Quantifying Sources of Variation in High-throughput Biology

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Franks, Alexander M. 2015. Quantifying Sources of Variation in High-throughput Biology. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:17463988">http://nrs.harvard.edu/urn-3:HUL.InstRepos:17463988</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Quantifying Sources of Variation in High-throughput Biology

A DISSERTATION PRESENTED
BY
ALEXANDER M. FRANKS
TO
THE DEPARTMENT OF STATISTICS
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2015
Quantifying Sources of Variation in High-throughput Biology

ABSTRACT

One of the central challenges in systems biology research is disentangling relevant and irrelevant sources of variation. While the relevant quantities are always context dependent, an important distinction can be drawn between variability due to biological processes and variability due measurement error. Biological variability includes variability between mRNA or protein abundances within a well defined condition, variability of these abundances across conditions (physiological variability), and between species or between subject variability. Technical variability includes measurement error, technological bias, and variability due to missing data. In this dissertation, we explore statistical challenges associated with disentangling sources of variability, both biological and technical, in the analysis of high-throughput biological data. In the first chapter, we present a careful meta-analysis of 27 yeast data sets supported by a multilevel model to separate biological variability from structured technical variability. In the second chapter, we suggest a simple and general approach for deconvolving the contributions of orthogonal sources of biological variability, both between and within molecules, across multiple physiological conditions. The results discussed in these two chapters elucidate the relative importance of transcriptional and post-transcriptional regulation of protein levels. Finally, in the third chapter we introduce a novel approach for modeling non-ignorable missing data. We illustrate the utility of this methodology on missing data in mRNA and protein measurements.
Contents

1 Introduction 1

2 Estimating a structured covariance matrix from multi-lab measurements in high-throughput biology 4
   2.1 Introduction ......................................................... 4
   2.2 Methods .............................................................. 12
   2.3 Theory .............................................................. 21
   2.4 Results .............................................................. 26
   2.5 Discussion ............................................................ 39

3 Quantification of post-transcriptional regulation across human tissues 44
   3.1 Introduction .......................................................... 44
   3.2 Results ............................................................... 46
   3.3 Methods ............................................................. 54
   3.4 Discussion ........................................................... 61

4 Conditionally specified Bayesian models for data missing not at random 62
   4.1 Introduction .......................................................... 62
   4.2 Models for Data Missing Not At Random ......................... 64
   4.3 Tukey’s Factorization ................................................ 67
   4.4 Bayesian Models For Missing Data ................................ 69
   4.5 Simulation: A Discrete-Continuous Mixture ...................... 75
   4.6 Application: Analysis of Transcriptomic and Proteomic Data . 81
   4.7 Discussion ............................................................ 86

Appendix A (Chapter 1) 89
   A.1 Proofs ............................................................... 89
   A.2 Additional figures and tables ...................................... 98
APPENDIX B (CHAPTER 2)  

APPENDIX C (CHAPTER 3)  
C.1 Derivations for Section 4.5 ........................................ 104 
C.2 Derivations for Section 4.6 ........................................ 106 
C.3 Additional results from Section 4.6 ............................. 108 

REFERENCES  

120
## Listing of figures

2.1 mRNA-protein correlations are structured .......................................................... 10  
2.2 Data is not missing at random .............................................................................. 11  
2.3 Structured covariance matrix ................................................................................ 16  
2.4 Asymmetric Laplace distributions ......................................................................... 28  
2.5 mRNA-protein correlation with technology random effects ................................. 35  
3.1 Simpson’s paradox of protein regulation .................................................................. 48  
3.2 Correlation simulation ............................................................................................ 49  
3.3 Fold-change correlations ......................................................................................... 51  
3.4 Error in quantifying proteins .................................................................................. 53  
3.5 Post-transcriptional regulation in two tissues ......................................................... 55  
4.1 Missing data parameters are highly correlated ....................................................... 75  
4.2 Simulation study data generating distributions ....................................................... 80  
4.3 Distortion due to selection factorization misspecification ...................................... 81  
4.4 Selection factorization vs Tukey’s factorization (protein) ..................................... 84  
4.5 Posterior complete data quantities (protein) .......................................................... 85  
4.6 Posterior samples of the ratio of complete data standard deviations ..................... 86  
A.1 True and misspecified covariance matrices ......................................................... 91  
A.2 Distribution of observed and imputed mRNA and protein levels ......................... 99  
A.3 PCA of mRNA replicates ......................................................................................... 99  
B.1 Between and within gene correlation histograms ............................................... 102  
B.2 Summary of rPTR variability .................................................................................. 103  
C.1 Selection factorization vs Tukey’s factorization (mRNA) ..................................... 108  
C.2 Posterior complete data quantities (mRNA) ......................................................... 109
Acknowledgments

First, I would like to thank my advisor, Professor Edoardo Airoldi. He has been an immensely supportive mentor and collaborator throughout my graduate studies and was integral to all of the work appearing in this dissertation. Edo has been incredibly helpful with identifying interesting research problems and working with me to further my career. I would also like to thank my other thesis committee members Don Rubin and Dr. Allan Drummond for their suggestions, general guidance, and for many interesting conversations.

Collaborative research has been at the heart of my graduate career, and I would like to extend my thanks to everyone I have worked with, including, but not limited to, Edward Wallace, Nikolai Slavov, and Florian Markowetz. I owe special thanks to my close collaborator Gabor Csardi.

I also need to thank the entire Airoldi lab, who sat through my presentations more times than I can count, and have provided invaluable feedback at different stages of my PhD. More generally, I would like to acknowledge the entire Department of Statistics for their support and wisdom. In particular, Professor Luke Bornn, Alex D’amour, Dan Cervone, Robin Gong, Alex Volfovsky and Andrew Miller have been critical to my development as a statistician. Most of what I have learned during my PhD has come from conversations with my peers in the department.

Finally, I would like to acknowledge my friends and family, all of whom have been extremely supportive throughout my PhD.
One of the central challenges in systems biology research is disentangling relevant and irrelevant sources of variation. While the “relevant” quantities are always context dependent, an important distinction can be drawn between variability due to true biological signal and that due to technical variation or measurement error. Biological variability includes variability between molecule abundances (e.g. mRNA or protein) within a well defined condition, variability of these abundances across con-
ditions (physiological variability), and between species or between subject variability. Technical variability includes measurement error, technological bias, and variability due to missing data. Generally, this is an “irrelevant” source of variability which we hope to remove. Understanding the scientific implications of biological variability is usually the primary aim. In practice, these multiple sources of variability are often confused, leading to spurious scientific conclusions. In this dissertation, we explore some statistical challenges associated with disentangling sources of variability, both biological and technical, in the analysis of high-throughput biological data.

First, in Chapter 1, we consider the problem of quantifying the degree of coordination between transcription and translation, in a well defined condition in yeast. Several studies have reported a surprising lack of coordination over the years, in organisms as different as yeast and human, using diverse technologies. However, a close look at this literature suggests that the lack of reported correlation may not reflect the biology of regulation, but rather superfluous variation due to measurement error and missing data. These reports do not control for between-study biases and structure in the measurement errors, ignore key aspects of how the data connect to the estimand, and systematically underestimate the correlation as a consequence. We design a careful meta-analysis of 27 yeast data sets and address the impact of noise, model mis-specification and non-ignorable missing data on estimates of the correlation. Ultimately, we show that the correlation between mRNA and protein levels is quite high under the studied condition suggesting that post-transcriptional regulation plays a less prominent role than previously thought.

In Chapter 2, we further investigate the degree of coordination between transcription and translation, now using data from human tissues. Specifically, we focus on the implications of two orthogonal sources of biological variation in transcriptomic and
proteomic data. We largely confirm our conclusions from the first chapter with human data by showing that the correlation between mRNA transcripts and protein abundances in any given condition is large. However, for any individual gene mRNA and protein levels across conditions are much less variable, and thus it is more difficult to assess the significance of any functional relationships. We gain some traction on this problem by pooling information between molecules by conducting a rigorous functional gene set analysis. We estimate the factors determining these two orthogonal variations and demonstrate that the variance between the levels of different proteins is dominated by transcriptional regulation while in stark contrast the physiological variance across conditions is largely determined by translational regulation. By disentangling distinct types of biological variation, we are able to reconcile existing estimates in the literature, caution against estimating protein fold-changes from mRNA fold-changes, and highlight the dominance of translational regulation in determining the physiological variation of a protein across conditions.

Finally, in Chapter 3, we present a novel method for addressing variability due to non-ignorable missing data. In any analysis involving missing data, we must establish reasonable assumptions about the missingness mechanisms. Unfortunately, when the data is missing not at random (MNAR), assumptions about the missing data mechanism are unverifiable. As such, it is necessary to propagate uncertainty about the missingness mechanism into our uncertainty about the estimands of interest. To address these issues, we introduce “Tukey’s factorization”, a modeling approach which is based on specifying a joint through conditional distributions. We demonstrate how these models are flexible, facilitate sensitivity analysis and model checking, and are generally computationally tractable. We evaluate the utility of this approach in simulation and in the analysis of high-throughput biological data.
2

Estimating a structured covariance matrix from multi-lab measurements

2.1 Introduction

We consider the problem of estimating the degree of coordination between transcription and translation, in yeast. A credible estimate would have two important sub-
stantive implications. It would help assess the extent to which analyses that take measures of transcription as proxies for measures of translation, are valid. A credible estimate would also help quantify the relative roles of transcriptional versus post-transcriptional regulation.

Several studies have addressed this problem over the years, in organisms as different as yeast and human, with diverse technologies\textsuperscript{40,1,15,47,910,92}. Typically, transcription is quantified in terms of the concentration of messenger RNA (mRNA), corresponding to different genes, while translation is quantified in terms of the ratio of protein abundance to mRNA. If rates of translation and degradation did not vary by gene, then protein-mRNA ratios would be constant, and mRNA-protein levels would be perfectly correlated\textsuperscript{24}. Accordingly, the correlation between the vectors of mRNA and protein concentrations has been used as a proxy for the degree of post-transcriptional regulation. Published estimates of the correlation are low, mostly between 0.3 and 0.6, and do not seem to increase with more modern technologies. Thus, the consensus is that there is significant regulation of protein levels after transcription, especially in higher organisms and mammals. This finding is quite surprising. The community agrees the extent to which mRNA and protein levels correlate is still unclear\textsuperscript{111}.

A close look at this literature suggests that the lack of reported correlation is not surprising after all. These studies are not based on a careful design, nor they carry out statistical analyses carefully, and ignore key aspects of how the data connect to the estimand. For instance, analyses are often limited to complete cases, discarding mRNAs and proteins with missing measurements. They ignore that missing measurements are more likely to be taken on mRNAs and proteins that are rare in cell. Structure in the variability of measurements, often referred to as \textit{batch effects}\textsuperscript{59}, is not
accounted for. Arguably, the low reported correlations are more likely to be due to limitations in the designs and analyses, rather than to limitations in the technology, or to aspects of regulation.

Conceptually, we can decompose the correlation into contributing components that should inform an appropriate study design and analysis. Namely, the main components that contribute to variation in the observed correlation between mRNA and protein levels are: differences in strain, technology and growth rate, the amount of alternative splicing, additional variability structured according to experiments, replicated measurements within an experiment, and actual biological variation\textsuperscript{83,113}.

Here, we design an original meta-analysis of 27 yeast data sets, supported by a multilevel model, full uncertainty quantification, a suite of sensitivity analyses and novel theory, to produce a more accurate estimate of the correlation between mRNA and protein levels. Briefly, the proposed design controls for strain and reported growth rate, includes multiple technologies for measuring mRNA and protein levels. A simple multilevel model accounts for the structure in the meta variance-covariance matrix, and includes a non-ignorable missing data mechanism for missing measurements\textsuperscript{34,89,46}. A limited amount of splicing in yeast\textsuperscript{77} and other sources of variation contribute to the residual error. The strategy for the meta-analysis is to first fit a simple normal-normal multilevel model, in which technologies are assumed as exchangeable. While this model is theoretically identifiable in the absence of missing data, or in the presence of data missing completely at random, properties of the inference under under non-ignorable missing data are uncertain. We show empirically that inference achieves nominal frequentist coverage for a number of key parameters in the presence of non-ignorable missing data, using posterior predictive meta data, in Section 2.4.1, and that the model is robust to departures from normality, in Section 2.4.2. In Section
2.4.3 we use this model to estimate the correlation between mRNA and protein levels. We then explore the impact of relaxing the exchangeable technologies assumption on the correlation estimates, in Section 2.4.4.

From a statistical perspective, this problem motivates new theory on the impact of noise, model mis-specifications and non-ignorable missing data on estimates of the correlation, in Section 2.3. These theoretical results are illustrated by the analysis presented in Section 2.4.2. It is worthwhile noting that, while standard theory exists that characterizes the impact of measurement noise and model mis-specifications on mean coefficients, and in some cases variance coefficients, there is no theory that characterizes the impact of such specifications on the covariance or correlation between high-dimensional responses, e.g., mRNA and protein concentrations, which is the estimandum of interest in the problem we consider.

From a substantive perspective, we find that the correlation between mRNA and protein levels is quite high, in yeast, suggesting that post-translational regulation plays a less prominent role than previously thought.

2.1.1 Data collection and exploratory data analysis

We gathered 16 data sets that measure mRNA expression and 11 that measure protein concentrations, mostly published, yielding a total of 58 high-throughput measurements on 5,308 genes and their corresponding proteins in yeast. The measurements were taken on yeast cultures using different technologies including custom and commercial microarrays, high-throughput sequencing and mass spectrometry.

The goal of the analysis is to study the steady state correlation of mRNA and protein levels. Thus it is important to use data that were collected under similar experimental conditions; from haploid yeast S. cerevisiae growing exponentially in rich
shaken liquid medium with 2% glucose between 22 and 30°C. Additional sources of variation are treated as noise for the purpose of the analysis.

Details of the data sets are summarized in Table 2.1. Throughout the chapter we work with the natural logarithm of the raw data, as this is approximately normally distributed. This is standard in mRNA expression and protein abundance studies.\textsuperscript{29}

The data sets in Table 2.1 have features that, if unaccounted for, are likely to result in poor estimates of the correlation of interest. First, the measurements are inherently noisy. Both biological and technical noise attenuate correlation estimates; we define attenuation as bias towards zero. Second, the measurements are structured. We refer to an “experiment” to indicate a set of replicated measurements, whether technical or biological, which were obtained with a specific biotechnology and published in a specific paper e.g.,\textsuperscript{47,61}. The data we collected can be grouped according to biotechnology and experiment. As expected, the variability of the mRNA expression values is larger between experiments than between replicated measurements within an experiment (Figure 2.1). Interestingly, the range of the observed mRNA-protein correlations is almost the same as the between-experiment correlations, for both mRNA and protein levels. Principal component analysis of the replicates (see Figure A.3 in the Appendix) confirms that experiment effects are large.

