_i$ be the vector

of missing data indicators for case k; $\mathbf{B}$ = number of distinct missing data patterns $\mathbf{m}_k$ in the data set; $S_b \equiv$ set of cases corresponding to pattern $b, b = 1, \ldots, B$; $r_b =$ number of cases in S_b, $\sum r_b = n_i$; $v_b =$ number of observed variables for the cases in S_b; and $\mathbf{D}_b$ be the $(p_i \times v_b)$ matrix indicating which variables are observed for pattern b. Each column of $\mathbf{D}_b$ represents a variables that was observed for pattern b and consists of $p_i - 1$ zeros and one 1 corresponding to the variable identified.

Now let, $\bar{\mathbf{y}}_b^{obs} \equiv r_b^{-1} \sum_{h \in S_b} \mathbf{y}_h^{obs}$ be the $(1 \times v_b)$ vector of means of observed variables for pattern b, and $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ be the maximum likelihood estimates of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$. Then a likelihood ratio based test for testing the MCAR assumption can be based on the test statistic

$$d^2 = \sum_{b=1}^{B} r_b \left(\bar{\mathbf{y}}_b^{obs} - \hat{\boldsymbol{\mu}}_b^{obs}\right) \tilde{\boldsymbol{\Sigma}}_b^{obs^{-1}} \left(\bar{\mathbf{y}}_b^{obs} - \hat{\boldsymbol{\mu}}_b^{obs}\right)^T, \tag{5.5}$$

where, $\hat{\boldsymbol{\mu}}_b^{obs} = \hat{\boldsymbol{\mu}}\mathbf{D}_b$, $\tilde{\boldsymbol{\Sigma}}_b^{obs} = \mathbf{D}_b^T \tilde{\boldsymbol{\Sigma}} \mathbf{D}_b$, and $\tilde{\boldsymbol{\Sigma}} = n_i \hat{\boldsymbol{\Sigma}}/(n_i - 1)$ is the maximum likelihood estimate of $\boldsymbol{\Sigma}$ with a correction for degrees of freedom.

Under the null hypothesis that the data are MCAR, and assuming that the distribution of $\mathbf{y}_k$ has finite fourth moments, d^2 is asymptotically chi-squared distributed with degrees of freedom, $\sum_{b=1}^{B} v_b - p_i$. For large d^2, we reject the null hypothesis in favor of an alternative model in which the means of the observed variables are allowed to vary across the missingness patterns.

5.2.3 A Multivariate General-MAR Model for Incomplete Peptide Data

In this section, we present a statistical framework for estimating the mean and covariance matrix of $\mathbf{Y}$ under a multivariate model that assumes an *ignorable* mechanism and has a *general* pattern of missingness. Our interest here is only in obtaining valid estimates of peptide means, variances and covariances, and not in making inferences

about these estimated parameters.

Under the *ignorability* assumption, the likelihood takes the form

$$L_{ign}\left(\boldsymbol{\lambda} \mid \mathbf{Y}^{obs}\right) = c\prod_{k=1}^{n_i}\int f\left(\mathbf{Y}_k \mid \boldsymbol{\lambda}\right) d\,\mathbf{Y}_k^{mis} = c\prod_{k=1}^{n_i} f\left(\mathbf{Y}_k^{obs} \mid \boldsymbol{\lambda}\right) \tag{5.6}$$

Note that (5.6) does not require a model for $\mathbf{M}$ to be specified and does not have many of the identifiability issues associated with model (5.2) (Little and Rubin, 2002)[73].

We now make the additional assumptions that the $\mathbf{y}_k$ are independent and

$$\left(\mathbf{y}_k \mid \boldsymbol{\theta}, w_k\right) \sim N_{p_i}\left(\boldsymbol{\mu}, \boldsymbol{\Psi}/w_k\right)\ ;\ k = 1,\ \ldots,\ n_i, \tag{5.7}$$

where the w_k are unobserved i.i.d. positive scalar random variables with known density $h(w_k)$, and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Psi})$. Letting each case have it's own dispersion factor w_k enables us to assign a weighting mechanism that downweights cases that are outside of the normative range of values. Note that this model setup we requires $n_i \geq 2$. For our data, this condition is satisfied since even if only one peptide is available, the elution profile of that peptide will contain multiple peak pairs. So, in practice, n_i is always greater than one.

Under this setup, likelihood inferences about $\boldsymbol{\theta}$ can be based on the marginal distribution of $\mathbf{Y}^{obs}$, without modeling the missing data mechanism (Little (1988b))[70]. Maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ can be found by applying the EM algorithm, treating $\mathbf{Y}^{mis}$ and $\boldsymbol{w} = (w_1,\ \ldots,\ w_{n_i})$ as missing data.

When the data are complete and $\boldsymbol{w}$ is observed, the maximum likelihood estimates of $(\boldsymbol{\mu}, \boldsymbol{\Psi})$ can be constructed using the complete-data sufficient statistics

$$s_0 = \sum_{k=1}^{n_i} w_k,\ \ s_y = \sum_{k=1}^{n_i} w_k\,\mathbf{y}_k\ \ \text{and}\ \ s_{yy} = \sum_{k=1}^{n_i} w_k\,\mathbf{y}_k\,\mathbf{y}_k^T.$$

These estimates are given by

$$\hat{\boldsymbol{\mu}} = s_y/s_0 = \frac{\sum_{k=1}^{n_i} w_k \mathbf{y}_k}{\sum_{k=1}^{n_i} w_k}, \tag{5.8}$$

$$\hat{\boldsymbol{\Psi}} = \left(s_{yy} - s_y\, s_y^T/s_0\right)/n_i = \frac{1}{n_i}\sum_{k=1}^{n_i} w_k \left(\mathbf{y}_k - \hat{\boldsymbol{\mu}}\right)\left(\mathbf{y}_k - \hat{\boldsymbol{\mu}}\right)^T. \tag{5.9}$$

Note that $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Psi}}$ are in fact the classical weighted least squares estimators of $(\boldsymbol{\mu}, \boldsymbol{\Psi})$.

When $\mathbf{Y}^{mis}$ and $\boldsymbol{w}$ are unobserved, the maximum likelihood estimates can be obtained by applying the EM theory for exponential families. The $(t+1)^{st}$ iteration of the EM algorithm proceeds as follows:

E-step : Estimate s_0, s_y and s_{yy} by their conditional expectations, given $\mathbf{Y}^{obs}$ and current estimates $\boldsymbol{\theta}^{(t)} = \left(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Psi}^{(t)}\right)$ of $\boldsymbol{\theta}$.

(1) $E\left(s_0 \mid \boldsymbol{\theta}^{(t)}, \mathbf{Y}^{obs}\right) = E\left(\sum_{k=1}^{n_i} w_k \mid \boldsymbol{\theta}^{(t)}, \mathbf{Y}^{obs}\right) = \sum_{k=1}^{n_i} E\left(w_k \mid \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}\right) = \sum_{k=1}^{n_i} w_i^{(t)}$

(2) The j^{th} component of $E\left(s_y \mid \boldsymbol{\theta}^{(t)}, \mathbf{Y}^{obs}\right)$ is

$$\begin{aligned} E\left(\sum_{k=1}^{n_i} w_k\, y_{kj} \mid \boldsymbol{\theta}^{(t)}, \mathbf{Y}^{obs}\right) &= \sum_{k=1}^{n_i} E\left[w_k\, E\left(y_{kj} \mid \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}, w_k\right) \mid \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}\right] \\ &= \sum_{k=1}^{n_i} w_k^{(t)}\, E\left(y_{kj} \mid \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}\right) \\ &= \sum_{k=1}^{n_i} w_k^{(t)} \hat{y}_{kj}^{(t)}, \text{ and} \end{aligned}$$

(3) The $(j,l)^{th}$ element of $E\left(s_{yy} \mid \boldsymbol{\theta}^{(t)}, \mathbf{Y}^{obs}\right)$ is

$$E\left(\sum_{k=1}^{n_k} w_k\, y_{kj}\, y_{kl} \mid \boldsymbol{\theta}^{(t)}, \mathbf{Y}^{obs}\right) = \sum_{k=1}^{n_k} E\left[w_k\, E\left(y_{kj}y_{kl} \mid \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}, w_k\right) | \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}\right]$$
$$= \sum_{k=1}^{n_k} E\left[w_k\, \left\{\hat{y}_{kj}^{(t)}\hat{y}_{kl}^{(t)} + \mathrm{cov}\left(y_{kj}, y_{kl} \mid \mathbf{y}_k^{obs}, w_k\right)\right\} \mid \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}\right]$$
$$= \sum_{k=1}^{n_k} \left(w_k^{(t)}\, \hat{y}_{kj}^{(t)}\, \hat{y}_{kl}^{(t)} + \psi_{jl.obs,k}^{(t)}\right)$$

where $w_k^{(t)} = E\left(w_k \mid \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}\right)$ and $\psi_{jl.obs,i}^{(t)}$ is zero if y_{ij} or y_{il} are observed, and w_i times the residual covariance of y_{ij} and y_{il} given $\mathbf{y}_i^{obs}$, if both y_{ij} and y_{il} are missing.

M-step : Compute new estimates $\boldsymbol{\theta}^{(t+1)} = \left(\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Psi}^{(t+1)}\right)$ from (1) and (2), with s_0, s_y, and s_{yy} replaced by their estimates from the E-step.

5.2.4 A Robust Alternative to the Multivariate Normal Estimation

The estimates discussed in the previous section are affected to a certain degree by violations of the multivariate normal assumption, and hence tend to be sensitive to outliers. Robust alternatives to the multivariate normal model have been suggested previously by Little and Smith (1987)[74], and Devlin *et al.*, (1981)[33]. These methods were primarily *ad hoc* modifications to complete-data procedures, and do not possess the optimal asymptotic properties associated with maximum likelihood. In this section, we present an alternative representation of the model given in (5.6) due to Little (1988b)[70] that is based on a multivariate t model.

The multivariate t extension can be achieved with only minor modification to the EM algorithm described in the previous section. More specifically, the extension is achieved by declaring a model for w_k, and attributing a particular form for the

weights, $w_k^{(t)}$. We have chosen a model for w_k that lets the marginal distribution of $\mathbf{y}_k$ to be the t distribution with v degrees of freedom.

Suppose that w_k is such that $w_k v$ is distributed as χ_v^2. Then marginally

$$\mathbf{y}_k \sim t_{p_i}\left(\boldsymbol{\mu}, \boldsymbol{\Psi}, v\right),$$

with unknown degrees of freedom v. Under the t model, the weights $w_k^{(t)}$ are given by

$$w_k^{(t)} = E\left(w_k \mid \boldsymbol{\theta}^{(t)}, \mathbf{y}_k^{obs}\right) = (v^{(t)} + p_{ik})/(v^{(t)} + d_k^{2(t)}), \tag{5.10}$$

where $p_{ik}\,;\, 0 \leq p_{ik} \leq p_i$ is the number of peptides observed at the k^{th} time protein i was quantified, and

$$d_k^{2(t)} = \left(\mathbf{y}_k^{obs} - \boldsymbol{\mu}_{obs,k}^{(t)}\right)^T \boldsymbol{\Psi}_{obs,k}^{(t)^{-1}} \left(\mathbf{y}_k^{obs} - \boldsymbol{\mu}_{obs,k}^{(t)}\right),$$

is the squared Mahalanobis distance from the mean of the observed peptides, evaluated at the current estimates of the parameters, $\left(\boldsymbol{\mu}_{obs,k}^{(t)}, \boldsymbol{\Psi}_{obs,k}^{(t)}\right)$. The EM algorithm for the t model requires an additional computational step in the M-step for computing the degrees of freedom, $v^{(t+1)}$. This is a one-dimensional maximization problem and can be achieved by maximizing the observed loglikelihood $\ell\left(\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Psi}^{(t+1)}, v \mid \mathbf{Y}_{obs}\right)$ with respect to v using a grid search or a Newton-Raphson step.