Third, a considerable portion of the data in any given experiment is missing. On average, over 25% of the values are missing in any replicated measurement, for either mRNAs or proteins, with some experiments missing over 95% of the values. The data sets with a very large number of measurements missing may be of questionable value for estimating the mRNA-protein correlation but they are included for completeness. These are classic data sets that originally led to the conclusion that mRNA and protein levels correlated poorly, and so their inclusion is natural.
<table>
<thead>
<tr>
<th>ID</th>
<th>Reference</th>
<th>Technology (measurements)</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAUS</td>
<td>Causton et al. (^\text{16})</td>
<td>commercial microarray (x5)</td>
<td>19-22%</td>
</tr>
<tr>
<td>DUD</td>
<td>Dudley et al. (^\text{27})</td>
<td>custom microarray (x4)</td>
<td>5%</td>
</tr>
<tr>
<td>GARC</td>
<td>Garcia-Martínez et al. (^\text{32})</td>
<td>custom microarray</td>
<td>1%</td>
</tr>
<tr>
<td>HOLS</td>
<td>Holstege et al. (^\text{44})</td>
<td>commercial microarray</td>
<td>12%</td>
</tr>
<tr>
<td>ING1</td>
<td>Ingolia et al. (^\text{47})</td>
<td>RNA-Seq (mRNA rich) (x2)</td>
<td>9-10%</td>
</tr>
<tr>
<td>ING2</td>
<td>Ingolia, unpublished</td>
<td>RNA-Seq (rq) (x2)</td>
<td>4-5%</td>
</tr>
<tr>
<td>ING3</td>
<td>Ingolia, unpublished</td>
<td>RNA-Seq (ca) (x2)</td>
<td>6-8%</td>
</tr>
<tr>
<td>LIP1</td>
<td>Lipson et al. (^\text{61})</td>
<td>RNA-Seq (x6)</td>
<td>1%</td>
</tr>
<tr>
<td>LIP2</td>
<td>Lipson et al. (^\text{61})</td>
<td>commercial microarray</td>
<td>4%</td>
</tr>
<tr>
<td>MAC</td>
<td>MacKay et al. (^\text{67})</td>
<td>custom microarray</td>
<td>28%</td>
</tr>
<tr>
<td>MIUR</td>
<td>Miura et al. (^\text{70})</td>
<td>cPCR (x4)</td>
<td>26-29%</td>
</tr>
<tr>
<td>NGAL</td>
<td>Nagalakshmi et al. (^\text{72})</td>
<td>RNA-Seq</td>
<td>22%</td>
</tr>
<tr>
<td>PELE</td>
<td>Pelechano &amp; Pérez-Ortín (^\text{79})</td>
<td>custom microarray</td>
<td>14%</td>
</tr>
<tr>
<td>ROTH</td>
<td>Roth et al. (^\text{86})</td>
<td>commercial microarray (x2)</td>
<td>59-70%</td>
</tr>
<tr>
<td>VELC</td>
<td>Velculescu et al. (^\text{109})</td>
<td>SAGE</td>
<td>58%</td>
</tr>
<tr>
<td>YASS</td>
<td>Yassour et al. (^\text{120})</td>
<td>RNA-Seq (x4)</td>
<td>5%</td>
</tr>
<tr>
<td>FUTR</td>
<td>Fütcher et al. (^\text{31})</td>
<td>2D gel</td>
<td>99%</td>
</tr>
<tr>
<td>GHAM</td>
<td>Ghaemmaghami et al. (^\text{37})</td>
<td>Western blot</td>
<td>34%</td>
</tr>
<tr>
<td>GODO</td>
<td>de Godoy et al. (^\text{23})</td>
<td>LC MS/MS</td>
<td>25%</td>
</tr>
<tr>
<td>GYGI</td>
<td>Gygi et al. (^\text{40})</td>
<td>2D gel</td>
<td>98%</td>
</tr>
<tr>
<td>LEE</td>
<td>Lee et al. (^\text{57})</td>
<td>LC MS/MS (x3)</td>
<td>67-76%</td>
</tr>
<tr>
<td>LU</td>
<td>Lu et al. (^\text{66})</td>
<td>LC MS/MS</td>
<td>83%</td>
</tr>
<tr>
<td>NGAR</td>
<td>Nagaraj et al. (^\text{73})</td>
<td>LC MS/MS (x6)</td>
<td>31%</td>
</tr>
<tr>
<td>NEWM</td>
<td>Newman et al. (^\text{74})</td>
<td>GFP</td>
<td>60%</td>
</tr>
<tr>
<td>PENG</td>
<td>Peng et al. (^\text{80})</td>
<td>LC MS/MS</td>
<td>74%</td>
</tr>
<tr>
<td>THAK</td>
<td>Thakur et al. (^\text{106})</td>
<td>LC MS/MS (x3)</td>
<td>84-85%</td>
</tr>
<tr>
<td>WASH</td>
<td>Washburn et al. (^\text{117})</td>
<td>LC MS/MS</td>
<td>77%</td>
</tr>
</tbody>
</table>

**Table 2.1:** List of mRNA data sets (above the midline) and protein concentration data sets (below the midline). If the data set has multiple measurements, the number of replicates in each data set is given after the technology name, in parentheses. ‘2D gel’ stands for two-dimensional gel electrophoresis, and ‘MS’ for mass-spectrometry. The last column is the missingness rate out of the 5,308 genes in our data set.

Notably, it is harder to obtain mRNA expression and protein concentration values for mRNA transcripts and proteins that are rare in the cell. A quick analysis of
**Figure 2.1:** mRNA expression data (left panel) and protein concentrations data (right panel) are highly structured. The plots show naive, biased Pearson correlation estimates between pairs of replicated measurements on the intersection of observed mRNAs/proteins; separately for replicates within experiments (solid) and across experiments (dashed). The thin black line in each panel shows the naive correlations between mRNA expression and protein replicates. The observed mRNA expression–protein correlations are comparable to the between-experiment correlations for both mRNA expression and protein levels. The top labels indicate the mean pairwise correlation between and within experiments.

replicated measurements suggests that the fraction of missing values appears to be inversely related to the average observed values of both mRNA and protein concentrations. This analysis is illustrated in Figure 2.2 and in Table A.1 in the Appendix.

We give some theoretical insights in Section 2.3 on how each of these three effects attenuate the observed correlation, and also perform an analysis of simulated data in Section 2.4.2.

### 2.1.2 Contributions of this work

We estimate the degree of coordination between transcription and translation, in yeast. To accomplish that, we have curated a collection of 27 yeast data sets about
**Figure 2.2:** Unavailable values are not missing at random. The bars show the average observed mRNA levels (left panel) and protein concentration values (right panel), standardized, plotted as a function of the number of missing values for each mRNA (out of 38 total), or protein (out of 20 total). Bar widths are proportional to the number of genes (proteins) in each bin.

mRNA and protein levels, in Table 2.1. We developed an original meta analysis strategy to estimate the amount of coordination, which we quantify in terms correlation between latent de-noised representations of mRNA and protein levels. This correlation is a parameter in a simple multilevel model that accounts for measurement error structure due to experimental protocols, replicated measurements, and technology biases. The analysis involves Bayesian confidence intervals, a suite of sensitivity analyses including an evaluation of frequentist coverage, robustness of the estimates to departure from key assumptions, such as normality and correct specifications of the covariance structure, and the effects of technological bias on the estimates. We also develop novel theory that provides analytical insights into the results of the sensitivity analyses we perform. Namely, we quantify the expected reduction in correlation as a function of (1) noise in the data; (2) experiment effects and model misspecifica-
tion; and (3) non-ignorable missing data. This theory extends Spearman’s correction for the attenuation of correlation \(^{102}\) between two quantities to a multivariate setting while accounting for experiment effects. In particular, while corrections for the effect of missing data on exploratory analyses have been explored \(^{118}\), we are the first, to our knowledge, to discuss the estimation of correlation from multiple measurements each with different non-ignorable missing data mechanisms.

2.2 Methods

We posit a simple model to estimate a covariance matrix between high-dimensional responses, in the presence of structured measurement errors and non-ignorable missing data, and we develop a Markov chain Monte Carlo algorithm to perform inference. Models of this sort are well established in statistical applications e.g., see \(^{90,48}\). We chose a combination of simple specifications to be able to develop novel theory for the estimated correlations, in Section 2.3. In Section 2.4.3, this model is used to carry out an original meta-analysis of the experiments listed in Table 2.1.

2.2.1 A structured covariance model of high-dimensional responses

While the model we detail below is generally applicable for the estimation of a covariance matrix among multiple responses, we specify the data generating process for our goal of estimating the amount of coordination between mRNA transcription and protein translation. In this application, we consider two high-dimensional responses, with approximately 5,300 dimensions, corresponding to mRNA expression and protein abundance in yeast. Each response is measured multiple times in a number of experiments, where each experiment consists of one or more replicates. Let \(X_{i,j}\) de-
note the measurement for mRNA/protein $i$ in replicate $j$. Replicates, experiments and response variables form a three-layer hierarchy of nested groups. Specifically, we have $N_L$ latent variables at the top of the hierarchy (two in this work, representing mRNA and abundance), $N_E$ experiments measuring one of the latent quantities, and $N_R$ total replicates across experiments. To write down the model, we define two functions that map replicates to the other two layers. The function $l[j]$ maps a replicate to the response type (mRNA expression or protein abundance) and the function $k[j]$ maps replicates to experiments. These mappings are such that $k[j_1] = k[j_2]$ implies $l[j_1] = l[j_2]$; that is, replicates of the same experiment measure the same response.

The model has two components: an observation model $p(I_{i,j}|X_{i,j})$, which provides the probability of observing a value for mRNA/protein $i$ in replicate $j$, given the latent mRNA/protein level, and a hierarchical model $p(X_{i,j} \ldots)$ for the latent mRNA/protein levels themselves. We posit

\begin{align}
X_{i,j} &= L_{i,l[j]} G_{k[j]} + E_{i,k[j]} + R_{i,j} + \nu_j \quad (2.1) \\
L_i &\sim \mathcal{N}_{N_L}(0, \Psi) \quad (2.2) \\
E_{i,k} &\sim \mathcal{N}(0, \xi_k) \quad (2.3) \\
R_{i,j} &\sim \mathcal{N}(0, \theta_j) \quad (2.4) \\
p(I_{i,j} = 0|X_{i,j} = x) &= \frac{1}{1 + \exp(-\eta^0_k - \eta^1_k X_{i,j})} \quad (2.5)
\end{align}

where the random variables $L_{i,l}$ specifies the latent mRNA expression and abundance, for mRNA and protein $i = 1, \ldots, N$, and $L_i = [L_{i,1}, \ldots, L_{i,N_L}]'$. The random variables $E_{i,k}$ capture experiment effects for experiment $k = 1, \ldots, N_E$, and $R_{i,j}$ are measurement noise for replicate $j = 1, \ldots, N_R$. Effects between experiments are independent, $\text{Cov}(E_{i_1,k_1}, E_{i_2,k_2}) = 0$ if $k_1 \neq k_2$. Measurement noise is independent between repli-
cates, \( \text{Cov}(R_{i1,j_1}, R_{i2,j_2}) = 0 \) if \( j_1 \neq j_2 \). The parameter \( \nu_j \) reflects replicate specific bias common to all mRNAs/proteins. The coefficient \( G_k \) is an experiment specific scaling factor for the latent expression and abundance. The indicator variable \( I_{i,j} \) denotes whether the value for \( X_{i,j} \) was observed and accounts for non-ignorable missing data as detailed in Section 2.2.1.

The estimand of interest, \( \Psi \), specifies the correlation matrix of the response variables. For our application, \( N_L = 2 \) and \( \psi_{1,2} \) represents the correlation between the true mRNA and protein levels. The diagonal of \( \Psi \) is fixed to one for identifiability. The parameters \( \xi_k \) and \( \theta_j \) specify the variances of the effects for experiment \( k \), and the measurement noise for replicate \( j \), respectively.

To write down the likelihood, let \( \mathbf{X}_i = [X_{i,1}, \ldots, X_{i,N_R}]' \) denote all measurements (both observed and missing) across replicates for mRNA/protein \( i \), and let \( \mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_N] \) denote the \( N \times N_R \) complete data matrix of all measurements. Then, \( \mathbf{X} \sim \mathcal{N}(\nu, \mathbf{I} \otimes \Sigma) \). Here the column covariance, \( \Sigma \) corresponds to the between experiment covariance. Since we assume independence between genes (but see Section 2.4.2), the row covariance is simply the \( N \times N \) identity matrix.

Similarly, define \( \mathbf{I} \) as the binary observation matrix of dimension \( N \times N_R \), and define the vectors \( \eta^0 = [\eta^0_1, \ldots, \eta^0_{N_E}] \), \( \eta^1 = [\eta^1_1, \ldots, \eta^1_{N_E}] \), and \( \nu = [\nu_1, \ldots, \nu_{N_R}] \). Then the complete data likelihood for the proposed model is

\[
\mathcal{L}(\mathbf{X}, \mathbf{I} | \Sigma, \eta^0, \eta^1, \nu) \propto \prod_{i=1}^{N} \left[ |\Sigma|^{-1/2} \exp \left( -\frac{1}{2} (\mathbf{X}_i - \nu)' \Sigma^{-1} (\mathbf{X}_i - \nu) \right) \right] \times \mathcal{L}(\mathbf{I} | \mathbf{X}, \eta^0, \eta^1),
\]

(2.6)
where

\[
\mathcal{L}(\mathbf{I}|\mathbf{X}, \eta_0, \eta_1) = \prod_{i=1}^{N} \prod_{j=1}^{N_R} \left( \frac{1}{1 + \exp(-\eta_{k[j]} - \eta_{q[j]} X_{i,j})} \right)^{I_{i,j}} \left( \frac{\exp(-\eta_{k[j]} - \eta_{q[j]} X_{i,j})}{1 + \exp(-\eta_{k[j]} - \eta_{q[j]} X_{i,j})} \right)^{1-I_{i,j}},
\]

and where \( \Sigma \) is a structured covariance matrix of size \( N_R \times N_R \) detailed in Section 2.2.1.

Note that Table 2.1 lists a few experiments that contain only one replicate. For these experiments we simplify Equation 2.1 by removing the random effect for the replicate, \( R_{i,j} \). This ensures that all the quantities remain identifiable.

**Covariance structure**

The nested response-experiment-replicate grouping leads to a structured covariance matrix, \( \text{Cov} \mathbf{X} = \Sigma \), for the complete data. Assuming that replicates are ordered according to response type and experiment \( (l[j] \text{ and } k[j]) \) in \( \mathbf{X} \), \( \Sigma \) consists of \( N_L \) large blocks corresponding to the response variables; and each large block is a block diagonal plus rank one matrix, with one block for each experiment. Covariance matrices with this structure, illustrated in Figure 2.3, are often referred to as “similarity matrices” e.g., see\footnote{Reference}. In our model, \( \Sigma \) is a function of \( \Psi \), \( \xi_j \), \( \theta_j \) and \( G_k \). The marginal variance of each observation is

\[
\sigma_{i,j}^2 = G_{k[j]}^2 + \xi_{k[j]} + \theta_j.
\]

Two replicates \( j_1 \) and \( j_2 \) within the same experiment \( k = k[j_1] = k[j_2] \) also have \( l = l[j_1] = l[j_2] \) and their covariance is \( G_k^2 + \xi_k \). The replicates are exchangeable within experiments but not between experiments.
OBSERVATION MODEL

Figure 2.2 suggests that the fraction of missing data is negatively correlated with the average observed values for both mRNA expression and protein concentrations. This

Figure 2.3: A) Responses, experiments and replicates form a nested group structure. B) These groups define a “similarity matrix”, a covariance matrix characterized by a block structure for Var $\mathbf{X} = \Sigma$. The $\sigma_j^2$ marginal variances are given by Equation 2.8, $\epsilon_k = G_k^2 + \xi_k$ is the within experiment covariance. $G_{k[j_r]}$ is the scaling factor for the experiment of the replicate corresponding to row $j_r$ of the matrix, $G_{k[j_c]}$ is the same for column $j_c$. 

16
is evidence that the measurements are missing not at random (MNAR)\textsuperscript{89}.

We follow a well established approach to model this type of missing data mechanism, by means of a generalized linear model\textsuperscript{46}. Equation 2.5 models the probability that measurement $X_{i,j}$ is missing, $p(I_{i,j} = 0)$, as a logistic function of the value of the measurement. The parameters of the missing data mechanism, $\eta^0_k$ and $\eta^1_k$, are shared by replicates within an experiment; they uniquely determine the probability that measurements are observed, conditional on $X_{i,j}$.

This observation model is flexible enough to include sharp censoring at a certain mRNA/protein value or to capture very little or no dependence of missingness on mRNA/protein levels. Importantly, the observation model parameters vary by experiments. See Figure A.2 in the Appendix for some examples on how the observation model fits to various experiments.

**Prior specifications**

To complete the model specifications we place priors on $\Psi$, $\xi_k$, $\theta_j$, $\eta^0_k$ and $\eta^1_k$. Recall that referenced works report correlation in the 0.3–0.6 range. In developing an independent meta-analysis, we use either flat, or weakly informative, to produce estimates that are unaffected by previous results that arguably depend on problematic assumptions and methods. For the parameters $\eta^0_k$ and $\eta^1_k$ of the logistic observation model we use a Cauchy prior with mean zero and scale 2.5, after scaling the data (at each imputation step) to have mean zero and standard deviation 1/2, as suggested by Gelman et al.\textsuperscript{35}. We assume flat priors on the scaling factors, $G_k$, and the measurement bias parameters $\nu_j$. For the replicate and experiment variances $\theta_j$ and $\xi_k$ we use independent conjugate scaled inverse $\chi^2$ priors with 3 degrees of freedom and scale 1/5. This is equivalent to an Inv-Gamma(3/2, 3/10) prior.
Since the primary estimand of interest is the correlation matrix $\Psi$, the choice of prior is particularly important. One option is to use the inverse Wishart prior, scaled to have unit variance. The inverse Wishart prior is the standard conjugative prior for covariance matrices, but it is quite restrictive. For instance, the inverse Wishart specifies the same degrees of freedom for every entry in the matrix. Crucially, with the inverse Wishart prior higher variances are associated with higher correlations.

As such, using a scaled inverse Wishart distribution to specify a prior actually corresponds to an informative prior on the correlations. To avoid this, we assume that the correlation and variance are independent. This is consistent with the separation strategy introduced by Barnard et al. (2000). This strategy involves putting a flat prior (Unif$[-1,1]$) on the correlation in the proposed model. The coverage studies of Section 2.4.1 indicate that the estimated correlation is not biased by this choice of prior.

### 2.2.2 Inference via Markov chain Monte Carlo

We fit the hierarchical model and the observation model jointly using a Gibbs sampler. Algorithm 1 provides an overview of the sampling strategy. A more detailed description of the individual steps follows.

Step 1. Since $\mathbf{L}_i$ and $\mathbf{X}_i$ are multivariate normal, $\mathbf{L}_i$ conditional on the other parameters is also multivariate normal. Specifically, \[(\mathbf{L}_i|\mathbf{X}_i, G_k, \xi_k, \theta_j, \nu_j) \sim \mathcal{N}_{N_k}(\operatorname{Cov}(\mathbf{X}_i, \mathbf{L}_i)\Sigma^{-1}(\mathbf{X}_i-\nu), \Psi-\operatorname{Cov}(\mathbf{X}_i, \mathbf{L}_i)\Sigma^{-1}\operatorname{Cov}(\mathbf{X}_i, \mathbf{L}_i)' ), \tag{2.9}\]

and $\operatorname{Cov}(\mathbf{X}_i, \mathbf{L}_i)$ can be easily calculated from Equation 2.1 and the parameters $G_k$, $\xi_k$, $\theta_j$. 

18
MCMC inference via Gibbs sampling

repeat

1. Draw multivariate responses:
   for $i \in 1, \ldots, N$ do
   Draw $L_i$ from a conditional multivariate normal.

2. Draw covariance matrix, conditional on $L$.

3. Draw experiment level random variables:
   for $k \in 1, \ldots, N_E$ do
   Draw $G_k$, $\xi_k$ and $E_{i,k}$ for all $i$ via Bayesian linear regression and normal and Inv-$\chi^2$ draws.

4. Draw replicate level random variables:
   for $j \in 1, \ldots, N_R$ do
   Draw $\nu_j$, $\theta_j$ and $R_{i,j}$ via Bayesian linear regression and normal and Inv-$\chi^2$ draws.

5. Impute missing data, see text.

6. Draw observation model parameters:
   for $k \in 1, \ldots, N_E$ do
   Draw $\eta^1_k$ and $\eta^0_k$ via Bayesian logistic regression and normal draws.

until desired number of samples

Algorithm 1: The Gibbs sampler
Step 2. Given \( \mathbf{L}_i \), we then draw \( \Psi \) using a Metropolis-Hastings random walk sampler. To sample the correlation, we use a truncated normal proposal, centered on the current value. Barnard et al.\(^5\) suggest setting the variance of the proposal distribution to a value inversely proportional to the number of measurements; after tuning, we set it to \( 1/(10N) \). When sampling from a bivariate covariance matrix, the truncation points for the proposal are simply \(-1 \) and \(+1 \), and the general formula is given by Barnard et al.\(^5\).

Step 3. The random effects and the variance parameters are drawn using Bayesian linear regression. First, for each experiment \( k \), we draw \( G_k, \xi_k \) and \( E_{i,k} \). Notice that \( X_{i,j} - R_{i,j} - \nu_j \) is the same for all \( j \) replicates that belong to the same experiment \( k \). So, we regress \( X_{i,j} - R_{i,j} - \nu_j \) on \( L_{i,[j]} G_{k[j]} \) for an arbitrary \( j \) for which \( k[j] = k \) holds and for all \( i \in 1, \ldots, N \). For the conjugate scaled Inv-\( \chi^2 \) prior the posterior of \( \xi_k \) is also scaled Inv-\( \chi^2 \). \( G_k \) is drawn from a normal, see Gelman et al.\(^33\), Sec. 14.2 for details. \( E_{i,k} \) correspond to the residuals of this regression.

Step 4. Similarly, we draw \( \nu_j, \theta_j \) and \( R_{i,j} \) for each replicate \( j \), by regressing \( X_{i,j} \) on \( L_{i,[j]} G_{k[j]} + E_{i,k[j]} \), \( i \in 1, \ldots, N \). \( \nu_j \) are drawn from a normal, \( \theta_j \) are drawn from a scaled Inv-\( \chi^2 \), and the residuals of the regression correspond to \( R_{i,j} \). Again, this is according to the textbooks, see Gelman et al.\(^33\), Sec. 14.2.

Step 5. Given these parameters, we impute the missing data. The conditional density for a missing measurement, \( i \), in replicate \( j \) and experiment \( k[j] \) is proportional to the product of the logistic CDF and a normal density. That is:

\[
p(X_{i,j}^{\text{missing}} | L_{i,[j]}, E_{i,k[j]}, \eta^0_{k[j]}, \eta^1_{k[j]}, \theta_j) \propto \exp \left( - \frac{(X_{i,j} - (L_{i,[j]} + E_{i,k[j]} + \nu_j))^2}{2\theta_j^2} \right) \frac{1}{1 + \exp(-\eta^0_{k[j]} - \eta^1_{k[j]} X_{i,j})} \tag{2.10}
\]
While this density does not correspond to a simple conditional draw, it can be approximated by a normal. For low missingness probabilities, or censoring that occurs far out in the tails, the density is very nearly normal. For more extreme censoring, it is closer to the truncated normal density. Since we do not observe sharp missingness patterns, typically the observed data distribution is close to normal. We use a Metropolis-Hastings independence sampler with a normal proposal centered at the mode of the PDF and variance equal to the Hessian at the mode. We get over 90\% acceptance using this approach.

Step 6. The parameters of the observation model are drawn from a normal, after Bayesian logistic regression on the missing and observed values to get the means and variances.$^{35}$

### 2.3 Theory

In Section 2.2.1, we developed a simple high-dimensional random effects model for the latent measurement, with a missing data mechanism specified through a logistic regression. While standard theory exists that explores identifiability and the effects of noise, structured errors, and non-ignorable missing data on estimates of the regression coefficients of models of this sort e.g., see$^{115}$, to the best of our knowledge, no theory exists that explores the effects on estimates of the correlation. In this Section, we establish a few novel theoretical results in this directions. They provide insights into the results of Section 2.4.

We state mild conditions under which the parameters of our model are expected to be identifiable, in Section 2.3.1. We then demonstrate three ways in which an analysis that disregards key aspects of the data leads to attenuated estimates of the correla-
tion, $\psi_{1,2}$. In Section 2.3.2 we specify, in the context of our model, the known result that noise attenuates correlation. In Section 2.3.3 we go further, proving that it is not enough to simply incorporate noise into the model— if we don’t model the correlation structure of the noise between replicates, we still underestimate correlation. Finally, in Section 2.3.4 we state a condition under which ignoring missing data also coincides with negatively biased estimates of $\psi_{1,2}$. Below, we state and discuss the main results. The proofs are provided in the Appendix.

Ultimately, all three results suggest that any analysis which ignores measurement error, covariance structure, or missing data will typically understate the magnitude of linear dependence between the response variables. Since all of the biases are in the same direction, the errors do not cancel out. These results are consistent with the relatively moderate correlations reported in previous analyses, none of which account for these three features. As such, these theoretical insights further support our finding in Section 2.4.3 that the true correlation between mRNA expression and protein abundance is larger than previously reported.

### 2.3.1 Identifiability

Lee\textsuperscript{58}, Sec. 2.2.2 states the conditions under which Gaussian random effects models (without missing data) are identifiable. For instance, a sufficient condition is that we fix $\text{diag}(\Psi) = 1$. According to this condition, the random effect portion of the model proposed in Section 2.2.1 is identifiable, up to a sign change, for all $L_{i,t[j]}$ and $G_{k[j]}$, since our model contains a single response variable for both mRNA expression and abundance levels.

The situation is more complicated for the observed data model because of the non-ignorable missing data mechanism. Simulation results in Section 2.4.1, obtained with
parameters specified in Table 2.6, show near nominal frequentist coverage of the Bayesian posterior intervals obtained using our MCMC inference strategy. These empirical results suggest that identifiability is not an issue whenever measurements are missing according to Equation 2.5.

2.3.2 Attenuation due to noise

In this Section, we state how the correlation between any two measured responses is smaller in magnitude than the true correlation between the responses, as long as the measurement noise is non-negligible. While this general result has long been established\textsuperscript{102}, we identify the specific parameters in the proposed model which govern the degree of attenuation. Specifically, the amount of attenuation depends on the scaling factors, $G_k$, as well as the replicate and experiment noise, $\theta_k$ and $\xi_k$.

**Theorem 1.** Consider two observed replicates, $X_1, X_2$, from two different experiments, measuring different response variables. For simplicity, let $l[j] = j$ and $k[j] = j$, so that for instance, $X_{i,1} = L_{i,1}G_1 + E_{i,1} + R_{i,1} + \nu_1$. As specified in section 2.2.1, we assume without loss of generality that $\text{Var}(L_i) = \psi_{i,i} = 1$. Given $\xi_k > 0$ and $\theta_k > 0$, for $k = 1, 2$;

$$\text{Cor}(X_1, X_2) = \frac{\psi_{1,2}}{\sqrt{1 + (\xi_1 + \theta_1)/G_1^2\sqrt{1 + (\xi_2 + \theta_2)/G_2^2}}} < \psi_{1,2}$$

holds for $\psi_{1,2} > 0$.

2.3.3 Attenuation due to model mis-specification

In this Section we show that even if we account for noise by incorporating data from multiple experiments, if we do not account for the presence of structured noise within
experiments, we still underestimate correlation. We prove this for a simplified case, where our model parameters are assumed to be homogeneous across responses, experiments and replicates.

We consider a model, \(\mathcal{M}\), of the form as in Equations 2.1–2.4, with two response variables \((N_L = 2)\), two experiments in each response \((N_E = 4)\), and \(n/2\) replicates for each experiment, \((N_R = 2n\) replicates in total):

\[
\mathcal{M} = (\Psi, [G_k], [\xi_k], [\theta_j]), \quad k = 1, \ldots, 4\text{ and } j = 1, \ldots, 2n
\]  

(2.11)

We assume that the parameters are homogeneous across response variables, experiments and replicates: \(\xi = \xi_k \geq 0\) and \(\theta = \theta_j > 0\) for all \(k\) and \(j\), and also assume \(G = G_k\) for all \(k\).

Let \(\tilde{\mathcal{M}}\) be another model, again, of the form of equations 2.1–2.4, but without an experiment specific random effect:

\[
\tilde{\mathcal{M}} = (\tilde{\Psi}, [\tilde{G}_k], [\tilde{\xi}_k = 0], [\tilde{\theta}_j])
\]  

(2.12)

As above, we assume that \(\tilde{\theta} = \tilde{\theta}_j > 0\) and \(\tilde{G} = \tilde{G}_k = 1\) for all \(j\) and \(k\). Aside from having no experiment specific random effect, the model two models are identical. That is, \(\tilde{\mathcal{M}}\) has the same structure, \(\tilde{N}_L = N_L = 2\), \(\tilde{N}_R = N_R = 2n\).

**Theorem 2.** Consider data generated by model \(\mathcal{M}\). Let \(\tilde{\psi}_{1,2}^{PM}\) denote the posterior mean estimator of \(\psi_{1,2}\) under the misspecified model, \(\tilde{\mathcal{M}}\). The posterior mean asymptotically underestimates the true correlation as \(N\), the number of mRNAs and proteins goes to infinity. That is,

\[
\lim_{N \to \infty} \tilde{\psi}_{1,2}^{PM} \leq \psi_{1,2},
\]  

(2.13)
with equality only if $\xi = 0$.