Note that the above model downweights cases with large squared distances. Therefore the t model can be expected to yield maximum likelihood estimates of the mean, $\boldsymbol{\mu}$ and covariance matrix, $v\boldsymbol{\Psi}/(v-2)$ of $\mathbf{y}_k$ that are more resistant to non-normality and outliers in the observed data. However, if the data are in fact normal, the t model is less efficient. Fortunately, this loss in efficiency has been shown to be relatively small (Little (1988b)[70].

5.2.5 Estimating the True Relative Protein Expression Ratio

Once we have valid estimates for the means and variances of the constituent peptides, they can be used to construct many different summary estimates for the true relative protein expression ratio. For example, we can obtain many summary estimates by using the following property of linear transformations of the multivariate t distribution.

If $\mathbf{Y} \sim t_{p_i}(\boldsymbol{\mu}, \boldsymbol{\Psi}; v)$ and $\mathbf{X} = \mathbf{a}\mathbf{Y}$, where $\mathbf{a}$ is a $(1 \times p_i)$ non-empty vector, then

$$\mathbf{X} \sim t_1\left(\mathbf{a}\boldsymbol{\mu}, \mathbf{a}\boldsymbol{\Psi}\mathbf{a}^T; v\right).$$

Choosing $\mathbf{a} = (1/p_i, \ldots, 1/p_i)$ yields the common mean $\tilde{\boldsymbol{\mu}}$ of the p_i variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_{p_i}$, each representing a peptide level estimate for protein i as $\tilde{\boldsymbol{\mu}}_i = \mathbf{a}\hat{\boldsymbol{\mu}}$ with associated variance $\sigma^2_{\tilde{\boldsymbol{\mu}}_i} = \mathbf{a}\hat{\boldsymbol{\Psi}}\mathbf{a}^T$.

5.3 A Missing Data Model for Single Peptide Proteins

In Section (5.2.3), we presented a multivariate general-MAR model that yields valid estimates about the true peptide level protein abundances when we can assume that data are MAR and have a general pattern of missingness. In that model we declared as missing any peptide for which we were not able to quantify both the *light* and corresponding *heavy* signals. Furthermore, partially observed peptides, i.e., peptides for which the *light* signal is quantified but the matching *heavy* signal is not, or vice versa, were also considered as missing.

However, when $p_i = 1$, i.e., when protein i is only identified using a single peptide, the multivariate model loses its appeal since it reduces to an univariate complete-case only analysis. In this section we develop a more efficient data analysis framework for

dealing with these single peptide proteins that makes use of not just the information available from the complete-cases but also at least some information recovered from the partially observed cases.

5.3.1 Setup of the data

Let, $\mathbf{Y}^* = (L, H)$ be the bivariate normal random variable representing the observed *light* and *heavy* signals corresponding to the single peptide that is available for identifying a given protein. Suppose that we have n such bivariate observations or *cases* across the experiment. Now let $\mathbf{M} = \{m_i\}$, where $m_i = (m_{il}, m_{ih})$ denotes the missing data pattern for case i with four possible values (0,0), (0,1), (1,0), and (1,1). For simplicity, we adopt an the integer labeling scheme for these four patterns; $r_i = r(m_i)$ such that $r(0,0) = 0$, $r(1,0) = 1$, $r(0,1) = 2$, and $r(1,1) = 3$. Under this scheme $r_i = 0$ and $r_i = 1$ correspond to the complete-case pattern and the completely-missing pattern, respectively. We also define the notation: $P(r_i = r) = \pi_r$, $\boldsymbol{\pi} = \{\pi_r\}$; $n = \sum_{r=0}^{3} n_r$, where n_r = number of cases falling under each pattern; and $S_r \equiv$ set of sample cases following pattern r.

Let $\boldsymbol{\theta} = (\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}^* = (\mu_l, \mu_h)$, $\boldsymbol{\Sigma} = (\sigma_{jk})$; $j, k = l, h$ be the mean vector and unrestricted covariance matrix of $\mathbf{Y}^*$. For each missingness pattern r, let $\phi^{(r)} = \left(\mu_l^{(r)}, \mu_h^{(r)}, \sigma_{ll}^{(r)}, \sigma_{hh}^{(r)}, \sigma_{lh}^{(r)}\right)$ denote the means, variances, and covariance of L and H. We also define the notation: $\phi_l^{(r)}$ and $\phi_{h.l}^{(r)}$ to represent the parameters of the marginal distribution of L and the conditional distribution of H given L; and $\phi_h^{(r)}$, $\phi_{l.h}^{(r)}$ to represent the parameters of the marginal distribution of H and the conditional distribution of L given H.

When all four patterns are present in the data, i.e., when the data are *saturated* with respect to missingness patterns, the dimension of $\boldsymbol{\phi} = \{\phi^{(r)}\}$ is $5 \times 4 = 20$. When n_r, is small, as is the case with our single peptide proteins, estimates for a majority of these 20 parameters become unreliable. However, an effective compromise can be

found if we restrict the analysis to only the cases corresponding to pattern $r = 0$ or $r = 1$. The choice of these two patterns is motivated by the fact that upwards of 90% of our data consist of these two patterns. In essence, this restricted analysis is an improvement on a complete-case only analysis in all situations, and is an effective compromise when the two patterns $r = 0, 1$ account for a majority of the data.

More formally, let N be the number of cases corresponding to the two patterns, $r = 0, 1$. Of the N cases, N_0 are complete on L and H, say $\mathbf{Y}_0^* = \{(l_i, h_i), i = 1, \ldots, N_0\}$, and $N_1 = N - N_0$ are complete $\{l_i, i = N_0 + 1, \ldots, N\}$ on L only. Note that this data setup corresponds to a *monotone* pattern of missingness as described in Section 5.1.2. Now let $\mathcal{M} = \{m_i\}$ be the missing-data indicator variable for this reduced data set. For case i, $m_i = 0$ if h_i is observed, and $m_i = 1$ if h_i is missing, $i = 1, \ldots, N$, Our ultimate interest is in estimating the true relative protein expression ratio, $R = \log_2(\mu_l/\mu_h) = \log_2 \mu_l - \log_2 \mu_h$. Possible estimates for this quantity include the (a) *complete-case* (CC) estimate $\log_2\bar{l} - \log_2\bar{h}$, where $\bar{l}$, $\bar{h}$ are the sample means of L, H from the n_0 complete cases; and (b) *available-case* (AC) estimate $\log_2\hat{\mu}_l - \log_2\bar{h}$, where $\hat{\mu}_l = \sum_{i=1}^{N} l_i/N$ is the sample mean of L from all the cases.

5.3.2 A Test of MCAR for Bivariate Normal Monotone-Missing Data

As already mentioned in Section (5.2.1), a test of the MCAR assumption can be based on a two sample t test that allows for unequal variances. In fact the t test is a simple and quite effective approach especially for our bivariate normal data with only two patterns of missingness. However, for completeness, we note here that the test statistic discussed in Section (5.2.2) can also be adapted to yield a small sample version of d^2 that is applicable to a *monotone* pattern of missingness. Little (1988)[71]

showed that for bivariate data the test statistic reduces to

$$d^2 = (N-1)\mathcal{F}/(N-2+\mathcal{F}), \tag{5.11}$$

where, $\mathcal{F} \sim F_{1,N-2}$, is the F statistic from the ANOVA of L on the missingness pattern r, under the null hypothesis of MCAR and assuming that values of L are normal. This test is in fact equivalent to the classical two sample t test that assumes equal variances.

5.3.3 A Bivariate Normal Monotone-MAR Model

If we assume that the missingness in the data is MAR, i.e., missingness of H can depend on L, but conditional on L it does not depend on H, then we can consider the maximum likelihood estimates under independence and the *ignorable* normal model parameterized by $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$\mathbf{Y} \equiv (L, H) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5.12}$$

$$\boldsymbol{\theta} = (\mu_l, \mu_h, \sigma_{ll}, \sigma_{hh}, \sigma_{lh}) \tag{5.13}$$

The likelihood equations for this model do not have an obvious solution. Anderson (1957)[5], and Little and Rubin (2002)[73] derived the maximum likelihood estimates by factoring the joint distribution of L_i and H_i into the marginal distribution of L_i and the conditional distribution of H_i given L_i as:

$$f(l_i, h_i \mid \boldsymbol{\theta}) = f(l_i \mid \mu_l, \sigma_{ll})\, f(h_i \mid l_i, \beta_{h0.l}, \beta_{hl.l}, \sigma_{hh.l}) \tag{5.14}$$

$$\equiv N(\mu_l, \sigma_{ll}) \times N(\beta_{h0.l} + \beta_{hl.l} l_i, \sigma_{hh.l}) \tag{5.15}$$

where,

$$\beta_{hl.l} = \sigma_{lh}/\sigma_{ll} \tag{5.16}$$

$$\beta_{h0.l} = \mu_h - \sigma_{lh}\beta_{hl.l}\,\mu_l \tag{5.17}$$

$$\sigma_{hh.l} = \sigma_{hh} - \sigma_{lh}^2/\sigma_{ll} \tag{5.18}$$

Under this setup, we get

$$\mu_h = \beta_{h0.l} + \beta_{hl.l}\mu_l \tag{5.19}$$

$$\sigma_{lh} = \beta_{hl.l}\sigma_{ll} \tag{5.20}$$

$$\sigma_{hh} = \sigma_{hh.l} + \beta_{hl.l}^2\sigma_{ll} \tag{5.21}$$

and the new transformed parameter space $\mathbf{\Phi} = (\mu_l, \sigma_{ll}, \beta_{h0.l}, \beta_{hl.l}, \sigma_{hh.l})$. We can now factor the density of the observed data as:

$$f\left(\mathbf{Y}^o \mid \mathbf{\Phi}\right) = \prod_{i=1}^{N} f\left(l_i \mid \mu_l, \sigma_{ll}\right) \prod_{i=1}^{N_0} f\left(h_i \mid l_i, \beta_{h0.l}, \beta_{hl.l}, \sigma_{hh.l}\right). \tag{5.22}$$

Since given the data, we assume that knowledge of (μ_l, σ_{ll}) does not provide any information about $(\beta_{h0.l}, \beta_{hl.l}, \sigma_{hh.l})$, maximum likelihood estimates of $\mathbf{\Phi}$ can be obtained by independently maximizing the likelihood of the marginal and conditional components. If $\hat{\mathbf{\Phi}}$ are the resultant MLEs, then since $\mathbf{\Phi}$ is a one-to-one function of

$\boldsymbol{\theta}$, we recover the MLEs of $\boldsymbol{\theta}$ as

$$\hat{\mu}_l = \sum_{i=1}^{N} l_i/N \tag{5.23}$$

$$\hat{\sigma}_{ll} = \sum_{i=1}^{N} \left(l_i - \hat{\mu}_l\right)^2 /N \tag{5.24}$$

$$\hat{\mu}_h = \bar{h} + \hat{\beta}_{hl.l} \left(\hat{\mu}_l - \bar{l}\right) \tag{5.25}$$

$$\hat{\sigma}_{lh} = \hat{\beta}_{hl.l} \hat{\sigma}_{ll} \tag{5.26}$$

$$\hat{\sigma}_{hh} = s_{hh.l} + \hat{\beta}^2_{hl.l} \left(\hat{\sigma}_{ll} - s_{ll}\right), \tag{5.27}$$

where $\bar{l}$, $\bar{h}$ are now the sample means of L, H from the N_0 complete cases, $s_{ll} = \sum_{i=1}^{N_0} \left(l_i - \bar{l}\right)^2 /N_0$, $s_{hh} = \sum_{i=1}^{N_0} \left(h_i - \bar{h}\right)^2 /N_0$, $s_{lh} = \sum_{i=1}^{N_0} \left(l_i - \bar{l}\right)\left(h_i - \bar{h}\right) /N_0$, $s_{hh.l} = s_{hh} - s^2_{lh}/s_{ll}$, and $\hat{\beta}_{hl.l} = s_{lh}/s_{ll}$ is the regression coefficient of L_i from a regression of H_i on L_i, based only on the N_0 complete cases.