2.3.4 \textbf{Attenuation due to missing data}

In this Section we explore the implications of neglecting to model a non-ignorable missing data mechanism. Since correlation cannot be computed with incomplete pairs of observations, a complete case analysis by definition ignores all mRNAs and proteins for which either value in the pair is missing. We consider a simplified complete case analysis, with a missingness mechanism on only one of the random variables, which induces missingness in the other. The result below states that when the the missingness mechanism generates an observed data distribution which has smaller variance than the complete data distribution, the complete case analysis (on observed pairs) leads to an underestimate of the true correlation.

This condition is generally consistent with the missing data mechanism we posit in Eq 2.5. That is, with a logistic missingness mechanism, the variance of the observed data is smaller than that of the complete data. As such, this result suggests that previous approaches that ignore the missing values for mRNA expression or protein abundance (complete case analyses) generally underestimate the correlation.

\textbf{Theorem 3.} Let $(X,Y)$ be a bivariate normal random variable. Consider a missingness mechanism on $X$ and denote the observed data, ignoring all censored observations, $X^{\text{obs}}$. Further, assume the missingness mechanism is such that $\text{Var}(X^{\text{obs}}) < \text{Var}(X)$. In a complete-case analysis, the missingness mechanism on $X$ also induces a stochastic censoring on $Y$, and only $Y^{\text{obs}}$ is observed. If $\text{Cor}(X,Y) > 0$, then

$$\text{Cor}(X^{\text{obs}}, Y^{\text{obs}}) < \text{Cor}(X,Y).$$

(2.14)
2.4 Results

We evaluate our methodology on synthetic and real data. In Section 2.4.1, we show that the Bayesian confidence intervals have good frequentist coverage, especially for the parameters of interest. In Section 2.4.2, we show that the proposed model is fairly robust to departures from normality of the log-mRNA or log protein abundance levels. We also empirically show that the basic structure of the model is necessary, consistent with theoretical results in Section 2.3. In Section 2.4.3, we present the results of the meta-analysis analysis on the data sets listed in Table 2.1 and compare our results to previous estimates of the correlation between mRNA expression and protein correlation in yeast. In Section 2.4.4, we incorporate technology information into the model, and check the sensitivity of the estimated correlation to different assumptions about the magnitude of technology bias.

2.4.1 Frequentist Coverage

We set out to evaluate frequentist coverage of the Bayesian intervals under realistic simulated data sets. We considered three scenarios for the true correlation, $\psi_{1,2} = 0.5, 0.8,$ and $0.9$. Each scenario consists of 27 simulated experiments, 11 measuring gene expression and 16 measuring protein abundance, each with a number of replicated measurement matching a real data set in Table 2.1, and each measurement with 5,300 dimensions—corresponding to distinct genes and proteins. The remaining parameters ($\eta_k, G_k, \xi_k, \theta_j$, for all $j, k$) were set to the posterior means reported in Table 2.6, which were obtained when fitting the model to the real data, to generate realistic data. Using these parameter values, we then simulated 100 replicated data collections for each correlation scenario.
Table 2.2 reports the frequentist coverage of the 50% and 95% Bayesian posterior intervals for the correlation $\psi_{1.2}$ and the other model parameters. For each of the three correlation scenarios ($\psi_{1.2} = 0.5, 0.8, \text{and } 0.9$), we report the fraction of times the posterior interval covers the true correlation. For $\xi_k$, $\eta_k^0$, $\eta_k^1$ and $G_k$ we report the average coverage, over the $N_E$ experiment specific parameters. For $\nu_j$ and $\theta_j$ we report the coverage averaged over all $N_R$ replicates in the data set. The coverage is excellent for most parameters, especially the main parameter of interest, $\psi_{1.2}$, and the experiment effect variances $\xi_k$.

<table>
<thead>
<tr>
<th>True $\psi_{1.2}$</th>
<th>Confidence</th>
<th>Coverage for parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_{1.2} = 0.5$</td>
<td>50%</td>
<td>43% 51% 39% 49% 40% 56% 54%</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>92% 95% 82% 94% 86% 96% 96%</td>
</tr>
<tr>
<td>$\psi_{1.2} = 0.8$</td>
<td>50%</td>
<td>43% 50% 39% 45% 41% 55% 54%</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>94% 95% 82% 92% 83% 96% 96%</td>
</tr>
<tr>
<td>$\psi_{1.2} = 0.9$</td>
<td>50%</td>
<td>49% 50% 40% 46% 39% 54% 54%</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>98% 95% 82% 93% 84% 96% 97%</td>
</tr>
</tbody>
</table>

*Table 2.2: Frequentist coverage of 50% (top row) and 95% (bottom row) Bayesian credible intervals, for various parameters. Data sets were generated with three true correlation levels: 0.5, 0.8, 0.9.*

2.4.2 Robustness to mis-specification

In this Section, we test the robustness of our model to departures from normality. Since it is not possible to observe the complete data, it is difficult to assess the left tail behavior of the complete data distribution for some data sets. To test how well our model performs for non-normal distributions with skew and heavier tails, we generate the mRNA expression and protein levels, $L_{i,t}$, using the asymmetric Laplace
distribution. A standard multivariate asymmetric Laplace has the representation

\[ Y = mX + X^{1/2}Z, \]  

(2.15)

where \( Z \sim \mathcal{N}_N(0, \Sigma) \) and \( X \) is exponentially distributed with mean one\(^{53}\). The asymmetric Laplace distribution is a continuous mixture of normals with exponentially distributed variance. The parameter \( m \) induces skewness. Figure 2.4 illustrates the univariate and bivariate asymmetric Laplace distributions for various values of the \( m \) skewness parameter.

We ran our algorithm on simulated data at three levels of correlation (0.5, 0.8 and 0.9) and varying skewness in the mRNA expression and protein levels. We again fixed the parameters to match those inferred from the true data (in Table 2.6) but this time generating \( L_i \) from an asymmetric Laplace (Equation 2.15) instead of the bivari-
<table>
<thead>
<tr>
<th>True Correlation</th>
<th>Asymmetric Laplace, skewness ( \mathbf{m} = [m_1, m_2] )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>([0, 0])</td>
</tr>
<tr>
<td>(\psi_{1,2} = 0.5)</td>
<td>0.49</td>
</tr>
<tr>
<td>(\psi_{1,2} = 0.8)</td>
<td>0.79</td>
</tr>
<tr>
<td>(\psi_{1,2} = 0.9)</td>
<td>0.90</td>
</tr>
</tbody>
</table>

**Table 2.3:** Robustness of the model to departures from normality. The table shows inferred posterior mean correlations for data sets with multivariate asymmetric Laplace distributions, with varying correlation and skewness, fit using the normal model, Equation 2.1. Standard deviations are 0.01 or less for all values.

The normal (Equation 2.2). Table 2.3 shows the inferred correlation for data generated using the multivariate asymmetric Laplace. While the model, as expected, gives biased correlation estimates for non-normal data, the bias is very small, even for very skewed and/or peaked data distributions.

Not only is the model robust to misspecification, but also, simpler models fail to give good estimates for at least some of the parameters. We conducted four kinds of experiments on synthetic data, the results of which are summarized in Table 2.4. All experiments we tested three different true \(\psi_{1,2}\) values: 0.5, 0.8 and 0.9, with 10 runs for each of these values. All experiments used 5000 mRNAs and proteins.

1. First we show that modeling noise is important because noise attenuates correlation. Ignoring noise results in a downward bias in the inferred correlation.

See Theorem 1. We generated noisy bivariate normal data with unit variance, one replicate for mRNA and one for protein levels, for 5000 mRNAs/proteins, with true correlations 0.5, 0.8 and 0.9. The noise level was \(\xi + \theta = 0.8\). Then we ignored the noise in our naive inference, i.e. we calculated the observed correlation of the noisy bivariate normal data.
2. Second, we show that ignoring the structure of the noise leads to attenuated correlation estimates. We use 16 mRNA expression and 16 protein replicates equally divided in 4 experiments for both. We generate noisy multivariate Normal data with this structure, with constant noise levels $\xi_{k[j]} = 0.6$ (experiment effects) and $\theta_j = 0.2$ (replicate effects). The $G_{t[j]}$ scaling parameter was one. Then we run the inference procedure by ignoring the experiment random effects, i.e. setting $\xi_k = 0$. See Theorem 2.

3a. Third, if part of the data is non-ignorably missing, then the correlation estimates are attenuated. We use 16 mRNA expression and 16 protein replicates equally divided in 4 experiments for both. We generate noisy multivariate Normal data with this structure, with constant noise levels $\xi_{k[j]} = 0.6$ (experiment effects) and $\theta_j = 0.2$ (replicate effects). The $G_{t[j]}$ scaling parameter was one. The parameters of the observation model were set arbitrarily in a way to get about 1000 completely observed mRNAs and proteins. In the inference we ignore the non-complete cases, and only use the (about 1000) completely observed mRNAs and proteins. See Theorem 3.

3b. Lastly, we show that imputing the missing data, but using a simpler, “missing at random” observation model fails to estimate correlation correctly. The synthetic data contained two experiments for both mRNA expression and protein levels, and two replicates for each experiments. The noise levels were set to $\xi_{k[j]} = 0.6$ and $\theta_j = 0.2$, the parameters of the observation model were tuned to obtain about 1000 completely observed mRNAs and proteins. The missing data was then imputed by fitting the model using a MAR assumption instead of Equation 2.5. We find that using more experiments and/or more replicates
Table 2.4: Features of the data that attenuate correlation: noise, noise structure, missing data and non-randomly missing data. See text for the complete description. Standard deviations are 0.01 or less, unless shown otherwise.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_{1,2}$</td>
<td>$\hat{\psi}_{1,2}$</td>
<td>$\hat{\psi}_{1,2}$</td>
<td>$\hat{\psi}_{1,2}$</td>
<td>$\hat{\psi}_{1,2}$</td>
</tr>
<tr>
<td>0.5</td>
<td>0.32</td>
<td>0.45</td>
<td>0.32 (± 0.07)</td>
<td>0.45 (± 0.02)</td>
</tr>
<tr>
<td>0.8</td>
<td>0.50</td>
<td>0.71</td>
<td>0.65 (± 0.02)</td>
<td>0.77 (± 0.03)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.56</td>
<td>0.80</td>
<td>0.81 (± 0.03)</td>
<td>0.88 (± 0.02)</td>
</tr>
</tbody>
</table>

What about correlation between observations? In this research we assume that the measurements on each gene are independent observations with between replicate covariance, $\Sigma$. We consider correlation between genes in the experiment effects. Certain functionally related genes may in fact vary together across experiments in which the data are actually obtained in some condition which is close to, but not exactly, the one defined. Let $\mathbf{E}$ be the $N \times N_E$ random matrix of experiment specific random effects. We can augment the model to incorporate “between gene” row correlation, $\Delta$, across the experiments:

$$\mathbf{E} \sim \mathcal{N}(\mathbf{0}, \Delta \otimes \boldsymbol{\xi}),$$

where $\boldsymbol{\xi}$ is the diagonal matrix of experiment specific variances. We evaluate the effect of non-identity row correlation, $\Delta$, in simulation. We consider simulations involving three different correlation structures between genes to evaluate how this influences the inference of latent mRNA-protein correlation at three different $\psi_{1,2}$ levels: 0.5,
0.8, and 0.9. In the first two cases we assume that the genes have a block correlation structure and that within blocks the genes are correlated at level 0.9. In two different simulations, we block the genes into 10 groups and 100 blocks of roughly equal size.

In the third simulation we generate data with the gene correlation structure estimated from an independent Yeast data set under multiple conditions\textsuperscript{12}.

Row-wise correlation essentially decreases the effective sample size, leading to over-confidence in the inference\textsuperscript{28}. Table 2.5 shows the 95\% interval of the sampling distribution as well as the coverage of the 95\% credible interval. As expected, there is significant loss of coverage, but the model estimates are essentially unbiased and the error is small. Thus the substantive conclusions on the data in Table 1 are not expected to change much in the presence of row-wise correlation in the noise.

Finally, we test how robust our model is to misspecification of the missingness mechanism. In particular, we assume a rather simple logistic form for the missingness of both mRNA and protein levels. There is evidence of more complicated missing-

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
\multicolumn{1}{|c|}{\text{Parameter}} & \multicolumn{1}{c|}{\text{Value}} & \text{2.5\%} & \text{Mean} & \text{97.5\%} & \text{Coverage} \\
\hline
$\psi = 0.5$, 10 blocks & 0.46 & 0.50 & 0.52 & 0.81 \\
$\psi = 0.5$, 100 blocks & 0.47 & 0.50 & 0.52 & 0.91 \\
$\psi = 0.5$, Brem et al & 0.48 & 0.50 & 0.52 & 0.90 \\
$\psi = 0.8$, 10 blocks & 0.77 & 0.79 & 0.81 & 0.59 \\
$\psi = 0.8$, 100 blocks & 0.79 & 0.80 & 0.81 & 0.89 \\
$\psi = 0.8$, Brem et al & 0.78 & 0.80 & 0.82 & 0.74 \\
$\psi = 0.9$, 10 blocks & 0.86 & 0.89 & 0.91 & 0.28 \\
$\psi = 0.9$, 100 blocks & 0.89 & 0.90 & 0.91 & 0.76 \\
$\psi = 0.9$, Brem et al & 0.88 & 0.90 & 0.91 & 0.60 \\
\hline
\end{tabular}
\caption{The 95\% interval of the sampling distribution of the posterior mean when experiment noise is correlated between genes. The results show that the estimate of $\psi_{1,2}$ is essentially unbiased but the variation increases as the degree of between gene correlation increases. The fourth column shows the coverage of the 95\% credible interval. While there is significant undercoverage, the error is small.}
\end{table}
ness mechanisms, especially in studies using LC MS/MS to measure protein abundance. Here, a two-stage missingness mechanism, capturing both informative and non-informative censoring may be more appropriate\(^5\). To account for a possible misspecification of this type, we generate data assuming that every protein is missing with a 20\% probability, independent of its abundance in addition to the logistic censoring specified in Equation 2.5. The data was generated in a way such that the total fraction of missingness matched the true data. We generate data at three levels of true correlation (0.5, 0.8 and 0.9) and estimate this correlation using the one-level informative missingness model. There is no bias in the estimates, even though the missing data mechanism is slightly misspecified (0.50 ± .02, 0.80 ± .01, 0.9 ± .006). Since the marginal probability of missingness can be well approximated by our two-parameter observation model, the procedure is robust to more complex mechanisms.