Note that the maximum likelihood estimate (5.25) of μ_h is also the average of observed and imputed values when the missing values of h_i are imputed with predictions from the regression of H_i on L_i, computed only from the N_0 complete cases. Finally, we get

$$\hat{R} = \log_2 \hat{\mu}_l - \log_2 \hat{\mu}_h.$$

5.3.4 Small sample inference

Under the MAR assumption, large-sample maximum likelihood inferences can be made under the normal approximation

$$\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right),$$

where $\mathbf{\Sigma}_\theta$ is now given by

$$\{-\partial^2 \log L\left(\boldsymbol{\theta} \mid \mathbf{M}, \mathbf{Y}^{obs}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\}^{-1}.$$

However, these estimates are not ideal when N is relatively small, which is typically the case with proteomics data. We therefore use a Bayesian approach to estimate the standard error of $\hat{R}$ using its posterior distribution. This Bayesian approach was motivated by results that were first derived by Lindley (1965)[69].

If we assume μ_l, σ_{ll}, $\beta_{h0.l}$, $\beta_{hl.l}$, and $\sigma_{hh.l}$ are *a priori* independent with reference prior

$$f\left(\mu_l, \sigma_{ll}, \beta_{h0.l}, \beta_{hl.l}, \sigma_{hh.l}\right) \propto \sigma_{ll}^{-a}\sigma_{hh.l}^{-c}, \tag{5.28}$$

then Lindley (1965) showed that the following results hold:

(1) $N\,\hat{\sigma}_{ll}/\sigma_{ll} \sim \chi^2_{N+2a-3}$

(2) Posterior distribution of μ_l given σ_{ll} is $N\left(\hat{\mu}_l\,,\,\sigma_{ll}/N\right)$

(3) $N_0\,\hat{\sigma}_{hh.l}/\sigma_{hh.l} \sim \chi^2_{N_0+2c-4}$

(4) Posterior distribution of $\beta_{hl.l}$ given $\sigma_{hh.l}$ is $N\left(\hat{\beta}_{hl.l}\,,\,\sigma_{hh.l}/\left(N_0 s_{ll}\right)\right)$

(5) Posterior distribution of $\beta_{h0.l}$ given $(\beta_{hl.l}, \sigma_{hh.l})$ is $N\left(\bar{h} - \beta_{hl.l}\,\bar{l}\,,\,\sigma_{hh.l}/N_0\right)$, and

(6) (μ_l, σ_{ll}) and $(\beta_{h0.l}, \beta_{hl.l}, \sigma_{hh.l})$ are *a posteriori* independent.

The posterior distribution of any function $g\left(\mathbf{\Phi}\right)$ of $\mathbf{\Phi}$ can then be simulated by creating draws $g_d,\ d = 1, \ldots, D$, where $g_d = g\left(\mathbf{\Phi}^{(d)}\right)$ and $\mathbf{\Phi}^{(d)} = \left(\mu_l^{(d)}, \sigma_{ll}^{(d)}, \beta_{h0.l}^{(d)}, \beta_{hl.l}^{(d)}, \sigma_{hh.l}^{(d)}\right)$. If we take $g\left(\mathbf{\Phi}\right) = R \equiv \log_2 \mu_l - \log_2 \mu_h = \log_2 \mu_l - \log_2\left(\beta_{h0.l} + \beta_{hl.l}\mu_l\right)$, then we can draw from the posterior distribution of R using the scheme outlined below:

(1) Draw independently x_{1t}^2 and x_{2t}^2 from chi-squared distributions with $N + 2a - 3$ and $N_0 + 2c - 4$ degrees of freedom, respectively.

(2) Draw independently three standard normal deviates z_{1t}, z_{2t}, and z_{3t}.

(3) For the d^{th} draw, compute $\mathbf{\Phi}^{(d)}$, where

$$\sigma_{ll}^{(d)} = N\,\hat{\sigma}_{ll}/\chi_{1t}^2 \tag{5.29}$$

$$\mu_l^{(d)} = \hat{\mu}_l + z_{1t}\sqrt{\left(\sigma_{ll}^{(d)}/N\right)} \tag{5.30}$$

$$\sigma_{hh.l}^{(d)} = N_0\,\hat{\sigma}_{hh.l}/\chi_{2t}^2 \tag{5.31}$$

$$\beta_{hl.l}^{(d)} = \hat{\beta}_{hl.l} + z_{2t}\sqrt{\sigma_{hh.l}^{(d)}/\left(N_0 s_{ll}\right)} \tag{5.32}$$

$$\beta_{h0.l}^{(d)} = \bar{h} - \beta_{hl.l}^{(d)}\,\bar{l} + z_{3t}\sqrt{\sigma_{hh.l}^{(d)}/N_0} \tag{5.33}$$

$$\mu_h^{(d)} = \bar{h} + \beta_{hl.l}^{(d)}\left(\mu_l^{(d)} - \bar{l}\right) \tag{5.34}$$

$$\sigma_{lh}^{(d)} = \beta_{hl.l}^{(d)}\sigma_{ll}^{(d)} \tag{5.35}$$

$$\sigma_{hh}^{(d)} = s_{hh.l} + \beta_{hl.l}^{2(d)}\left(\sigma_{ll}^{(d)} - s_{ll}\right) \tag{5.36}$$

(4) Compute $g_d = g\left(\mathbf{\Phi}^{(d)}\right) = R^{(d)} = \log_2 \mu_l^{(d)} - \log_2 \mu_h^{(d)}$.

5.4 Results

5.4.1 Estimating Relative Protein Expression from Incomplete Peptide Data

We illustrate our multivariate incomplete data methods on a number of selected proteins obtained from either Sample A or Sample B.

First, we randomly select three yeast proteins that have different levels of missingness, and apply the test described in Section 5.2.2 to test for violations of the MCAR assumption. The set of constituent peptides for each of the proteins is set at five; these five peptides being the ones that have the highest empirical frequency of observation for the protein in this particular experiment. Note that a proteomics practitioner can use many other scientific criteria and expert judgement in choosing these set of peptides. Secondly, we select a single protein, *YGR192C* (again with data limited to five *proteotypic* peptides), that has no missing values and artificially create three incomplete versions of the protein by introducing 10%, 20%, and 30% missingness. Missing data are created using a random number generator that randomly selects peptide level data points for deletion. The four proteins under investigation and their five most *proteotypic* peptides are listed in Table 5.1. We then apply the estimation procedure described in Section 5.2.3 to the four data sets: the original data set with no missing values, and the three incomplete versions with different levels of missingness. For comparison, we also derive complete-case (CC) only estimates for each of the data sets using the MINQUE procedure presented in Section 3.2.3.

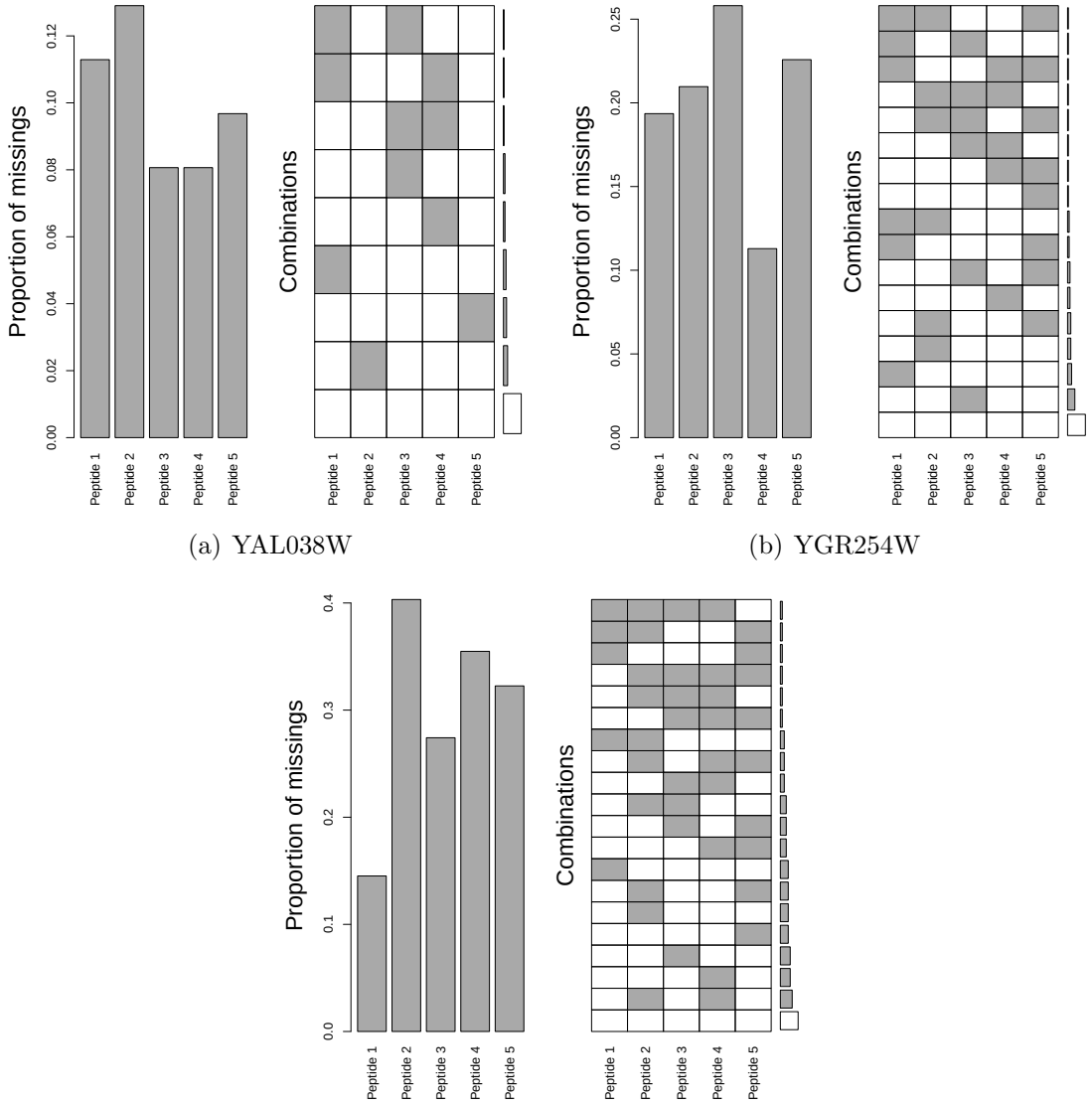
Peptide List	**Protein List**			
	YAL038W	**YGR254W**	**YHR174W**	**YGR192C**
Peptide 1	AIIVLSTSGTTPR	AADALLLK	AADALLLK	HIDAGAK
Peptide 2	GDLGIEIPAPEVLAVQK	AVDDFLISLDGTANK	AVYAGENFHHGDK	IVSNASCTTNCLAPLAK
Peptide 3	GVFPFVFEK	GNPTVEVELTTE	IEEELGDK	TASGNIIPSSTGAAK
Peptide 4	LTSLNVVAGSDLR	IGLDCASSEFFK	IGLDCASSEFFK	VPTVDVSVVDLTVK
Peptide 5	YRPNCPIILVTR	IGSEVYHNLK	PTVEVELTTEK	YDSTHGR

Table 5.1: Four yeast proteins and their *proteotypic* peptides.