2.4.3 Quantifying the transcriptional control of protein production

The main focus of this research is to identify the underlying true correlation between mRNA expression and protein abundance in exponentially growing yeast at steady state. Thus, we fit our model on the data listed in Table 2.1. When fitting the model on this data, we initialize our chains using standard software\(^8\) to find the EM solution, assuming data missing at random\(^4\), and use this as a starting point for our Gibbs sampler. To save disk space we save every 50th sample, and use over 5000 samples to generate posterior estimates. We checked the convergence of the MCMC simulation for the \(\psi_{1,2}\) samples, using two MCMC chains and the \(\hat{R}\) statistics of Gelman & Rubin\(^3\). In our real data fits, \(\hat{R}\) was close to 1 (less than 1.01), indicating very good convergence. The effective sample size for the inferred correlation, \(\psi_{1,2}\), is 1427. The
average effective sample size for the experiment noise within mRNA expression experiments is 3368 and for protein expression experiments is 1609.

After accounting for the measurement structure, biological and technical noise, and missing data, we estimate the true posterior mean correlation to be 0.82 (± 0.01). This estimate is significantly larger than almost all previous estimates\textsuperscript{40,47} or estimates derived from naive complete-case analyses between single measurements (Figure 2.5A).

Some of our data sets have a very large number of measurements missing. To check that including them does not bias our results, we also fitted the model with excluding experiments with (1) more than 80% and (2) more than 60% of missingness. In both cases the inferred $\psi_{1,2}$ value was 0.83 ± 0.01, comparable to the result obtained on the full data set, 0.82 ± 0.01, in fact slightly higher.

These results have implications for our understanding of the role of post-transcriptional regulation in yeast at steady-state. In particular, they suggest that this type of regulation is not as pervasive as previously thought. Additionally, the data and our results suggest that, using the current technologies, yeast mRNA expression levels are not much worse for predicting protein abundance values in a given experiment than another protein abundance measurement from another lab. This is important because measuring mRNA expression levels is simpler and cheaper than measuring protein abundances. Thus, mRNA levels may in fact be a reasonable proxy for protein abundance, at least in steady state. A list of all experiment specific parameters are given in Table 2.6. The parameters, $\eta$, reflect the inferred missingness pattern by experiment, and the noise parameters $\xi$ and $\theta$ reflect how much each experiment and replicate deviate from the inferred true gene expression or protein levels.
Figure 2.5:  A) mRNA expression–protein correlation estimates. The lines with filled and empty circles show all naive pairwise correlation estimates, using mRNAs measured in both data sets only, and the posterior distribution for the correlation, inferred via our structured covariance model, respectively. Dashed vertical lines correspond to mean values. The correlation of the (naive) average protein and mRNA expression levels over measurements is also shown. B) Posterior distribution of the correlation using the technology extension to our model, and discrete technology variance priors on all combinations over the weights 1, 2 and 5. The vertical lines show three weight configurations: the ones with the smallest and largest mean inferred correlation and the equal weighting (all weights equal to 1). C) Posterior distributions of mRNA expression–protein correlations, conditional on exactly one up weighted technology ($W = [1,1,1,1,5]$).
<table>
<thead>
<tr>
<th>ID</th>
<th>$\eta_0^k$</th>
<th>$\eta_1^k$</th>
<th>$G_k$</th>
<th>$\xi_k$</th>
<th>$\theta_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAUS</td>
<td>-3.09</td>
<td>12.41</td>
<td>1.4</td>
<td>0.4</td>
<td>(0.09, 0.13, 0.58, 0.05, 0.03)</td>
</tr>
<tr>
<td>DUD</td>
<td>-3.96</td>
<td>-13.55</td>
<td>0.84</td>
<td>0.85</td>
<td>(0.53, 0.21, 0.29, 0.37)</td>
</tr>
<tr>
<td>GARC</td>
<td>-1.08</td>
<td>1.05</td>
<td>1.03</td>
<td>-</td>
<td>0.87</td>
</tr>
<tr>
<td>HOLS</td>
<td>-1.6</td>
<td>-3.92</td>
<td>1.46</td>
<td>-</td>
<td>0.27</td>
</tr>
<tr>
<td>ING1</td>
<td>-0.52</td>
<td>-0.45</td>
<td>1.41</td>
<td>0.46</td>
<td>(0.04, 0.02)</td>
</tr>
<tr>
<td>ING2</td>
<td>-1.16</td>
<td>0.32</td>
<td>1.43</td>
<td>0.47</td>
<td>(0.05, 0.01)</td>
</tr>
<tr>
<td>ING3</td>
<td>-1.38</td>
<td>1.18</td>
<td>1.56</td>
<td>0.33</td>
<td>(0.04, 0.02)</td>
</tr>
<tr>
<td>LIPS1</td>
<td>-2.34</td>
<td>-4.58</td>
<td>1.35</td>
<td>0.79</td>
<td>(0.01, 0.01, 0.01, 0.01, 0.01, 0.01)</td>
</tr>
<tr>
<td>LIPS2</td>
<td>-0.65</td>
<td>0.42</td>
<td>1.13</td>
<td>-</td>
<td>0.58</td>
</tr>
<tr>
<td>MAC</td>
<td>-0.2</td>
<td>-0.48</td>
<td>1.22</td>
<td>-</td>
<td>2.18</td>
</tr>
<tr>
<td>MIUR</td>
<td>-1.27</td>
<td>-2.53</td>
<td>1.21</td>
<td>3.81</td>
<td>(0.25, 0.01, 0.03, 0.06)</td>
</tr>
<tr>
<td>NGAL</td>
<td>-0.8</td>
<td>0.87</td>
<td>1.3</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>PELE</td>
<td>-1.61</td>
<td>-2.23</td>
<td>0.9</td>
<td>-</td>
<td>0.76</td>
</tr>
<tr>
<td>ROTH</td>
<td>-3.46</td>
<td>-8.11</td>
<td>1.56</td>
<td>1.06</td>
<td>(0.22, 0.03)</td>
</tr>
<tr>
<td>VELC</td>
<td>-20.04</td>
<td>-2.43</td>
<td>0.94</td>
<td>-</td>
<td>1.16</td>
</tr>
<tr>
<td>YASS</td>
<td>-1.24</td>
<td>1.69</td>
<td>1.36</td>
<td>0.37</td>
<td>0.06</td>
</tr>
<tr>
<td>FUTR</td>
<td>-2.43</td>
<td>7.75</td>
<td>4.62</td>
<td>-</td>
<td>6.8</td>
</tr>
<tr>
<td>GHAM</td>
<td>-1.02</td>
<td>6.06</td>
<td>1.68</td>
<td>-</td>
<td>1.41</td>
</tr>
<tr>
<td>GODO</td>
<td>-0.74</td>
<td>11.25</td>
<td>3.52</td>
<td>-</td>
<td>1.86</td>
</tr>
<tr>
<td>GYGI</td>
<td>-2.33</td>
<td>4.68</td>
<td>3.53</td>
<td>-</td>
<td>5.49</td>
</tr>
<tr>
<td>LEE</td>
<td>-1.23</td>
<td>8.72</td>
<td>4.1</td>
<td>1.68</td>
<td>(0.81, 0.79, 0.94)</td>
</tr>
<tr>
<td>LU</td>
<td>-0.59</td>
<td>5.82</td>
<td>1.35</td>
<td>-</td>
<td>3.01</td>
</tr>
<tr>
<td>NGAR</td>
<td>-2.04</td>
<td>21.32</td>
<td>3.53</td>
<td>-</td>
<td>(0.16, 0.26, 0.13, 0.21, 0.17, 0.21)</td>
</tr>
<tr>
<td>NEWM</td>
<td>-5.25</td>
<td>21.12</td>
<td>1.93</td>
<td>-</td>
<td>2.03</td>
</tr>
<tr>
<td>PENG</td>
<td>-2.26</td>
<td>-13.8</td>
<td>2.1</td>
<td>-</td>
<td>1.3</td>
</tr>
<tr>
<td>THAK</td>
<td>-1.44</td>
<td>7.51</td>
<td>5.72</td>
<td>4.99</td>
<td>(0.6, 0.3, 0.33)</td>
</tr>
<tr>
<td>WASH</td>
<td>-3.44</td>
<td>-23.24</td>
<td>2.82</td>
<td>-</td>
<td>6.57</td>
</tr>
</tbody>
</table>

**Table 2.6:** List of inferred posterior means for mRNA expression (above the midline) and protein concentration (below the midline) for every experiment. The total variance can be found using Equation 2.8. These parameter values are also used to generate the simulated data in Section 2.4.1 and 2.4.2. Refer to Table 2.1 for details about the individual data sets.

### 2.4.4 Assessing the impact of different measurement technologies

In our initial analysis, we assumed that for both mRNA expression and protein levels, all of the experiment level variables, $E_{i,k}$, are exchangeable. However, in reality
there is further distinguishing information, namely, the technology that is used by each lab. In the literature, in addition to lab level effects, there is evidence of different systematic biases in the technologies\textsuperscript{116,84,121}. Incorporating these effects implies that experiments are only exchangeable if they are conducted using the same technology.

By introducing technology specific variables into the model, we can assess how sensitive the estimate of $\boldsymbol{\Psi}$ is to a model incorporating technology. For this analysis, we assume that each technology, $t$, has some bias, $T_t$, which is normally distributed with a technology specific variance. However, as noted by Larsson et al.\textsuperscript{56}, the extent of technology specific bias and variation is not completely understood. As such, in our model, the technology specific biases have unknown variance terms that are impossible to infer without external data or prior knowledge. Thus, we perform a sensitivity analysis to check how our inferred correlation changes with different assumptions about this bias. We amend our model to incorporate technology information as follows:

\begin{align*}
X_{i,j} &= T_{i,t[j]}G_{k[j]} + E_{i,k[j]} + R_{i,j} + \nu_{j,t[j]} \\
T_{i,t[j]} &\sim \mathcal{N}(L_{i,t[j]}, \tau_{t[j]}/W_{t[j]}),
\end{align*}

with the rest of the model as defined in equations 2.2–2.5. Here, $t[j]$ indexes a particular technology used for measuring replicate $j$. Technologies and experiments form nested groups. All replicates in a given experiment were performed using the same technology. Each technology is only used to measure either mRNA or protein levels. As before, $E_{i,k}$ and $R_{i,j}$ represent experiment and replicate specific effects.

$W_{t}$ is a technology specific weight which can be fixed a priori or drawn from a distribution. The measured data alone cannot inform us about which technologies give
more biased estimates. Accordingly, we fit our model, in separate runs, using different pre-chosen sets of weights, \( \mathbf{W} \), to explore the sensitivity of our results to possible biases in technology.

We consider three technologies for measuring mRNA expression (custom microarray, commercial microarray and RNA-Seq) and two technologies for measuring protein abundance (two-dimensional gel electrophoresis and mass spectrometry). For each technology, we assume \( W_{t[j]} \) is iid uniform over the set \{1, 2, 5\}. The values 1, 2, and 5 are arbitrary but representative of possible moderate and large technology specific biases. Under this assumption, the heavily weighted technology \( (W_t = 5) \) has bias with average magnitude that are \( \sqrt{5} \) times smaller than the technologies assigned weight 1. Figure 2.5B shows the posterior mean correlation mixed over all combinations of weights. The mean correlation is slightly larger and more variable, but the qualitative results are qualitatively similar to those presented in Section 2.4.3. Figure 2.5C shows five conditional posteriors each with exactly one technology assigned weight 5 and the rest assigned weight 1.

Interestingly, the results in Figure 2.5C are nearly identical between protein abundance technologies, suggesting that mass spectrometry and the 2D gel technique imply biases of similar magnitude on \( \hat{\Psi} \). The results are more variable for the mRNA expression technologies. Weighting our estimate toward RNA-Seq yields the lowest correlation estimate (0.80) while weighting the estimate toward custom microarray yields a higher estimate (0.85). Crucially, when all technologies are given equal weight, the posterior mean correlation is close to the highest, at 0.86. Consistent with previous studies\(^66\), this suggests that by combining data from experiments involving diverse technologies, we may in fact get better estimates than any one technology could give us on its own.
2.5 Discussion

We have presented an original meta-analysis of high-throughput biological data sets to quantify the coordination between transcription and translation, in yeast grown exponentially at steady state. Operationally, we have developed a hierarchical random effects model for log-transformed mRNA expression levels and protein concentrations, which includes a non-ignorable missing data mechanism. The correlation between latent representations of these two high-dimensional responses is the estimand of interest in our meta-analysis. This estimand is traditionally regarded as a nuisance parameter e.g.,\textsuperscript{115}, thus we develop theory to assess the effects of noise, structured measurements, and non-ignorable missing data on the estimates, in Section 2.3.

We defined the correlation between latent mRNA and protein levels as the estimand of interest, to quantify the notion of coordination between transcription and translation. Our study is necessarily restricted to a single state of a simple organism, and has no direct implications for post-translational regulation in other settings, dynamically changing environments, other organisms, or regulation that cannot be measured by correlation (e.g. amplification of effects). Alternative notions of coordination are possible, however, some more justifiable than others. The correlation between observable measurements is a poor choice, for instance. More sophisticated approaches could consider a notion of an underlying biologic signal, quantified by means of categorical, or even simply binary, signal\textsuperscript{78}. In the context of such approaches, it would then be natural define the correlation of these categorical, or binary, random variables as the estimand of interest.

Further evidence that illustrates the relevance and timeliness of estimates about the scalar estimand of interest here is given by a recent paper that targets the same esti-
mand, in human. In this paper, the authors report an estimate for the correlation between mRNA and protein levels of about 0.8, which is close to the estimate we report, but slightly lower, as can be expected given the complexity of a study in human.

Identifiability of random effects models is an outstanding issue and needs to be evaluated on a case-by-case basis. As detailed in Section 2.3.1, our model meets sufficient conditions for identifiability for the parameters $L_{i,j}$ and $G_{k,l}$ for all combinations of the indices $i,j,k,l$ e.g., Sec. 2.2.2, but the non-ignorable missing data mechanism complicates the situation beyond the reach of available theory. However, the frequentist coverage results in Section 2.4.1 suggest that all the key parameters are identifiable. While these results were obtained on simulated data sets, the design of experiments matched closely the properties of the data collected for the meta-analysis, and parameter values were set to the estimated values obtained on the real data, thus adding confidence to the empirical identification.

We choose not to include information on estimates of the correlation between mRNA and protein levels reported in previous studies, including those whose data we included in the meta-analysis presented in Sections 2.4.3 and 2.4.4. This choice is motivated by the questionable statistical choices previous results depend on, including the use of complete cases only in the presence of non-ignorable missing data (caused by the measurement protocols implemented in the various technologies), the lack of modeling assumptions about important sources of variation in the data, or the lack of a model altogether. By not including previously reported correlations, we aimed at producing an independent analysis, based on a simple model that can be expected to produce robust estimates.