The missingness patterns for each of the three proteins with missing values are graphically presented in Figure 5.4.1. The results of applying the likelihood ratio test described in Section 5.2.2 are shown in Table 5.2. The test provides more evidence against the MCAR assumption as the proportion of missing values in the data increases. When the proportion of missing values is greater than ten percent, the test

rejects the null hypothesis of equality of the peptide means across the B missingness patterns at the 0.1 level of significance.

The estimated peptide means and variances are presented in Table 5.3. The estimated overall relative expression ratio of protein *YGR192C* and it's variance corresponding to different levels of missingness are given in Table 5.4.



(a) YAL038W

(b) YGR254W

(c) YHR174W

Figure 5.1: Multivariate missingness patterns observed in three selected yeast proteins. Gray colored cells represent missing data values; gray colored bars represent the proportion of missing data in each peptide; horizontal rows represent missing data patterns; and the vertical bar to the right of each figure, indicates the relative frequency of each pattern.

Protein	Missing (%)	N-obs , N-miss	B	d^2	χ^2(df, p-value)
YAL038W	6.0	296 , 19	9	39.0306	(29 , 0.1011)
YGR254W	11.0	276 , 34	17	65.7350	(50 , 0.0670)
YHR174W	14.1	318 , 52	20	97.3933	(54 , 0.0003)

Table 5.2: Likelihood ratio test results for testing MCAR. B = number of missing data patterns; d^2 = value of the likelihood ratio test statistic.

Relative Peptide Expression Ratio Estimates

Missingness (%)	**Multivariate t Estimates ($\hat{\mu}$, $\hat{\sigma}$)**				
	Peptide 1	**Peptide 2**	**Peptide 3**	**Peptide 4**	**Peptide 5**
0	0.112, 0.706	0.121, 1.115	-0.126, 0.992	0.015, 0.593	-0.031, 1.306
10	0.132, 0.747	0.089, 1.078	-0.100, 1.031	0.086, 0.525	-0.206, 1.072
20	0.066, 0.717	0.014, 1.161	-0.316, 1.060	0.026, 0.733	-0.025, 1.238
30	0.124, 0.572	0.199, 1.478	-0.118, 0.937	-0.104, 0.887	-0.117, 1.566

Missingness (%)	**Complete-Case Only Estimates ($\hat{\mu}$, $\hat{\sigma}$)**				
	Peptide 1	**Peptide 2**	**Peptide 3**	**Peptide 4**	**Peptide 5**
0	-0.191, 1.007	-0.001, 0.827	-0.116, 0.729	-0.089, 1.028	0.021, 0.933
10	-0.111, 0.883	0.065, 0.850	0.227, 0.705	0.001, 0.732	0.029, 0.876
20	0.095, 1.204	0.101, 1.246	0.023, 1.037	-0.095, 0.933	-0.044, 0.891
30	-0.094, 0.781	-0.027, 0.848	-0.006, 1.076	-0.086, 0.971	-0.145, 1.135

Table 5.3: Relative Peptide Expression Ratio Estimates based on the MINQUE methodology and the robust multivariate t model.

	YGR192C $\left(\tilde{\boldsymbol{\mu}}, \sigma^2_{\tilde{\boldsymbol{\mu}}}\right)$	
Missingness (%)	Complete-Case	Multivariate t
0	-0.0756 , 0.1941	0.0182 , 0.1989
10	0.0423 , 0.1566	0.0001 , 0.1553
20	0.0160 , 0.1609	-0.0471 , 0.2011
30	-0.072 , 0.1714	-0.0034 , 0.2543

Table 5.4: Complete-case and robust multivariate t estimates of the relative expression ratio of *YGR192C*.

5.4.2 Estimating Relative Protein Expression from Single Peptide Data

Similar to the multivariate t analysis, we demonstrate the single peptide methods only on a few selected proteins. We illustrate the utility of the two sample t test as a means of testing the MCAR assumption in bivariate data with a monotone missingness pattern using three randomly selected single-peptide yeast proteins. Then we select a single protein, *YBL030C*, that has no missing values for its single detected peptide, *GFLPSVVGIVVYR*, and artificially created three incomplete data sets with 10%, 20%, and 30% missingness. The four proteins and the single peptide used to identify each of them are listed in Table 5.5. We then apply the estimation procedure discussed in Section 5.3.3 for bivariate monotone data to each of the four data sets separately.

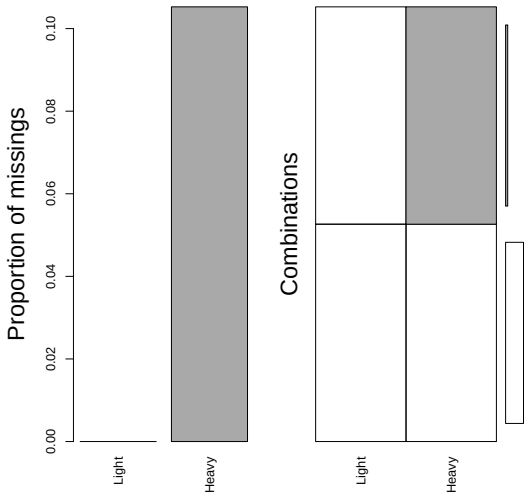
Protein	**Peptide**
YGR155W	LSGLVTLSELLR
YGR159C	GYGYVDFENK
YGR180C	IITEAVEIEK
YBL030C	GFLPSVVGIVVYR

Table 5.5: Four yeast proteins identified using a single peptide.

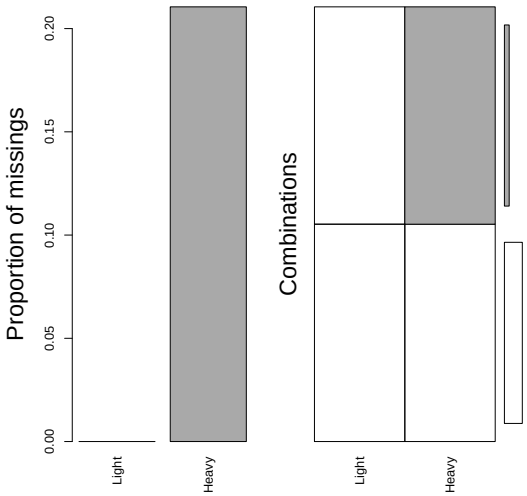
The missingness pattern for each of the three proteins with missing values are graphically presented in Figure 5.4.2. The results of applying the two sample t test to the three proteins (YGR155W, YGR159C, YGR180C) are shown in Table 5.6. The test provides some evidence against the MCAR assumption, when the number of missing values is relatively high for the *heavy* signal. The evidence is not definitive when there is only a few missing values.

The estimated means, variances, and covariances of the *light* and *heavy* signals for *YBL030C* corresponding to different levels of missingness are presented in Table 5.7. Both sample estimates of these quantities and ML estimates under our bivariate monotone-MAR model are given. The estimated overall relative expression ratio of

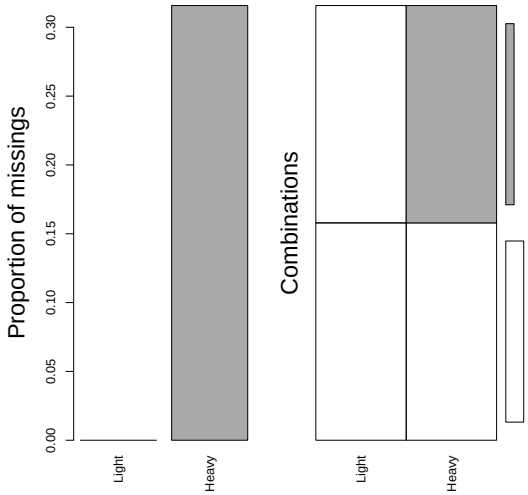
the protein, $\hat{R} = log_2(\hat{\mu}_l) - log_2(\hat{\mu}_h)$, is given in Table 5.8. At each level of missingness: 10%, 20%, and 30%, the Bivariate-Monotone-MAR (B-M-MAR) estimate of the relative protein expression ratio is closer to its 0% missingness value compared to its CC estimate. However, there is no clear pattern of divergence of the estimates from their 0% missingness values, as the proportion of missing values goes up.



(a) YGR155W



(b) YGR159C



(c) YGR180C

Figure 5.2: Bivariate monotone missingness patterns.

Protein	Missing (%)	N-obs , N-miss	t (value , d.f. , p-value)
YGR155W	8.8	31 , 3	t = -6.3858, d.f. = 8.564, p-value = 0.0002
YGR159C	13.2	33 , 5	t = 2.0964, d.f. = 3.474, p-value = 0.1270
YGR180C	23.4	36 , 11	t = 1.9195, d.f. = 15.893, p-value = 0.0731

Table 5.6: Two sample t test results for testing Bivariate - MCAR.

Missingness (%)	**Parameters**				
	μ_l	μ_h	σ_{ll}	σ_{hh}	σ_{lh}
	Sample Estimates				
0	14.0789	13.9977	3.6258	4.4199	2.8553
10		13.9703		4.6917	2.9599
20		13.5053		3.5266	1.9782
30		14.1204		4.3325	3.0578
	ML Estimates				
0	14.0789	13.9977	3.6258	4.4199	2.8553
10		14.0120		4.5941	2.8351
20		13.7189		3.7753	2.3602
30		14.1653		4.1998	2.8914

Table 5.7: Estimated means, variances, and covariances of the *light* and *heavy* signals for *YBL030C*.

Missingness (%)	**Estimated Relative Protein Expression Ratio, $\hat{R}$**	
	Complete-Case	**B-M-MAR**
0	0.0083	0.0083
10	0.0112	0.0069
20	0.0600	0.0373
30	-0.0104	-0.0088

Table 5.8: Estimated Relative Protein Expression Ratio, $\hat{R}$.

5.4.2.1 Small Sample Confidence Intervals

We calculate small sample 95% confidence intervals for the parameters, μ_h, σ_{hh}, σ_{lh}, and R, by obtaining the 2.5^{th} to 97.5^{th} percentiles of their Bayesian posterior distributions, each with 100,000 simulated values. The prior distribution of the parameters is set to the Jeffrey's prior for a factored density (Box and Tiao, 1973[22]), by setting $a = c = 1$ in the reference prior (5.28). Plots of the posterior draws and distribution for each parameter are given in Appendix 5.2.2. For comparison, we also calculate asymptotic 95% confidence intervals based on the inverse of the observed information matrix. The numerical results of the analysis are shown in Table 5.9.

	Parameter estimates and 95% Confidence Intervals		
	ML Estimate	**Asymptotic C.I.**	**Small Sample C.I.**
Missingness (%) = 10			
μ_h	14.0120	13.2746 , 14.7494	13.2722 , 14.7541
σ_{hh}	4.5941	2.0731 , 7.1151	2.9996 , 8.0229
σ_{lh}	2.8351	1.0007 , 4.6695	1.5509 , 5.1926
R	0.0069	-0.3307 , 0.3445 $^{\Delta}$	-0.0513 , 0.0657
Missingness (%) = 20			
μ_h	13.7189	13.0104 , 14.4274	13.0093 , 14.4347
σ_{hh}	3.7753	1.4512 , 6.0994	2.3306 , 6.9234
σ_{lh}	2.3602	0.6269 , 4.0935	0.9906 , 4.4529
R	0.0373	-0.3045 , 0.3791 $^{\Delta}$	-0.0685 , 0.05198
Missingness (%) = 30			
μ_h	14.1653	13.4087 , 14.9219	13.4065 , 14.9341
σ_{hh}	4.1998	1.6039 , 6.7957	2.6182 , 7.7610
σ_{lh}	2.8914	1.0184 , 4.7644	1.5783 , 5.2999
R	-0.0088	-0.3687 , 0.3511 $^{\Delta}$	-0.0269 , 0.1061

Table 5.9: Asymptotic and small sample confidence intervals. Δ indicates a C.I. based on a second order delta method approximation.