The exploratory data analysis summarized in Figure 2.2 suggests that the amount of missing data is inversely proportional to mRNA expression and protein concen-
tration. This is expected, since even modern high-throughput technology find it difficult to complete the measurement protocols successfully for rare transcripts and proteins\textsuperscript{114,100}. For convenience, we fully specified the non-ignorable missing data mechanism by means of a logistic regression, a well established approach e.g., see\textsuperscript{90,46}. Inference results were not sensitive to two alternative specifications of the (MNAR) missing data mechanism we considered; probit and log-log.

The assumption of normality of the log-transformed measurements of mRNA expression and protein concentration is another choice of convenience. We intended to carry out the meta-analysis with a model that included all the important sources of variation in the data, while simple enough to allow for some theoretical results on the correlation estimates. The multivariate normal distribution was an obvious choice. Exploratory data analysis suggested that log-transformed data are approximately normal. Goodness-of-fit evaluation by means of posterior predictive checks confirmed that the models in Sections 2.4.3 and 2.4.4 fit the data well. The simulation studies based on the multivariate asymmetric Laplace distribution for log-transformed data presented in Section 2.4.2, add further confidence that estimates of the correlation between mRNA and protein levels are robust to model mis-specifications.

2.5.1 Substantive Conclusions

The main result of our meta-analysis is that the correlation between mRNA and protein levels, when estimated with a reasonable model, is much higher than previously reported. Our analyses indicate that a more accurate estimate of such correlation is between 0.82 and 0.86, depending on which model variant is used, the most conservative estimate being 0.82±0.01. The proportion of variance explained is expected to increase if one were to remove some of the within experiment variation by design. This
could be accomplished, for instance, by using the same sample for both mRNA and protein quantification, by preparing the sample under conditions that are demonstrably steady-state and not altered by a transient stress response, or by using measurement technology with improved precision and accuracy. While our study is restricted to a simple organism and a well-defined condition, the analysis indicates that there has been widespread overestimation of the role of post-transcriptional regulation in these conditions\textsuperscript{40,47}, and that suggests that other dominant modes of regulation are not waiting to be discovered.

Interestingly, the sensitivity analysis that incorporates technology information into the model suggests that the highest estimated correlation is obtained when we assume a bias of equal magnitude across technologies. This result is consistent with previous work that suggest improved estimates can be achieved by averaging across technologies\textsuperscript{66}. While there is debate about the best high-throughput technology, this result suggests that consolidating data from different sources, under the assumption that all technologies are equally good, balances out the biases from any individual approach. In other words, new technology is not necessarily better, than older but more mature technology.

Technology alone, however, does not explain all of the variability between different experiments. We hypothesize that much of the between experiment variability is due to disparity in growth rates at time of harvest. Even though the studies in our data collection claim to analyze samples from exponentially growing yeast, it is plausible that the growth rates differ due to experimental protocols. As evidence of this, preliminary results suggest that the scaling factors, $G_{k[j]}$, are highly correlated with independent estimates of growth rate\textsuperscript{2}. We further explore this hypothesis elsewhere\textsuperscript{20}.

Ultimately, our meta-analysis analysis highlights the dangers of casually using
correlations between observables to estimate the strength of the coordination between processes in the cell. We have shown that noise, missing data and model misspecification can lead to spurious conclusions, in theory, and they actually do in practice.
3

Quantification of Post-transcriptional regulation across human tissues

3.1 Introduction

The relative ease of measuring mRNA levels has facilitated numerous investigations of how cells regulate their gene expression across different pathological and physi-
ological conditions.\textsuperscript{101,97,103,98,94,26} However, often the relevant biological processes depend on protein levels and mRNA level are merely their proxies.\textsuperscript{3} If a gene is regulated mostly transcriptionally, its mRNA level is a good proxy of its protein level. Conversely, post-transcriptional regulation (PTR) can set protein levels independently from mRNA levels, as in the cases of classical regulators of development,\textsuperscript{54} cell division,\textsuperscript{43,82} and metabolism.\textsuperscript{22,96} Thus understanding the relative contributions of transcriptional and post-transcriptional regulation is essential for understanding their trade-offs and principles of biological regulation, as well as for assessing the feasibility of using mRNA levels as proxies for protein levels.

Some studies of these relative contributions have concluded that protein levels depend mostly on the mRNA levels\textsuperscript{49,19} while other studies have concluded the opposite, i.e., that protein levels depend mostly on PTR.\textsuperscript{40,99,93} These conflicting views arise because of differences in the systems, methods, and the quantified protein variance. In particular, correlations between absolute levels of mRNA and protein conflate many sources of variation, including variation between the levels of different proteins, variation within the same protein across different conditions, and the variation due to measurement error and technological bias.

However, these different variances have very different biological interpretations and implications. A major source of the variability in measured protein and mRNA levels is due to differences between the mean (average across tissue-types and physiological conditions) levels of proteins, which we refer to as \textit{mean-level variability}. This mean-level variability reflects the fact that some proteins, like ribosomal proteins, are highly abundant across all profiled conditions while other proteins are \texttimes 10^8 times less abundant across all profiled conditions. Another principal source of variation in protein levels, orthogonal to mean-level variability, is the variability within a protein across
different physiological conditions or cell-types. This variation reflects normal physiological regulation, which we refer to as physiological variability, and is usually smaller in magnitude. However, the physiological variability is frequently the most relevant source of variation for understanding different physiological phenotypes across cells types and physiological conditions.

We separately quantify the contributions of transcription and PTR to the mean-level variability and to the physiological variability. Our results indicate that the physiological variability across human tissues is dominated by post-transcriptional regulation, while the mean-level variability is dominated by transcriptional regulation. These results reconcile existing estimates in the literature and highlight the dominance of PTR in determining the variation in the levels of a protein across conditions. Furthermore, we suggest a simple and general approach for deconvolving the contributions of transcriptional and post-transcriptional regulation to measured protein levels.

3.2 Results

3.2.1 Conflation of Orthogonal Sources of Variability

A common approach to analyzing the respective roles of transcriptional and post-transcriptional regulation of protein abundances relies on directly correlating estimates of absolute levels of RNA and protein, often on a logarithmic scale.\textsuperscript{40,19,93} If the ratio between protein and mRNA levels did not vary by gene, then absolute levels of mRNA and protein would be perfectly correlated. This protein/mRNA ratio has been referred to as a gene’s “translational efficiency” because it reflects, in large part, the translational rate of the gene. Since this ratio also reflects other layers of regulation, such as protein degradation, and hence we will refer to it as post-transcriptional
regulation (PTR). Wilhelm et al.\textsuperscript{119} suggested that the protein level of a gene in a particular tissue can be predicted by scaling the corresponding mRNA level by the median PTR ratio (across tissues) for that gene. In the absence of tissue-specific PTR, the gene-specific PTR ratios will be constant across tissue-types, and thus simple scaling of mRNA by the PTR ratio – we refer to it as “scaled mRNA” – will accurately predict tissue-specific protein levels.

To investigate this suggestion, we compute the correlation between scaled mRNA and measured protein and its squared value ($R_T^2$), which quantifies the fraction of the total protein variance explained by mRNA levels between genes. As previously observed\textsuperscript{93,119}, the variance explained between scaled mRNA and measured protein is in fact large ($R_T^2 = 0.77$, across 6104 measured proteins, Figure 3.1a). This fraction, however, conflates the the mean-level variance and the physiological variance of protein levels. Figure 3.1b shows this conflation graphically for a subset of 100 genes measured across 12 tissues. The physiological variance is captured by the variation \textit{within} the regression fits and the mean-level variance is captured by the variation \textit{between} the regression fits. Such conflation of variation, where different subgroups of the data show different trends, is known as Simpson’s (amalgamation) paradox and can lead to counter-intuitive results\textsuperscript{10}. To illustrate Simpson’s paradox, we specifically chose a subset of genes for which the scaled mRNA and measured protein were negatively correlated across tissues, but for which the mean-level variance spanned the full dynamic range. On this subset of genes, the between protein correlation on this subset is still large and positive, despite the fact that all within gene trends are negative. Over all genes, the empirical correlations between scaled mRNA and protein for a single protein across tissues, $R_P$, span a wide range (Appendix, Figure B.1). As this makes clear, $R_T^2$ is not in general, informative about the variance explained by scaled
Figure 3.1: The fraction of the total variance of protein levels explained by scaled mRNA levels is not informative about the variance explained by scaled mRNA on protein’s levels across tissue-types or conditions. (a) Scaled mRNA correlate strongly with measured protein levels ($R^2_B = 0.77$ over 6104 measured proteins in each of 12 different tissues), (b) subset of 100 genes are used to illustrate an example Simpson’s paradox: regression lines reflect within gene physiological variation across each tissue. Despite the fact that the overall correlation between scaled mRNA and measured protein levels, $R_T$, is positive and large, for any single gene in this set, scaled mRNA is negatively correlated with measured protein levels ($R_P < 0$).

mRNA on protein levels within gene ($R^2_P$).

Importantly, the biologically relevant quantity is usually the change in protein abundance between different tissues or physiological conditions. These changes happen on a dynamic range of about $2 - 10$ fold variation which is dwarfed by the $10^9$ fold dynamic range of abundances across different proteins. Thus, even 10 fold errors in the estimates of protein levels are consistent with very large $R^2_T$. To further demonstrate the implications of this vast difference in the dynamic ranges, we generate data from a simple model and tune the between-tissue variation of the PTR ratio. We use mRNA levels and estimated median PTR ratios from Wilhelm et al.¹¹⁹
Figure 3.2: Simulated data demonstrating change in $R_T^2$ and $R_P^2$ as a function of the across-tissue variance in the PTR ratio ($\tau^2$). As $\tau^2$ increases (corresponding to more between tissue PTR), the fraction of the total protein variance explained by scaled mRNA decreases only slightly (red), while the fraction of the physiological variance that can be explained by mRNA levels plummets sharply (black, dashed). The grey region represents the approximate range for the Wilhelm et al.\textsuperscript{119} data. Across-tissue variation in PTR explains between 16 and 20 percent of the total protein variance (Section 3.3)

to simulate plausible protein levels from a log-normal distribution (see Methods).

Since the dynamic range between proteins is large, a small change in the variation in the across-condition PTR ratios has negligible impact on $R_T^2$. However, the dynamic range within gene is significantly smaller, and as such, even for moderate variation in the PTR ratio across conditions has a dramatic impact on the within-gene explained variance, $R_P^2$ (Figure 3.1c). As we demonstrate below, this simple model is consistent with what we observe in real data: mRNA and estimated median PTR ratios together explain a large fraction of variance in mean protein levels because proteins span a large dynamic range. Within gene, small variations in the PTR ratio accounts for a large fraction of the variance in that protein’s abundance.
3.2.2 Variance Conflation in Practice

In Section 3.2.1, we illustrate the statistical problems with using the fraction of the total protein variance explained by mRNA levels ($R_{2B}^2$) as an indication about the extent to which mRNA changes contribute to protein changes across tissues (e.g. $R_{2W}^2$). Still, it remains to be seen how big of a problem this is in practice. First, we again demonstrate that in three different tissues, $R_{2B}^2$ is high (Figure 3.3, top). However, if each gene is indeed associated with a single PTR ratio, this quantity has no effect on the relationship between mRNA fold-change and protein fold-change. Specifically, mRNA and protein ratios should be perfectly correlated (up to measurement noise) if there is no variation due to post-transcriptional regulation (Methods, Section 3.3).

We explore how these fold-change correlations vary in pair-wise comparisons between three human tissues (Figure 3.3, bottom). These comparisons show that in fact, the fraction of variance explained in protein-fold changes by mRNA fold change is usually small and depends on the pair of tissues. For instance, the mRNA fold-changes between the uterus and prostate have essentially no predictive power for protein fold changes in those same tissues. Nonetheless, other tissues show a moderate fold-change correlation (e.g. prostate vs. kidney and uterus vs. kidney). Across the 3 examples, although the fraction of the variance in protein fold-changes that can be explained by mRNA fold-changes varies significantly, $R_{2B}^2$ remains very high in each case, dominated by the mean-level variance. This result underscores the general problem of variance conflation in the context of gene regulation on real data.

In Figure 3.4a we extend our results on fold-change correlations from Figure 3.3 to all pairwise combinations of tissue-types. The range of correlations indicate that for some tissue-type pairs, physiological variation in mRNA explains a significant fraction
Figure 3.3: mRNA fold-changes are weakly correlated with protein fold-changes in some tissue-types but uncorrelated in others. While scaled mRNA is highly predictive of the absolute protein levels (top row), the accuracy of these predictions does not generally reflect the accuracy of protein fold-changes predicted from mRNA fold-changes (bottom row).
of the physiological protein variance (as much 58%) but for other tissue-type pairs it does not. This suggests that the physiological variation of many genes is regulated primarily post-transcriptionally in at least some tissues.

Still, it is is possible that the physiological variation of a subset of genes is regulated primarily transcriptionally and thus for these genes protein fold-changes can be predicted reliably from mRNA fold-changes. As such, we quantify the error in predicting protein fold-changes from mRNA fold-changes, again assuming that the PTR ratio of each gene is constant across all 12 tissue types. The cumulative distribution of errors (Figure 3.4b) indicates that the protein fold-changes for less that 1000 genes can be estimated from mRNA fold-changes with less than 100% error. For over 30% of proteins, estimating protein levels using a single gene-specific PTR ratio results in over 1000% error.

3.2.3 Post-transcriptional regulation in functional gene sets

The lack of correlation between protein and mRNA fold-changes can reflect large measurement noise rather than PTR\textsuperscript{30}. Because the physiological variation in the protein levels is relatively small, protein levels are measured at relatively low precision, and missing data is abundant, it is difficult to accurately quantify the true PTR variability for a single gene. However, we can gain some power to detect significant variability by pooling information from many related genes. As such, we investigate whether the PTR variability across tissue-types reflects systematic regulatory trends within sets of functionally related genes.

For this analysis, we define the “relative PTR” (rPTR) of a gene in a given tissue to be the PTR in that tissue divided by the median PTR of the gene across the other 11 tissues. We evaluated the significance of rPTR variation for a gene-set in
Figure 3.4: The physiological variation of protein levels can be predicted from the corresponding mRNA variation only for a small subset of genes (a) As in Figure 3.3, we compute all pairwise correlations between the fold-changes in mRNA and Protein for all 12 tissue-types. For some tissue pairs, mRNA fold change is predictive of protein fold-change, but for others it is not. (b) The cumulative distribution of the maximum fold error that results from estimating protein levels from scaled mRNA levels. The error in quantifying protein fold change from mRNA exceeds 10-fold for 1862 proteins (31%, red) and exceeds twofold for 5106 proteins (85%, blue); the fold-changes of less than 1000 proteins (%) can be estimated from mRNA fold-changes with less than twofold (100%) error.
each tissue-type by comparing the corresponding gene-set rPTR distribution to the rPTR distribution for those same genes pooled across the other tissues (Figure B.2); we use the KS-test to detect differences in the rPTR distributions; see Methods. Our results indicate that the genes from many GO terms\textsuperscript{17} have much higher rPTR in some tissues than in others. For example the ribosomal proteins of the small subunit (40S) have high rPTR in kidney but low rPTR in stomach (Figure 3.5a-b).