The asymptotic interval for μ_h is shorter than its small sample counterpart for all levels of missingness. The opposite is true for all other parameters. The small sample interval for R is noticeably shorter than it's asymptotic delta method approximation. Another advantage of the small sample intervals is that unlike the asymptotic intervals, they are not constrained to be symmetric. The small sample intervals are also adequately adjusted for simulation error since we draw a large number of posterior

draws for each parameter (i.e, 100,000).

5.5 Discussion

Standard statistical methodologies used in proteomics data analyses, do not consider the issue of dealing with non-expressed or undetected observations, beyond simple missing value imputation schemes such as mean or median imputation. In this work, we describe a model based framework for estimating the relative protein expression ratios from incompletely observed peptide level data, under the assumption that the data are missing at random (MAR). This type of data analyses is a significant improvement on complete-case only analyses, or multiple imputation schemes that rely on the overly restrictive missing completely at random (MCAR) assumption. In the MAR context, we propose both a multivariate t model for robustly estimating the true relative protein expression ratio when the data have a general pattern of missingness, and a bivariate normal model when the missingness pattern is monotonic. We also propose a Bayesian scheme for deriving small sample confidence intervals for parameters derived under the bivariate normal model.

The exact mechanisms responsible for generating missing values in proteomics data is not known. However there is no reason to assume that the data are missing completely at random. In our investigations, we propose to use a formal test of the MCAR assumption due to Little (1988)[71]. This test is limited to testing for differences in variable means across different missingness patterns in the data. However, we believe that any evidence the test can provide for or against the MCAR assumption is quite informative. For example, this information can be helpful in deciding which asymptotic inferences are applicable and valid since this decision depends on whether or not the data are MCAR.

The multivariate t model's primary utility is its ability to be more resistant to

outliers. We let each data vector representing the peptide level data observed for a protein have its own dispersion factor, thereby allowing the downweighting of cases that have a large squared Mahalanobis distance from their estimated peptide means. Our proposed model for the dispersion factors allows each data vector to be marginally distributed as t with an unknown number of degrees of freedom that is estimated from the data it self.

When a protein is identified using only one peptide, the multivariate t model loses its appeal since it reduces to an univariate complete-case only analysis. Therefore, we propose a bivariate normal model that works at the observed *light* and *heavy* signal level, assuming, without loss of generality, that only the *light* signal is fully observed for all cases. This assumption is reasonable since upwards of 90% of our data have at least one of the two signals recorded. Under this model, asymptotic confidence intervals based on the inverse of the observed information matrix are valid (under MAR). However, these estimates are not ideal since single-peptide proteins usually have small sample sizes. In our work, we propose an alternative Bayesian approach for constructing confidence intervals for the estimated model parameters.

We demonstrate our methods by artificially introducing different levels of missingness to a small number of selected proteins that have no missing values to begin with. Both the multivariate t model and the bivariate normal model, when applied to data with different levels of missingness, produce results that are generally consistent with estimates obtained from the original data with no missing values. In particular, the bivariate normal model estimate of the relative protein expression ratio is closer to its 0% missingness value compared to its CC estimate, for all considered levels of missingness. However, there is no clear pattern of divergence of the other parameter estimates from their values corresponding to 0% missingness, as the proportion of missing values increases. The small sample Bayesian intervals generally out perform their asymptotic counterparts except in the case of the interval for the estimated

mean of the variable with missingness, i.e., the *heavy* signal.

To summarize, in this work, we develop a data analysis framework for estimating relative protein expression ratios when data are allowed to be missing at random. We also propose methods for data analysis when only a single peptide is available to identify a protein. In this work, our intention is simply to demonstrate the applicability of these methods, and not arriving at definitive conclusions. Such conclusions will have to come from rigorous and extensive simulation studies, and is beyond the scope of our current research.

5.6 Future Work

The bivariate *ignorable* missing data method that we presented in Section 5.3.3 relied on the MAR assumption. That is, we allowed missingness of H to depend on the value of L, which is always observed, but not on the value of H, which is sometimes missing. While this is a much more reasonable assumption to make compared to the MCAR assumption, it is still worthwhile investigating other plausible missing-data models. For example, we could hypothesize that missingness could be related to not just L, but to some unknown extent on H. This scenario however clearly violates the MAR assumption, and needs to be studied under a Not Missing At Random (NMAR) or *non-ignorable* missing-data model. In our future research, we propose to investigate this problem within the context of fitting a Pattern Mixture Model - PMM (Glynn, Laird, and Rubin (1986)[50]; Little (1993)[72]). In this section, we present a detailed outline of this planned research.

5.6.1 A Pattern Mixture Model (PMM) for Single Peptide Proteins

PMMs explicitly model the effect of missing data mechanism by first identifying different patterns of missing data and then including parameters in a *non-ignorable* model that accounts for this effect.

Let us again make the assumptions that for our *monotone* data described in Section (5.3.1) that:

(a) $(L_i, H_i \mid m_i = r) = N_2\left(\boldsymbol{\mu}^{(r)}, \boldsymbol{\Sigma}^{(r)}\right)$ with parameters
$\phi^{(r)} = \left(\mu_l^{(r)}, \mu_h^{(r)}, \sigma_{ll}^{(r)}, \sigma_{lh}^{(r)}, \sigma_{hh}^{(r)}\right)$, $(r = 0, 1)$;

and also make the additional assumptions:

(b) m_i is marginally Bernoulli with $Pr(m_i = 1) = \pi$; $Pr(m_i = 0) = 1 - \pi = \pi_0$.

(c) Missingness of H given L and H depends only on an arbitrary function $g(H^*)$, $H^* = L + \lambda H$, for $\lambda \neq 0$.

Then under (a) and (b), we can specify a PMM for our data by re-expressing the joint distribution of $\mathbf{Y}$ and $\mathcal{M}$ as the product of the missing data pattern, $\mathcal{M}$, and a model of $\mathbf{Y}$ conditional on $\mathcal{M}$:

$$f(L, H, \mathcal{M} \mid \boldsymbol{\phi}, \pi) = f(L, H, \mid \mathcal{M}, \boldsymbol{\phi})\, f(\mathcal{M} \mid \pi), \tag{5.37}$$

where $\boldsymbol{\phi} = \{\phi^{(r)}\}$ and π are unknown parameters as defined in assumption (a), (b) above. This model setup implies that marginally (L, H) is a mixture of two normal distributions and that the marginal mean of (L_i, H_i) averaged over the two missingness patterns is $\boldsymbol{\mu} = (1-\pi)\boldsymbol{\mu}^{(0)} + \pi\boldsymbol{\mu}^{(1)}$. The parameter of interest $R = \log_2(\mu_l/\mu_h)$ is not

a parameter of the PMM, but is expressible as a function of the model parameters as

$$R = \log_2\left(\frac{\mu_l}{\mu_h}\right) \tag{5.38}$$

An important feature of PMMs is that they are by construction under-identified. Of the eleven parameters that define model (5.37), only eight can be identified from the data in the sense of appearing in the likelihood and having unique maximum likelihood estimates. The remaining three parameters of the regression of H on L for pattern $r = 1$ are not identifiable. Namely

$$\phi_{id} = \left(\pi, \mu_l^{(0)}, \mu_h^{(0)}, \sigma_{ll}^{(0)}, \sigma_{lh}^{(0)}, \sigma_{hh}^{(0)}, \mu_l^{(1)}, \sigma_{ll}^{(1)}\right) \tag{5.39}$$

$$\phi_{unid} = \left(\mu_h^{(1)}, \sigma_{lh}^{(1)}, \sigma_{hh}^{(1)}\right) \tag{5.40}$$

The likelihood under (5.37) takes the form

$$L\left(\phi_{id}\right) = \pi_0^{N_0}\pi^{N_1}\prod_{i=1}^{N_0} f\left(l_i, h_i \mid m_i = 0, \phi^{(0)}\right)\prod_{i=N_0+1}^{N} f\left(l_i \mid m_i = 1, \mu_l^{(1)}, \sigma_{ll}^{(1)}\right), \tag{5.41}$$

and yields maximum likelihood estimates

$$\hat{\pi} = (N - N_0)/N \tag{5.42}$$

$$\left\{\hat{\mu}_l^{(0)}, \hat{\mu}_h^{(0)}, \hat{\sigma}_{ll}^{(0)}, \hat{\sigma}_{lh}^{(0)}, \hat{\sigma}_{hh}^{(0)}\right\} = \left\{\bar{l}, \bar{h}, s_{ll}, s_{lh}, s_{hh}\right\} \tag{5.43}$$

$$\left\{\hat{\mu}_l^{(1)}, \hat{\sigma}_{ll}^{(1)}\right\} = \left\{\bar{l}^{(1)}, s_{ll}^{(1)}\right\}. \tag{5.44}$$

The set of statistics (5.43) is estimated from the N_0 complete cases while the second set (5.44) is estimated from the N_1 incomplete cases. Note that the three parameters of the conditional distribution of H given L for incomplete cases, $\phi_{h.l}^{(1)} = \left(\beta_{h0.l}^{(1)}, \beta_{hl.l}^{(1)}, \sigma_{hh.l}^{(1)}\right)$, do not appear in the likelihood given in (5.41) and need to be identified through imposing restrictions on the parameters or through prior informa-

tion.

In the absence of prior information, we can address this situation through the use of parameter restrictions, which can be imposed by setting inestimable parameters of the incomplete patterns equal to functions of the parameters describing the distribution of the complete cases. Under (c), the conditional distribution of L given H^* is independent of pattern, that is:

$$f\left(L \mid H^*, \phi^{(1)}, \mathcal{M}=1\right) = f\left(L \mid H^*, \phi^{(0)}, \mathcal{M}=0\right) \tag{5.45}$$

Little (1994) used the fact that

$$\phi^{(1)}_{h.h^*} = \phi^{(0)}_{h.h^*}, \tag{5.46}$$

where $\phi^{(r)}_{h.h^*} = \left(\beta^{(r)}_{h0.h^*}, \beta^{(r)}_{hl.h^*}, \sigma^{(r)}_{hh.h^*}\right)$ are the parameters representing the intercept, slope and residual covariance of the distribution of H given H^* for pattern r, to show that (5.46) leads to three parameter constraints on $\left(\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}\right)$ that are just sufficient to identify our bivariate PMM. The resulting maximum likelihood estimates of the mean and variance of H^* and the covariance of L and H^* are given by

$$\hat{\mu}_h = \bar{h} + b^{(\lambda)}_{hl.l}\left(\hat{\mu}_l - \bar{l}\right) \tag{5.47}$$

$$\hat{\sigma}_{hh} = s_{hh} + \left(b^{(\lambda)}_{hl.l}\right)^2\left(\hat{\sigma}_{ll} - s_{ll}\right) \tag{5.48}$$

$$\hat{\sigma}_{lh} = s_{lh} + b^{(\lambda)}_{hl.l}\left(\hat{\sigma}_{ll} - s_{ll}\right), \tag{5.49}$$

where $\hat{\mu}_l = \sum_{i=1}^{N} l_i/N$, and $b^{(\lambda)}_{hl.l} = \dfrac{\lambda s_{hh} + s_{lh}}{\lambda s_{lh} + s_{ll}}$.