Some of these trends can account for fundamental physiological differences between tissue types— the kidney is by far the most metabolically active (energy consuming) tissue among the 12 profiled tissues\textsuperscript{41} and it has very high rPTR for many gene sets involved in energy production (Figure 3.5a). In this case, post-transcriptional regulation plays a functional role in meeting the high energy demands of kidneys. Moreover, the fact that we observe a highly significant (posterior error probability < $10^{-10}$) mode of translational regulation (such as increased TF for mitochondrial genes and decreased TF for focal adhesion in kidney) indicates that at least some of the variation in translational efficiency and protein degradation across tissue-types reflects regulatory activity rather than measurement noise.

3.3 Methods

Data and Scaled mRNA

In this chapter, we use data from Wilhelm et al.\textsuperscript{119} consisting of mRNA and protein measurements for $N = 6104$ genes measured in each of twelve different human tissues: adrenal gland, esophagus, kidney, ovary, pancreas, prostate, salivary gland, spleen, stomach, testis, thyroid gland, and uterus. For these genes, about 8% of the mRNA measurements and about 40% of the protein measurements are missing.
Figure 3.5: Concerted variability in the PTR of functional gene-sets across tissue-types (a) mRNAs coding for the small ribosomal subunit, NADH dehydrogenase and respiratory proteins are translated much more efficiently in kidney as compared to the median across the other 11 tissues (FDR < 2%). In contrast mRNAs coding for focal adhesion are translated less efficiently (FDR < 2%). (b) The stomach also shows very significant rPTR variation, with low rPTR for the small ribosomal subunit and high rPTR for tRNA-aminoacylation (FDR < 2%).
First, denote \( m_{ij} \) the log mRNA levels for gene \( i \) in condition \( j \). Similarly, let \( p_{ij} \) denote the corresponding log protein levels. First, we normalize the columns of the data, for both protein and mRNA, to different amounts of total protein per sample. Any multiplicative factors on the raw scale correspond to additive constants on the log scale. Consequently, we normalize data from each tissue-type by minimizing the sum of squared differences between data from that tissue and the first tissue (chosen to serve as a baseline). Specifically, for all proteins and conditions \( j > 1 \), we normalize each measurement by setting

\[
p_{ij}^* \leftarrow p_{ij}^u - \frac{1}{N} \sum_i (p_{i1}^u - p_{ij}^u)
\]

Where \( p_{ij}^* \) and \( p_{ij}^u \) represent the normalized and non-normalized protein measurements respectively. We conduct the same normalization for mRNA. This normalization corrects for any multiplicative differences in the raw mRNA or protein.

After normalization, we define \( r_{ij} = p_{ij} - m_{ij} \) as the log PTR ratio of gene \( i \) in condition \( j \). If the post-transcriptional regulation the \( i^{th} \) gene were not tissue-specific, then the \( i^{th} \) PTR ratio would be independent of tissue-type and can be estimated as

\[
\hat{T}_i = \text{median}(p_{ij} - m_{ij})
\]

Then the log “scaled mRNA” (or mean protein level) can be defined as

\[
p_{ij} = m_{ij} + T_i
\]

On the raw scale this amounts to scaling each mRNA by its median PTR ratio and represents and estimate of the mean protein level. The residual difference between the
log mean protein level and the measured log protein level

\[ r_{ij} = p_{ij} - \overline{p}_{ij} \]

consists of both condition-specific PTR and measurement noise.

**Simulated Protein Levels**

We sought to evaluate the role of variance conflation in the context of gene regulation based on a simple model:

\[
\begin{align*}
\overline{p}_{ij} &= m_{ij} + T_i \\
p_{ij} &= \overline{p}_{ij} + r_{ij} \\
r_{ij} &\sim N(0, \tau^2)
\end{align*}
\]

\( r_{ij} \) corresponds to the log condition-specific deviation from the mean protein level. Here, \( \tau^2 \) controls the variation in PTR ratios across tissues (e.g. the amount of post-transcriptional regulation). When \( \tau^2 = 0 \), there is no variation in post-transcriptional regulation across tissue-types \( (r_{ij} = 0) \), and thus all of the variance in protein levels can be explained by mRNA levels and a single PTR ratio, both between and within proteins. That is, \( \overline{p}_{ij} = p_{ij} \). Using this notation, we explicitly state the total correlation between log scaled mRNA and log measured protein:
\[
R_T = \text{Cor}(p_{ij}, p_{ij})
\]

\[
= \frac{\text{Cov}(p_{ij} + r_{ij}, p_{ij})}{\sqrt{\text{Var}(p_{ij} + r_{ij})\text{Var}(p_{ij})}}
\]

\[
= \frac{\sqrt{\text{Var}(p_{ij})}}{\sqrt{\text{Var}(p_{ij} + r_{ij})\text{Var}(p_{ij})}}
\]

\[
= \frac{\sqrt{\text{Var}(p_{ij})}}{\sqrt{\text{Var}(p_{ij}) + \tau^2}}
\]

Where the last equality holds if \( r_{ij} \) is uncorrelated with \( p_{ij} \), that is, the amount of post-transcriptional regulation does not depend on the mean protein level. It is apparent that the correlation is close to 1 whenever \( \text{Var}(p_{ij}) \gg \text{Var}(r_{ij}) = \tau^2 \). This holds in practice because the mean-level variability is significantly larger than the within gene physiological variability.

On the other hand, the physiological correlation of scaled mRNA and measured protein for a given gene across tissues or conditions is significantly smaller. Although it has an analogous form, we now fix the gene index \( i = i_0 \):

\[
R_{P,i_0} = \text{Cor}(p_{i_0j}, p_{ij}\,|\,i = i_0)
\]

\[
= \frac{\text{Cov}(p_{i_0j} + r_{i_0j}, p_{i_0j})}{\sqrt{\text{Var}(p_{i_0j} + r_{i_0j})\text{Var}(p_{i_0j})}}
\]

\[
= \frac{\sqrt{\text{Var}(p_{i_0j})}}{\sqrt{\text{Var}(p_{i_0j} + r_{i_0j})}}
\]

\[
= \frac{\sqrt{\text{Var}(m_{i_0j} + T_{i_0})}}{\sqrt{\text{Var}(m_{i_0j} + T_{i_0}) + \tau^2}}
\]

58
Where the last line is due to the fact that now $T_{ij}$ is constant within gene. This correlation is small when the variation in mRNA levels, $Var(m_{ij})$, is small relative to the variation in the log PTR ratio, $\tau^2$.

In Figure 3.1c we plot $R_T^2$ and $R_P^2$ against $\frac{\tau^2}{Var(p_{ij}^*)}$. As $\tau^2$ increases, the fraction of the total protein variance explained by the mRNA levels decreases only slightly. A plot of $\bar{p}$ versus $p^\text{sim}$ is depicted in Figure 3.1c illustrating this phenomena. In data collected by Wilhelm et al.\textsuperscript{119}, we found variation in across-tissue PTR, $r_{ij}$ to account for almost 20% of total protein variance.

With real data, $\text{medianVar}(r_{ij}) \approx 0.17$ whereas $\text{medianVar}(m_{ij}) \approx 0.03$. This implies that for the median gene $R_P^2 = 0.15$ $(R_P = 0.39)$.

However, these variances are themselves quite variable across genes. As a consequence, $R_P$, $i$ is highly variable (Figure B.1b, Appendix).

**FUNCTIONAL GENE SET ANALYSIS**

To identify tissue-specific PTR for functional sets of genes, we analyzed the distributions of PTR ratios within functional gene-sets using the same methodology as Slavov & Botstein\textsuperscript{95}. We restrict our attention to functional groups in the GO ontology\textsuperscript{17} for which at least 10 genes were quantified by Wilhelm et al.\textsuperscript{119}. Let $k$ index one of these approximately 1600 functional gene sets. First, for every gene in every tissue we estimate the relative PTR (rPTR) or equivalently, the difference between log mean protein level and measured protein level:

$$\hat{r}_{ij} = p_{ij} - \text{median}(p_{ij'} - m_{ij'})$$

To exclude the possibility that $\hat{r}_{ij} = 0$ exactly, we require that $j' \neq j$. When the
estimated rPTR is larger than zero, the measured protein level in tissue $j$ is larger than the estimated mean protein level. Likewise, when this quantity is smaller than onezero, the measured protein is smaller than expected. Measured deviations from the mean protein level are due to both measurement noise and tissue specific PTR. To eliminate the possibility that all of the variation in the rPTR ratios is due to measurement we conduct a full gene set analysis.

For each of the gene sets we compute a vector of these estimated log ratios so that a gene set is comprised of

$$\mathcal{G}_{kj} = \{\hat{r}_{i_1j}, \ldots, \hat{r}_{i_{nk}j}\}$$

where $i_1$ to $i_{nk}$ index the genes in set $k$ and $j$ indexes the tissue type.

Let $KS(\mathcal{G}_1, \mathcal{G}_2)$ be the function that returns the p-value of the Kolmogorov-Smirnov test on the distribution in sets $\mathcal{G}_1$ and $\mathcal{G}_2$. The KS-test is a test for a difference in distribution between two samples. Using this test, we identify gene sets that show systematic differences in PTR ratio in a particular tissue ($j$) relative to all other tissues.

Specifically, the p-value associated with gene set $k$ in condition $j$ is

$$\rho_{kj} = KS(\mathcal{G}_{kj}, \bigcup_{j' \neq j} \mathcal{G}_{kj'})$$

To correct for multiple hypotheses, we computed the false discovery rate (FDR) for all gene sets in tissue $j^{105}$. In Figure 3.5, we present only the functional groups with FDR less than 2% and report their associated p-values. The significance of many of these groups, controlling for false discoveries suggests that not all of the variation in rPTR is due to measurement noise.
3.4 Discussion

Highly abundant proteins have highly abundant mRNAs. This dependence is consistently observed\textsuperscript{19,19,40,99,93} and dominates the explained variance in the estimates of absolute protein levels (Figure 3.3, top). This underscores the role of transcription for setting the full dynamic range of protein levels. In stark contrast, differences in the proteome between distinct human tissues are poorly explained by transcriptional regulation (Figure 3.3, bottom). Rather, the mechanisms shaping the tissue-specific proteomes involve post-transcriptional regulation. This result underscores the role of translational regulation for mediating physiological functions within the range of protein levels consistent with life.

The estimates of absolute protein levels are affected by technological biases and measurement error\textsuperscript{81,30} which can contribute to overestimating post-transcriptional regulation. These biases can difficult to estimate and influential\textsuperscript{19}, potentially leading to underestimates of the variance in protein levels explained by transcription. However, such systematic biases do not affect the relative changes of protein levels and the estimates of physiological variability. Indeed, the strong enrichment of rPTR within gene sets (Figure B.2) demonstrates a concerted regulation at the post-transcriptional level. It is thus unlikely that bias and measurement error completely explain the weak correlations between tissue-specific differences in mRNA and protein levels.
Conditionally Specified Bayesian Models
for Data Missing Not At Random

4.1 Introduction

Missing data problems arise in a number of areas including survey sampling, public health, policy and bioinformatics. How to model data with missingness depends cru-
cially on the mechanism by which missingness occurs. In particular, missing data problems can be divided into two classes: “ignorable” and “non-ignorable”. In this chapter, we focus on data which is missing not at random (MNAR), the most common type of non-ignorable data.

Data is said to be missing not at random when the missingness is dependent on the unobserved values. Examples of non-randomly missing data include informative dropout in longitudinal studies, nonresponse in surveys, or missingnessness in biological applications due to the technical challenges of measuring molecules that occur in low abundances.

In Bayesian inference, we can ignore the observation mechanism if the posterior distribution of the complete data estimands is conditionally independent of the response indicators given only the observed data. With non-ignorable missing data, on the other hand, we must specify a model for the joint distribution of the response indicators and values. The fundamental challenge with non-ignorable missing data, is that the missing data mechanism is not fully identifiable from the data without strong assumptions. As such, much of the literature on MNAR data focuses on the importance of assessing inferential sensitivity to different model specifications.

In this chapter, we explore flexible conditionally specified Bayesian models which we term “Tukey’s factorization models”. Under Bayesian inference, all sensitivity analyses can be conducted by specifying a prior over a set of parameters which index non-identifiable model components. Then, the posterior variance of all relevant parameters accurately reflects our ignorance about the true observation mechanism. This philosophy motivates the use of Bayesian models with Tukey’s factorization which facilitates prior specification for the most relevant factors.

In Section 4.5 we explore the utility of Tukey’s factorization in simulation. Then, in
Section 4.6 we demonstrate the use of conditionally specified models for missing high-throughput biological data. First, we start with a brief discussion of the two most common model specifications for MNAR data, before formally introducing the theoretical foundations of Tukey’s factorization in Section 4.3.

4.2 Models For Data Missing Not At Random

Mathematically, there are many equivalent ways to factorize a joint distribution. Nevertheless, model specification is usually motivated by the way in which the joint distribution is factorized. Models which can be easily specified using one factorization may be analytically intractable for another factorization of the same joint. As a general strategy, researchers should directly model the factors for which they have the most intuition or scientific expertise. This is especially important in missing data models for which certain assumptions are not checkable from the data, as is the case for non-likelihood identifiable missing data assumptions. This has motivated work on different ways of specifying non-ignorable missing data mechanisms. For simplicity, throughout this chapter, we focus on the the case in which there are no available covariates which inform missingness, though what follows can be extended to incorporate covariates.

Principally, the literature on non-random missing data has concentrated on two different ways of factorizing the joint distribution of missingness indicators and values. Throughout, let $Y = (y_1, y_2, \ldots, y_N)$ represent the complete data and $R = (r_1, r_2, \ldots, r_N)$ represent the response indicators for $Y$ (e.g. “missing” when $r=0$ or “observed” when $r=1$). Below, we describe these two classes which involve factorizing the joint distribution of $(R, Y)$ into marginal and conditional distributions. A more
detailed overview of these factorizations can be found in Molenberghs\textsuperscript{71}.

4.2.1 The Selection Factorization

Under the selection factorization, which is the most common way to specify models with non-ignorable missing data\textsuperscript{62}, the joint is factorized into the complete data distribution $P(Y)$ and the selection function $P(R|Y)$:

$$P(R, Y) = P(R|Y)P(Y)$$

where the selection function $P(R|Y)$ specifies the probability of observation as a function of $Y$. In practice, selection models are appealing because we often have strong intuition about the functional forms for these two factors. For instance, we may postulate that the complete data is unimodal and symmetric. By the central limit theorem, it may be reasonable in many cases to assume normality of the complete data– this is done often in practice\textsuperscript{42}. Furthermore, the estimands of interest are usually a function of the complete data distribution. In this sense, it is more natural to directly parameterizing the complete data in terms of quantities which are related to the relevant estimands.