We can then obtain an estimate for R by plugging in these maximum likelihood

estimates in (5.46):

$$\hat{R}^{(\lambda)} = \log_2\left(\frac{\hat{\mu}_l}{\hat{\mu}_h}\right) \tag{5.50}$$

$$= \log_2\left(\frac{\hat{\mu}_l}{\bar{h} + \hat{b}^{(\lambda)}_{hl.l}\left(\hat{\mu}_l - \bar{l}\right)}\right) \tag{5.51}$$

$$= \log_2\left[\frac{\left(1 - \frac{N_0}{N}\right)\bar{l}^{(0)} + \frac{N_0}{N}\bar{l}}{\bar{h} + \left(1 - \frac{N_0}{N}\right)\hat{b}^{(\lambda)}_{hl.l}\left(\bar{l}^{(0)} - \bar{l}\right)}\right], \tag{5.52}$$

where $\bar{l}^{(0)}$ is the mean of L for the N_1 cases missing H.

5.6.2 Choice of λ

Different choices of λ lead to different estimates of $\hat{R}^{(\lambda)}$. Note however that we are only interested in situations where $s_{lh} > 0$, since L and H, i.e, the matching *light* and *heavy* peptide signals for a given protein are always positively correlated. In this case, it is reasonable to assume that λ serves as a measure of the severity of the *non-ignorability* in the data. For positive λ, the value zero implies that missingness depends entirely on L, i.e., the data are MAR. This situation was discussed in Section (5.3.3) under a bivariate *monotone* MAR model. Higher values of λ are indicative of more extreme departures from MAR; and $\lambda \to \infty$ implies that missingness depends entirely on H.

The situation when $\lambda = 1$ is of particular interest since it corresponds to a scenario where the missingness mechanism depends on the sum of L and H. This is a reasonable characterization of empirical behavior since in practice the probability of missingness tends to be inversely proportional to the sum of the *light* and *heavy* signals. For example, in the case of a low abundant peptide, the sum of l and h is typically small, and the probability that h would not be detected above the signal-to-noise ratio even when l is detected, tends to be higher. On the other hand, if the sum of the two signals is large, then at least one of l, h must be relatively high. Then

given that l is detected h is also more likely to be non-missing.

An additional point of interest is the fact that the maximum likelihood estimate, $\hat{R}^{(\lambda)}$ reduces to the Complete-Case only (CC) estimate, $\log_2\left[\hat{\mu}_l/\left(\hat{\mu}_l + \bar{h} - \bar{l}\right)\right]$, when $\hat{b}^{(\lambda)}_{hl.l} = 1$, and $\lambda = (s_{ll} - s_{lh})/(s_{hh} - s_{lh})$. This value of λ reduces to $\lambda = 1$ when $s_{ll} = s_{hh}$. In other words, when the CC sample variances of L and H are equal and the missingness depends on the sum of L and H, the CC estimate of R is optimal.

5.6.3 Sensitivity Analysis

Since there is no data available to estimate the distribution of H given L for the incomplete cases, it is not possible to estimate λ from the data itself. In fact, the fit of the model to the observed data is identical for all choices of λ (Little and Rubin, (2002))[73]. The standard solution for handling this uncertainty about the choice of λ is to either specify a prior distribution or conduct a *sensitivity analysis* by specifying a range of plausible values for λ. In our analyses, we plan to adopt the latter approach to illustrate the sensitivity of results to three different choices for λ. Namely, $\boldsymbol{\lambda} = \{0.5, 1, \infty\}$.

We are particularly interested in studying the effect of λ on inference for $\mu_h = E(h_i)$. This is because for any 'null' protein, we expect $E(l_i) = E(h_i)$, and given that we have an estimate of $E(l_i) = \hat{\mu}_l$ from all N complete cases, we have a convenient means of investigating how close $\hat{\mu}_h$ is to $\hat{\mu}_l$ under different assumptions made about the missingness mechanism. Note that our choice of $\boldsymbol{\lambda}$ allows us to look at the behavior of $\hat{\mu}_h$ for progressively more severe violations of the MAR assumption. The planned *sensitivity analysis* requires that we construct normal 95% intervals for $\hat{\mu}^{(\lambda)}_h$. We achieve these intervals by using a Taylor series approximation to compute the variance of $\hat{\mu}^{(\lambda)}_h$:

$$\text{Var}\left(\hat{\mu}^{(\lambda)}_h\right) = \frac{1}{N}\hat{\sigma}_{hh} + \left(\hat{\mu}_l - \bar{l}\right)^2 var\left(b^{(\lambda)}_{hl.l}\right) + \frac{N_1}{N_0 N}\{s_{hh} - 2b^{(\lambda)}_{hl.l}s_{lh} + b^{(\lambda)^2}_{hl.l}s_{ll}\}, \quad (5.53)$$

where,

$$\mathrm{Var}\left(b_{hl.l}^{(\lambda)}\right) = \frac{\left(s_{ll}s_{hh} - s_{lh}^2\right)\left(\lambda^2 s_{hh} + 2\lambda s_{lh} + s_{ll}\right)^2}{N_0\left(\lambda s_{lh} + s_{ll}\right)^4}. \tag{5.54}$$

Appendices

5.1 Appendix

5.2 Chapter 5 - Appendices

5.2.1 Appendix A: Parameter Estimates of the Multivariate t Models Fitted to *YGR192C*

A.1. *Estimated parameters corresponding to 0% missingness*

$$\hat{v} = 119.354 \; ; \; \hat{\boldsymbol{\mu}} = \begin{bmatrix} 0.112 \\ 0.121 \\ -0.126 \\ 0.015 \\ -0.031 \end{bmatrix} \quad ; \quad \hat{\boldsymbol{\Psi}} = \begin{bmatrix} 0.706 & -0.015 & 0.201 & -0.049 & 0.064 \\ -0.015 & 1.115 & 0.127 & -0.055 & 0.090 \\ 0.201 & 0.127 & 0.992 & -0.096 & -0.036 \\ -0.049 & -0.055 & -0.096 & 0.593 & -0.098 \\ 0.064 & 0.090 & -0.036 & -0.098 & 1.306 \end{bmatrix}$$

A.2. *Estimated parameters corresponding to 10% missingness*

$$\hat{v} = 75.797 \; ; \; \hat{\boldsymbol{\mu}} = \begin{bmatrix} 0.132 \\ 0.089 \\ -0.100 \\ 0.086 \\ -0.206 \end{bmatrix} \quad ; \quad \hat{\boldsymbol{\Psi}} = \begin{bmatrix} 0.747 & -0.087 & 0.170 & -0.074 & 0.069 \\ -0.087 & 1.078 & 0.084 & -0.032 & -0.036 \\ 0.170 & 0.084 & 1.031 & -0.151 & -0.084 \\ -0.074 & -0.032 & -0.151 & 0.525 & -0.143 \\ 0.069 & -0.036 & -0.084 & -0.143 & 1.072 \end{bmatrix}$$

A.3. *Estimated parameters corresponding to 20% missingness*

$$\hat{v} = 67.033 \; ; \; \hat{\boldsymbol{\mu}} = \begin{bmatrix} 0.066 \\ 0.014 \\ -0.316 \\ 0.026 \\ -0.025 \end{bmatrix} \quad ; \quad \hat{\boldsymbol{\Psi}} = \begin{bmatrix} 0.717 & -0.055 & 0.038 & -0.117 & 0.046 \\ -0.055 & 1.161 & 0.153 & -0.005 & 0.011 \\ 0.038 & 0.153 & 1.060 & -0.031 & 0.127 \\ -0.117 & -0.005 & -0.031 & 0.733 & -0.109 \\ 0.046 & 0.011 & 0.127 & -0.109 & 1.238 \end{bmatrix}$$

A.4. *Estimated parameters corresponding to 30% missingness*

$$\hat{v} = 46.197 \; ; \; \hat{\boldsymbol{\mu}} = \begin{bmatrix} 0.124 \\ 0.199 \\ -0.118 \\ -0.104 \\ -0.117 \end{bmatrix} \quad ; \quad \hat{\boldsymbol{\Psi}} = \begin{bmatrix} 0.572 & 0.142 & 0.134 & 0.161 & 0.336 \\ 0.142 & 1.478 & -0.026 & -0.707 & 0.261 \\ 0.134 & -0.026 & 0.937 & 0.010 & 0.138 \\ 0.161 & -0.707 & 0.010 & 0.887 & 0.005 \\ 0.336 & 0.261 & 0.138 & 0.005 & 1.566 \end{bmatrix}$$

5.2.2 Appendix B: Posterior Distribution and Draws of μ_h, σ_{hh}, σ_{lh}, and $\hat{R}$

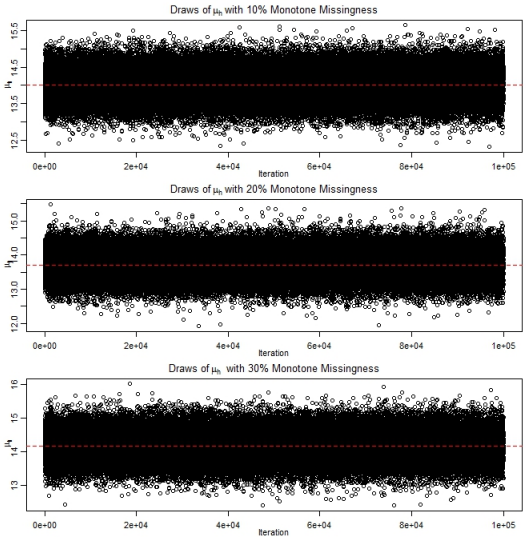


Figure 5.3: Draws of μ_h with 10%, 20%, and 30% missingness.

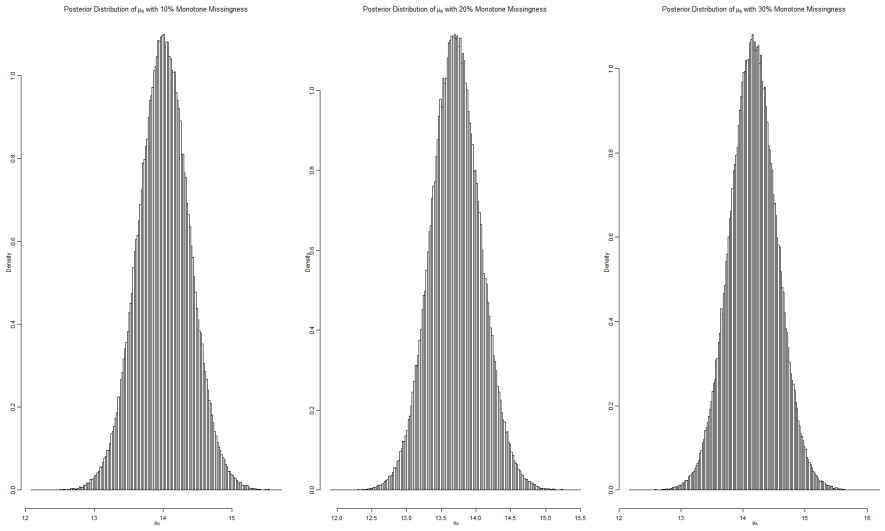


Figure 5.4: Posterior distribution of μ_h with 10%, 20%, and 30% missingness.

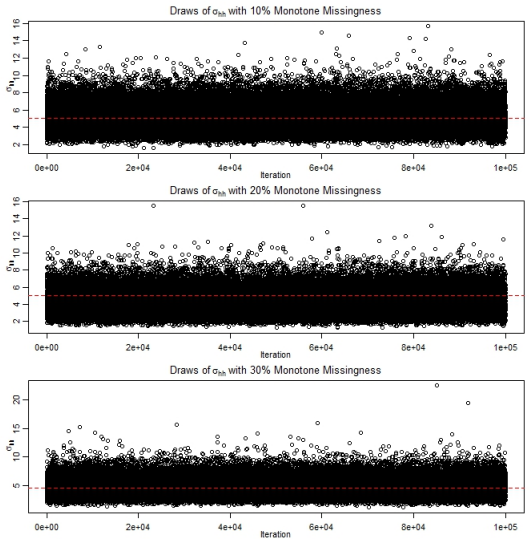


Figure 5.5: Draws of σ_{hh} with 10%, 20%, and 30% missingness.

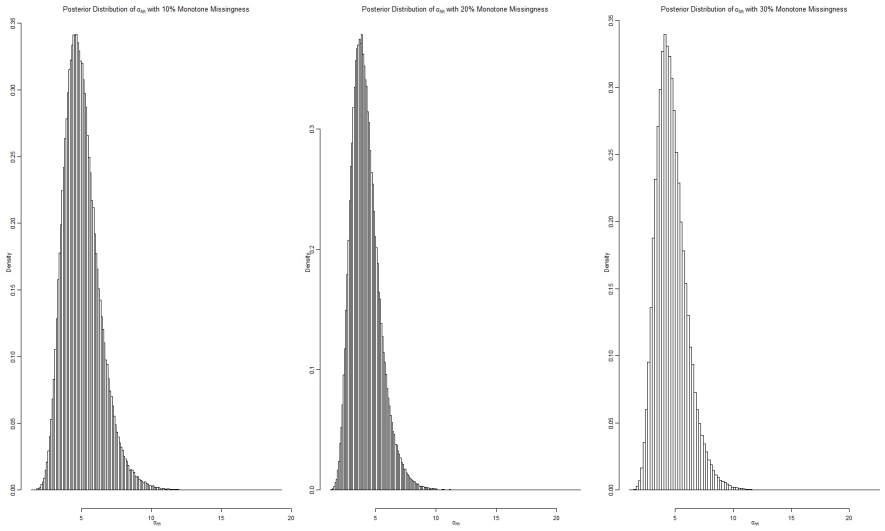


Figure 5.6: Posterior distribution of σ_{hh} with 10%, 20%, and 30% missingness.

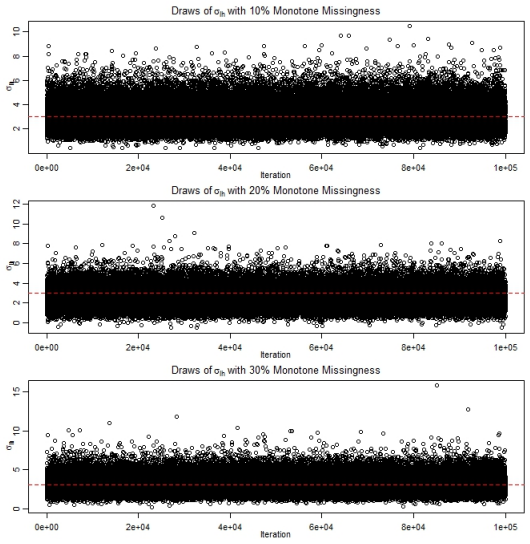


Figure 5.7: Draws of σ_{lh} with 10%, 20%, and 30% missingness.

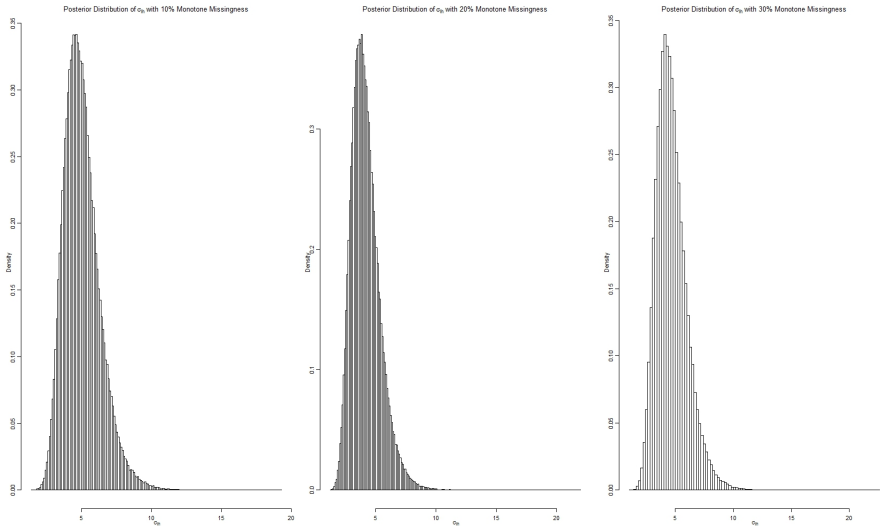


Figure 5.8: Posterior distribution of σ_{lh} with 10%, 20%, and 30% missingness.

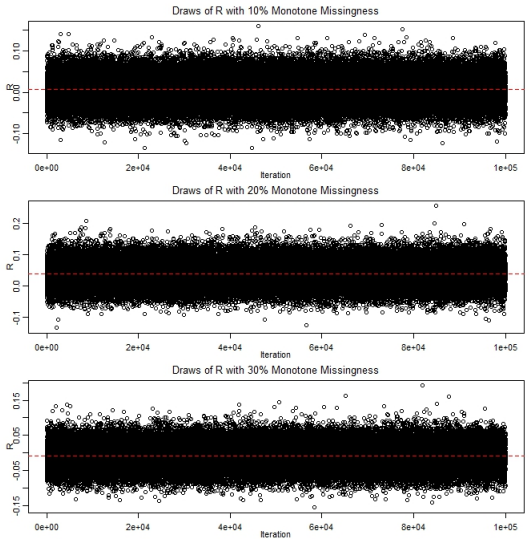


Figure 5.9: Draws of R with 10%, 20%, and 30% missingness.

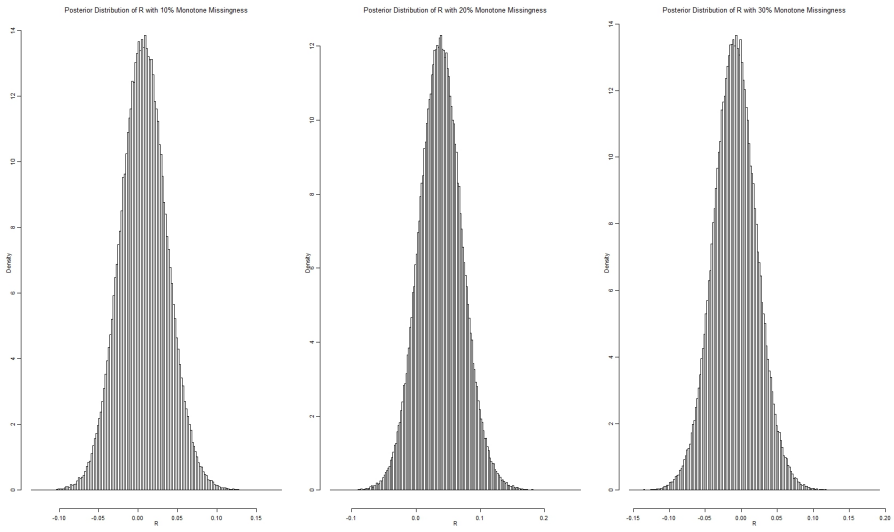


Figure 5.10: Posterior distribution of R with 10%, 20%, and 30% missingness.

Bibliography

[1] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions On Automatic Control*, 19(6):716–723, 1974.

[3] D. Allison, G. Gadbury, M. Heo, J. Fernández, C-K. Lee, T. Prolla, and R. Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1):1–20, 2002.

[4] N.L. Anderson, N.G. Anderson, L.R. Haines, D.B. Hardie, R.W. Olafson, and T.W. Pearson. Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (SISCAPA). *Journal of Proteome Research*, 3(2):235–244, 2004.

[5] T.W. Anderson, R. Gnanadesikan, and J.R. Kettenring. Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52:200–203, 1957.

[6] B.C. Arnold, R.J. Beaver, R.A. Groeneveld, and W.Q. Meeker. The Nontruncated Marginal of a Truncated Bivariate Normal Distribution. *Psychometrika*, 58(3):471–488, 1993.

[7] M.H. Asyali, M.M. Shoukri, O. Demirkaya, and K.S.A. Khabar. Assessment of

reliability of microarray data and estimation of signal thresholds using mixture modeling. *Nucleic Acids Research*, 32(8):2323–2335, 2004.

[8] J. Aubert, A. Bar-hen, J.J. Daudin, and S. Robin. Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics*, 5:125, 2004.

[9] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.

[10] A. Azzalini. Further results on a class of distributions which includes the normal ones. *Statistica*, 46:199–208, 1986.

[11] A. Azzalini and A. Capitaino. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society. Series B*, 65(2):367–389, 2003.

[12] V. Bafna and N. Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17:13–21, 2001.

[13] K.A. Baggerly, J.S. Morris, J. Wang, D. Gold, L.C. Xiao, and K.R. Coombes. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics*, 3(9):1667–1672, 2003.

[14] O.E. Barndorff-Nielsen. Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society London A*, 353:401–419, 1977.

[15] O.E. Barndorff-Nielsen and N. Shephard. Aggregation and model construction for volatility model. Working Paper 10, Centre for Analytical Finance, University of Århus, 1998.

[16] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1):289–300, 1995.

[17] R. Beran. Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697, 1988.

[18] R.J. Beynon, M.K. Doherty, J.M. Pratt, and S.J. Gaskell. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nature Methods*, 2(8):587–589, 2005.

[19] B. Blagoev, I. Kratchmarova, S.E. Ong, M. Nielsen, L.J. Foster, and M. Mann. A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nature Biotechnology*, 21(3):315–318, 2003.

[20] B. Blagoev, S.E. Ong, I. Kratchmarova, and M. Mann. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature Biotechnology*, 22(9):1139–1145, 2004.

[21] D. Böhning. *Computer-Assisted Analysis of Mixtures and Applications*. Chapman & Hall/CRC, 2000.

[22] G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA, 1973.

[23] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:11901208, 1995.

[24] J. Chang, H. Van Remmen, W.F. Ward, F.E. Regnier, A. Richardson, and J. Cornell. Processing of data generated by 2-dimensional gel electrophoresis

for statistical analysis: missing data, normalization, and statistics. *Journal of Proteome Research*, 3(6):1210–1218, 2004.

[25] T. Chen, M. Kao, M. Tepel, J. Rush, and G.M. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 8(3):325–337, 2001.

[26] X. Chen, P.L. Ulintz, E.S. Simon, J.A. Williams, and P.C. Andrews. Global topology analysis of pancreatic zymogen granule membrane proteins. *Molecular & Cellular Proteomics*, 7(12):2323–2336, 2008.

[27] K.R. Coombes, Jr. H.A. Fritsche, C. Clarke, J. Chen, K.A. Baggerly, J.S. Morris, L. Xiao, M. Hung, and H.M. Kuerer. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clinical Chemistry*, 49(10):1615–1623, 2003.

[28] J. Cox and M. Mann. Is proteomics the new genomics? *Cell*, 130(3):395–398, 2007.

[29] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26:1367–1372, 2008.

[30] V. Dančík, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.

[31] A.C. Davison, D.V. Hinkley, and B.J. Worton. Bootstrap likelihoods. *Biometrika*, 79(1):113–130, 1992.

[32] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete

data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[33] S.J. Devlin, R. Gnanadesikan, and J.R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76:354–362, 1981.

[34] E. Durr, J. Yu, K.M. Krasinska, L.A. Carver, J.R. Yates, J.E. Testa, P. Oh, and J.E. Schnitzer. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nature Biotechnology*, 22(8):985–992, 2004.

[35] E. Eberlein and U. Keller. Hyperbolic distributions in finance. *Bernoulli*, 1(3):281–299, 1995.

[36] E. Eberlein and E.V. von Hammerstein. Generalized hyperbolic and inverse gaussian distributions: Limiting cases and approximation of processes. Technical Report 80, University of Freiburg, http://www.stochastik.uni-freiburg.de/DYNSTOCH/papers/ghlimapprox2s.pdf, 2003.

[37] B. Efron. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002.

[38] B. Efron. Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377, 2007.

[39] B. Efron. Microarrays, Empirical Bayes and the Two-Groups Model. *Statistical Science*, 23(1):1–22, 2008.

[40] B. Efron and R. Tibshirani. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiologys*, 23(1):70–86, 2002.

[41] B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

[42] P.H.C. Eilers. Parametric time warping. *Analytical Chemistry*, 76(2):404–411, 2004.

[43] J.E. Elias and S.P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007.

[44] J.K. Eng, A.L. McCormack, and J.R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[45] B.S. Everitt and D.J. Hand. *Finite Mixture Distributions*. Chapman and Hall, London, 1981.

[46] C. Flecher, P. Naveau, and D. Allard. Estimating the closed skew-normal distribution parameters using weighted moments. *Statistics & Probability Letters*, 79(19):1977–1984, 2009.

[47] S. Frühwirth-Schnatter. *Finite Mixture and Markov Swithching Models*. Springer Verlag, 2006.

[48] M.G. Genton, L. He, and X. Liu. Moments of skew-normal random vectors and their quadratic forms. *Statistics & Probability Letters*, 51(4):319–325, 2001.

[49] S.A. Gerber, J. Rush, O. Stemman, M.W. Kirschner, and S.P. Gygi. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem

MS. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12):6940–6945, 2003.

[50] R.J. Glynn, N.M. Laird, and D.B. Rubin. *Drawing Inferences from Self-Selected Samples*. Springer-Verlag, New York, 1986.

[51] A. Gusnanto, S. Calza, and Y. Pawitan. Identification of differentially expressed genes and false discovery rate in microarray studies. *Current Opinion in Lipidology*, 18(2):187–193, 2007.

[52] A. Gut. *An Intermediate Course in Probability*. Springer, New York, 1995.

[53] S.P. Gygi, Y. Rochon, B.R. Franza, and R. Aebersold. Correlation between protein and mRNA abundance in yeast. *Molecular & Cellular Biology*, 19(3):1720–1730, 1999.

[54] D.K. Han, J. Eng, H. Zhou, and R. Aebersold. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology*, 19(10):946–951, 2001.

[55] E.L. Hendrickson, Q. Xia, T. Wang, J.A. Leigh, and M. Hackett. Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. *Analyst*, 131(12):1335–1341, 2006.

[56] N. Henze. A probabilistic representation of the "skew-normal" distribution. *Scandinavian Journal of Statistics*, 13:271–275, 1986.

[57] L.F. Hoogerheide, J.F. Kaashoek, and H.K. van Dijk. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks. *Journal of Econometrics*, 139(1):154–180, 2007.

[58] L.F. Hoogerheide and H.K. van Dijk. Possibly Ill-Behaved Posteriors in Econometric Models: On the Connection Between Model Structures, Non-Elliptical Credible Sets and Neural Network Simulation Techniques. Technical Report 036/4, Tinbergen Institute, Erasmus University Rotterdam, 2008.

[59] P.J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73101, 1964.

[60] S.R. Hurst, E. Platen, and S.T. Rachev. Option pricing for asset returns driven by subordinated processes. Working paper, Australian National University, School of Mathematical Sciences, Canberra, 1995.

[61] Y.Y. Jung, M.S. Oh, D.W. Shin, S.H. Kang, and H.S. Oh. Identifying differentially expressed genes in meta-analysis via Bayesian model-based clustering. *Biometrical Journal*, 48(3):435–450, 2006.

[62] M. Karas, A. Schmidt, and T. Dlcks. Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: When does ESI turn into nano-ESI? *Journal of the American Society for Mass Spectrometry*, 14(5):492–500, 2003.

[63] A. Keller, A.I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistrys*, 74(20):5383–5392, 2002.

[64] B. Kim, A.I. Nesvizhskii, P.G. Rani, S. Hahn, R. Aebersold, and J.A. Ranish. The transcription elongation factor TFIIS is a component of RNA polymerase II preinitiation complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 104(41):16068–16073, 2007.

[65] M.L.T. Lee, F. Kuo, G. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from

repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):9834–9839, 2000.

[66] X.J. Li, H. Zhang, J.A. Ranish, and R. Aebersold. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Analytical Chemistry*, 75(23):6648–6657, 2003.

[67] Tsung I. Lin, Jack C. Lee, and Shu Y. Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17:909–927, 2007.

[68] W.T. Lin, W.N. Hung, Y.H. Yian, K.P. Wu, C.L. Han, Y.R. Chen, Y.J. Chen, T.Y. Sung, and W.L. Hsu. Multi-Q: a fully automated tool for multiplexed protein quantitation. *Journal of Proteome Research*, 5(9):2328–2338, 2006.

[69] D.V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2, Inference*. Cambridge University Press, Cambridge, UK, 1965.

[70] R.J.A. Little. Robust estimation of the mean and covariance matrix from data with missing values. *Journal of the Royal Statistical Society. Series C*, 37(1):23–38, 1988.

[71] R.J.A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

[72] R.J.A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[73] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, 2002.

[74] R.J.A. Little and P.J. Smith. Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82:58–68, 1987.

[75] C. Liu and D.B. Rubin. ML estimation of the multivariate t distribution with unknown degrees of freedom. *Statistica Sinica*, 5:19–39, 1995.

[76] Q. Liu, B. Krashnapuram, P. Pratapa, X. Liao, A. Hartemink, and L. Carin. Identification of differentially expressed proteins using MALDI-TOF mass spectra, 2003.

[77] P. Mallick, M. Schirle, S.S. Chen, M.R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology*, 25(1):125–131, 2007.

[78] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66(24):4390–4399, 1994.

[79] M. Marelli, J.J. Smith, S. Jung, E. Yi, A.I. Nesvizhskii, R.H. Christmas, R.A. Saleem, Y.Y. Tam, A. Fagarasanu, D.R. Goodlett, R. Aebersold, R.A. Rachubinski, and J.D. Aitchison. Quantitative mass spectrometry reveals a role for the GTPase Rho1p in actin organization on the peroxisome membrane. *Journal of Cell Biology*, 167(6):1099–1112, 2004.

[80] G. Marsaglia and J. Marsaglia. Evaluating the anderson-darling distribution. *Journal of Statistical Software*, 9(2):1–5, 2004.

[81] E.H. Martínez, H. Varela, H.W. Gómez, and H. Bolfarine. A note on the likelihood and moments of the skew-normal distribution. *Statistics & Operations Research Transactions*, 32(1):57–66, 2008.

[82] G.J. McLachlan. On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Applied Statistics*, 36(3):318–324, 1987.

[83] G.J. McLachlan and K.E. Basford. *Mixture models. Inference and applications to clustering*. Dekker, New York, 1988.

[84] G.J. McLachlan, R.W. Bean, and L.B. Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformaticss*, 22(13):1608–1615, 2006.

[85] G.J. McLachlan and P. Jones. Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, 44(2):571–578, 1988.

[86] G.J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[87] P. Mertins, H.C. Eberl, J. Renkawitz, J.V. Olsen, M.L. Tremblay, M. Mann, A. Ullrich, and H. Daub. Investigation of protein-tyrosine phosphatase 1B function by quantitative proteomics. *Molecular & Cellular Proteomics*, 7(9):1763–1777, 2008.

[88] M.P. Molloy, E.E. Brzezinski, J. Hang, M.T. McDowell, and R.A. VanBogelen. Overcoming technical variation in quantitative proteomics. *Proteomics*, 3(10):1912–1919, 2003.

[89] A.I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistrys*, 75(17):4646–4658, 2003.

[90] M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8(1):37–52, 2001.

[91] N.V. Nielsen, J.M. Carstensen, and J. Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis

using correlation optimised warping. *Journal of Chromatography A*, 805(1-2):17–35, 1998.

[92] S.E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1(5):376–386, 2002.

[93] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546554, 2002.

[94] W. Pan, L. Lin, and C.T. Le. A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics*, 3(3):117–124, 2003.

[95] D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell. Probability based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[96] S. Pounds and S.W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):12361242, 2003.

[97] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 1992.

[98] A. Ramos-Fernández, D. López-Ferrer, and J. Vázquez. Improved method for differential expression proteomics using trypsin-catalyzed $^{1}8$o labeling with a correction for labeling efficiency. *Molecular & Cellular Proteomics*, 6(7):1274–1286, 2007.

[99] T.W. Randolph and Y. Yasui. Multiscale processing of mass spectrometry data. *Biometrics*, 62(2):589–597, 2006.

[100] J.A. Ranish, E.C. Yi, D.M. Leslie, S.O. Purvine, D.R. Goodlett, J. Eng, and R. Aebersold. The study of macromolecular complexes by quantitative proteomics. *Nature Genetics*, 33(3):349–355, 2003.

[101] C.R. Rao. Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65(329):161–172, 1970.

[102] C.R. Rao. Estimation of variance and covariance components - MINQUE theory. *Journal of Multivariate Analysis*, 1(3):257–275, 1971.

[103] C.R. Rao. Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67(337):112–115, 1972.

[104] P.S.R.S. Rao, J. Kaplan, and W.G. Cochran. Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association*, 76(373):89–97, 1981.

[105] P.L. Ross, Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D.J. Pappin. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12):1154–1169, 2004.

[106] D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[107] S. Ryu, B. Gallis, Y.A. Goo, S.A. Shaffer, D. Radulovic, and D.R. Goodlett. Comparison of a label-free quantitative proteomic method based on peptide ion current area to the isotope coded affinity tag method. *Cancer Informatics*, 6:243–255, 2008.

[108] G.A. Satten, S. Datta, H. Moura, A.R. Woolfitt, G. Carvalho, G.M. Carlone, B.K. De, A. Pavlopoulos, and J.R. Barr. Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, 20(17):3128–3136, 2004.

[109] A. Savitzky and M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[110] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

[111] S.M. Shah and M.C. Jaiswal. Estimation of parameters of doubly truncated normal distribution from first four sample moments. *Annals of the Institute of Statistical Mathematics*, 18(1):107–111, 1966.

[112] S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53:683690, 1991.

[113] J.D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445, 2003.

[114] J.A. Taylor and R.S. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9):1067–1075, 1997.

[115] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q. Le. Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, 20(17):3034–3044, 2004.

[116] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

[117] R.J.O. Torgrip, M. Aberg, B. Karlberg, and S.P. Jacobsson. Peak alignment using reduced set mapping. *Journal of Chemometrics*, 17(11):573582, 2003.

[118] B.B. Turnbull. Optimal estimation of false discovery rates. Technical report, Stanford University, http://www.stanford.edu/ bkatzen/optimal-FDR.pdf, 2007.

[119] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D.Ward, K. Williams, and H. Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003.

[120] Y. Yasui, D. McLerran, B.L. Adam, M. Winget, M. Thornquist, and Z.D. Feng. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Journal of Biomedicine & Biotechnology*, 4:242–248, 2003.

[121] Y. Yasui, M. Pepe, M.L. Thompson, B. Adam, G.L. Wright, Jr., Y. Qu, J.D. Potter, M. Winget, M. Thornquist, and Z. Feng. A data analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4(3):449–463, 2003.

[122] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao. Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Computational Biology & Chemistry*, 30(1):27–38, 2006.