We also usually have some intuition about the selection function, $P(R|Y)$. For instance, in many applied problems, the selection probabilities are montone in $Y$, and may be well approximated with a logistic function\textsuperscript{25}. Even when the missingness probabilities are non-monotone, it is almost always smooth in $Y$. Thus, it is reasonable to encode these beliefs about the joint distribution as model assumptions for selection factor directly.

Although the selection factorization is attractive for these reasons, there are of
course issues with the modeling strategy. For one, in practice, identifiability in this class of models is usually achieved by making specific parametric modeling choices for each factor. Consequently, one must check that the implied observed data distribution is consistent with the true observed data. In general, there are no guarantees that under the parametric constraints, the implied observed density will fit the data well. Even more, in this formulation, untestable assumptions are made about both factors, since neither factor alone is fully identifiable. Together, the complete data and selection function determine the observed data distribution, but individually, neither factor can be identified without the other. These concerns have been used as motivation for the other common class of non-ignorable missing data models.

4.2.2 The Pattern-Mixture Factorization

The pattern-mixture factorization is the other common alternative for missing data models. Here, we model the data as a mixture of observed and missing data components and specify the marginal probability of missingness:

\[ P(Y, R) = P(Y|R)P(R) \]

Both the observed data distribution, \( P(Y|R = 1) \), and the fraction of missing data, \( P(R) \), are well identified from the data. Thus, all unassessable assumptions in the pattern mixture formulation come from the specification of the missing data distribution, \( P(Y|R = 0) \). In this sense, sensitivity analysis under the pattern-mixture factorization is more straightforward.

On the other hand, although we have freedom to assess sensitivity of our estimates to different assumptions about the missing data distribution, it is very hard to sci-
entifically motivate any of these distributional choices. When the data belongs to a complicated (e.g. multimodal or non-continuous) distribution, we may have no insight into reasonable priors for the missing component. Rather, most of our intuition about the missing data distribution comes from our belief about the selection function. This suggests that we should utilize a model which involves the selection function. Below, we introduce a third approach which does just that, by combining some of the benefits of both selection and pattern-mixture models.

4.3 Tukey’s Factorization

While selection factorization and pattern-mixture factorization are the most obvious ways to factor the joint distribution of response indicators and data, there are other ways to factorize this distribution. An appealing alternative was first suggested by John Tukey in a discussion of the work by Glynn et al.\textsuperscript{39}. Specifically, Tukey demonstrated, using Bayes rule, that the joint can also be specified in terms of conditionals:

\[ P(Y, R) \propto \frac{P(R|Y)P(Y|R = 1)}{P(R = 1|Y)} \]  \hspace{1cm} (4.1)

with normalizing constant \( P(R = 1) \) ensuring that the joint distribution integrates to one. This factorization is appealing in part because it represents a hybrid of the selection factorization and the mixture factorization.

First, this factorization includes an explicit term for the selection function, \( P(R|Y) \). As noted above, we generally have more domain knowledge and intuition about the form of the missingness mechanism. In most real problems, even when the complete data distribution is complicated and possibly non-continuous the missingness mechanism is smooth. Second, the factorization includes a term for observed data distri-
bution, \(P(Y|R = 1)\), which is well identified from the data. Unlike the selection factorization, all of the unidentifiable uncertainty is associated with the selection mechanism. Finally, unlike the standard pattern mixture formulations, there is no need to explicitly make unjustifiable choices about the missing data distribution.

When there is little direct knowledge about the form of the complete or missing data distributions, but we have some intuition about the missingness mechanism, the focus of the model should be on developing reasonable priors for this missingness mechanism. Further, models for the observed data distribution are easy to check and fully identified from the data. This motivates the utility of Tukey’s factorization.

Interestingly, Tukey’s factorization can be derived through a simple application of Brook’s lemma, which equates a joint distribution to the set of full conditionals\(^\text{13,7}\). Specifically, the formula specifies that, for some joint distribution \(P(\mathbf{W})\)

\[
P(\mathbf{W}) = \prod_{i=1}^{n} \frac{P(W_i|W_1, \ldots, W_{i-1}, w_{0,i+1}, w_{0,n})}{P(w_{0,i}|W_1, \ldots, W_{i-1}, w_{0,i+1}, w_{0,n})}
\]  

(4.2)

where \(w_0\) is any point in the support of \(\mathbf{W}\). In the missing data context, if we let \(\mathbf{W} = (Y, R)\) and \(w_0 = (y, 1)\), Brook’s lemma simplifies to (4.1). In general, the ordering of the variables in Equation 4.2 are arbitrary. The cleverness of Tukey’s factorization is that, by ordering the variables appropriately and fixing \(w_{0,2}\) at \(R = 1\), the specification only involves one factor which is not likelihood identifiable.

Although Brook’s lemma is more commonly referenced in the theory of spatial autoregressive models\(^\text{18}\), its connection to Tukey’s factorization is relevant because much of the established theory about specifying joint distributions through conditional distributions carries over. Importantly, Brook’s Lemma is only applicable when the positivity condition is satisfied. This condition ensures that the denominator of Equation

68
4.2 is always nonzero. With Tukey’s factorization, this means that for all \((r, y)\)

\[
\text{If: } P(R = r) > 0 \text{ and } P(Y = y) > 0
\]

\[
\text{Then: } P(R = r, Y = y) > 0
\]

Intuitively, the observed data distribution cannot be informative about the missing data distribution when the support of missing data is not a subset of the support of the observed data. This condition is not trivially satisfied in all missing data problems. Importantly, Tukey’s factorization cannot be applied to models where \(P(R = 1|Y < c) = 0\), deterministically, for some cutoff \(c\). Consequently, we necessarily focus on problems where \(P(R = 1|y) > 0\) for all \(y\), that is, where the support of the missing data is contained in the support of the observed data.

Finally, the full conditional distributions specified in Equation 4.2 must also imply an integrable joint density\(^7\). With Tukey’s factorization, the integrability condition constrains the rate at which the tails of the observed data decrease relative to the rate at which the odds of missingness increase. We give an explicit example of the integrability condition in Section 4.4.1 and briefly discuss the implications of this condition in Section 4.5. In this chapter, we focus on inference under Bayesian models for non-ignorable missing data using Tukey’s factorization and thus begin by formulating it in this context.

### 4.4 Bayesian Models for Missing Data

Throughout this chapter, let \(\beta\) be a set of parameters specifying the selection mechanism, \(r(\beta, y) = P(R Y = y, \beta)\) and let \(\theta\) be a set of parameters identifying the observed data distribution, \(f_{obs} = P(Y|R = 1, \theta)\). Using Equation 4.1 we can write
the complete data likelihood as

\[ L(\beta, \theta; y, R) = r(y, \beta)^{R-1}(1 - r(y, \beta))^{1-R} f_{\text{obs}}(y|\theta_1)p(\theta_1, \beta) \]  

(4.3)

with inverse normalizing constant

\[
p(\theta_1, \beta) = \frac{1}{\left( \sum_{R=0}^{1} \int r(y, \beta)^{R-1}(1 - r(y, \beta))^{1-R} f_{\text{obs}}(y|\theta_1)dy \right)} \]  

(4.4)

Here \( p(\theta, \beta) \) corresponds to the population fraction of observed data. For i.i.d data, the observed likelihood then, is simply:

\[
L(Y_{\text{obs}}, R) = \prod_{i} [f_{\text{obs}}(y_i|\theta_1)p(\theta_1, \beta)]^R(1 - p(\theta_1, \beta))^{(1-R)}
\]  

(4.5)

\[
= (1 - p(\theta_1, \beta))^{N_{\text{miss}}} p(\theta_1, \beta)^{N_{\text{obs}}} \prod_{i=1}^{N_{\text{obs}}} f_{\text{obs}}(y_i|\theta_1)
\]  

(4.6)

Finally, if we equate the marginals in the mixture factorization and Tukey’s factorization, then we get the implied missing data distribution for this factorization as shown in Wainer\textsuperscript{112} and\textsuperscript{71}:

\[
f_{\text{miss}}(y|\beta, \theta_1) = \frac{p(\theta_1, \beta)}{1 - p(\theta_1, \beta)} \frac{1 - r(y, \beta)}{r(y, \beta)} f_{\text{obs}}(Y, \theta_1)
\]  

(4.7)

When the selection function does not depend on, \( y \), then the odds are constant in \( y \) and the missing data distribution matches the observed data distribution. Here
the distinction between MNAR and MCAR (or MAR when covariates are available) can be viewed explicitly as a function of the odds of missingness. When \( \frac{1-r(y,\beta)}{r(y,\beta)} \) has low variance, it may be a reasonable approximation to assume the data is MCAR (or MAR).

Unfortunately, the expression for the joint density in conditionally specified models involve a normalizing constant that is generally difficult to compute (Equation 4.4). Consequently, we start by introducing a class of models for which computation of this normalizing constant is tractable.

### 4.4.1 Exponential Family Models

In this section, we assume that the observed data belongs to an exponential family and that the missingness probabilities are logistic in the sufficient statistics of the observed data family. This class of models is closely related to the exponential tilt pattern-mixture models introduced by Birmingham et al.\(^8\). In this setting, we are able to analytically relate the missing data parameters to the population fraction of missing data.

Let \( f_{\text{obs}}(y|\theta) \) be of exponential family form. First, we reparameterize this density in terms of the natural parameters, \( \eta \). Then,

\[
f_{\text{obs}}(y|\eta) = h(x)g(\eta)e^{\eta T(y)}
\]

where, \( g(\eta) \) is the partition function, \( \eta \) is the natural parameter, and \( T(y) \) is sufficient statistic.

Also, we assume that

\[
r(y,\beta) = \text{logit}^{-1}(\beta T(y))
\]

71
Under these assumptions, the normalizing constant from equation 4.4 can be computed analytically:

\[
p(\eta, \beta) = \frac{1}{1 + \int h(x) g(\eta) e^{\eta T(y)} e^{\beta T(y)}}
\]

\[
= \frac{1}{1 + g(\eta) \int h(x) e^{(\beta + \eta) T(y)}}
\]

\[
= \frac{1}{1 + \frac{g(\eta)}{g(\beta + \eta)}}
\]

\[
= \frac{g(\beta + \eta)}{g(\beta + \eta) + g(\eta)}
\]

Given this expression for the normalizing constant, we can now analytically specify the posterior density of \((\eta, \beta)\). Moreover, all complete data estimands can be expressed as a function of these parameters, and thus any uncertainty about the missingness mechanism, specified through the prior on \(\beta\) will be propagated into the posterior uncertainty about the complete data estimands of interest.

Using Equation 4.7, it is easy to show that the missing data distribution is of the same exponential family form with natural parameter \(\eta^* = (\beta + \eta)\):

\[
f_{\text{miss}}(y|\eta, \beta) = h(x) g(\eta + \beta) e^{(\eta + \beta) T(y)}
\]

(4.9)

This results demonstrates that missing data imputation for these models is straightforward. Note that not all models for \(f_{\text{obs}}(y|\eta)\) and \(r(y, \beta)\) lead to proper joint distributions. In particular, in this setting, the integrability condition implies that \((\eta + \beta)\) is in the natural parameter space of the exponential family. We discuss this condition with a concrete example in Section 4.5.
Although this result is built on relatively strong parametric assumptions, they are assumptions that are used often in practice. Moreover, these assumptions can be generalized in two useful ways. First, we can take $f_{\text{obs}}$ to be a mixture of exponential families. In this case, we can build arbitrarily complex observed data models, while maintaining the analytic tractability of the posterior density. Relaxing assumptions about the missingness mechanism is more difficult. Still, we can define $r(y, \beta)$ to be a mixture of logistic functions or model the log-odds of missingness using quadratic splines to build even more flexible missingness mechanisms, if necessary. We explore models which exploit some of these generalizations in simulation, in Section 4.5, and on high-throughput biological data in Section 4.6.

4.4.2 Bayesian Inference for Non-Ignorable Missing Data

In addition to the modeling tradeoffs for the factorizations presented in Sections 4.2 and 4.3, it is important to consider the computational tradeoffs. One of the challenges with Bayesian inference under models specified using the selection factorization, is that the implied observed data density, $f_{\text{obs}}(y, \theta)$, often cannot be computed analytically. Accordingly, it is common to conduct Bayesian inference through data augmentation\textsuperscript{30,46}. That is, rather than specify the observed data posterior directly, sampling is done via Monte Carlo integration of the missing data by imputing missing values at each iteration of the sampler. Imputation can be difficult when the missing data density is complicated and convergence can be slow when the fraction of missing data is large.

Conversely, both Tukey’s factorization and pattern-mixture factorization models typically do not require Monte Carlo missing data integration since neither of these factorizations involve an explicit model for the complete data. Because the models
are specified for the observed data directly, no explicit missing data marginalization is necessary. As is the case for collapsed Gibbs samplers, the computational benefits of marginalization can be significant.

Thus, in theory, given an analytic solution to the normalizing constant, with models specified using Tukey’s factorization we can use standard MCMC methods to sample from the posterior distribution, without augmenting the parameter space with the number of missing values. Also, when the number of missing values is unknown, as is the case for stochastically truncated data, the prior on the parameters of the missingness mechanism implies a prior on the fraction of missing data. This also suggests a reasonable prior for the number of missing values.

With censored data, on the other hand, the amount of missing data is known and inference is generally more challenging. In this case, Equation 4.4 induces a soft moment constraint that implies that the posterior over the missing data parameters will concentrate around a manifold of dimension $D - 1$ for missing data parameters $\beta \in \mathbb{R}^D$. Since $p_{obs}$, the population fraction of observed data is usually highly identified from the data, we effectively lose one degree of freedom when specifying the prior parameters of the missingness mechanism. Figure 4.1 illustrates this phenomena for the model presented in Section 4.6. Sampling from this posterior is difficult, even with more advanced Bayesian methods. Nevertheless, both Riemannian manifold Hamiltonian Monte Carlo, which exploits the geometry of the posterior, or sequential Monte Carlo have been suggested as useful inferential methods with unidentifiable models and may provide a practical solution in this context.

In our analysis, we take a more computationally feasible approach. Rather than specify a prior on all $D$ parameters of the missingness mechanism, we specify a prior on $p_{obs}$ and $D - 1$ dimensions of the missing data mechanism and then use Equa-
\[
\logit r(y, \phi) = \phi_1 y + \phi_0
\]

Figure 4.1: The posterior distribution for missing data parameters are highly correlated when the population fraction of missing data is known to moderate precision. Here, the observed data is drawn from a standard normal and the missingness is logistic. The width of the ridge is determined by the information about the fraction of missing data, \( p_{obs} \). In this example \( N_{obs} = 100 \) and \( N_{miss} = 100 \).

tion 4.8 to numerically solve for the last missing data parameter. When this normalizing constant is analytically tractable, as is the case for the class of models in Section 4.4.1, this is a fast and simple calculation. Below, we use this approach to demonstrate the utility of Tukey’s factorization in simulation and problems in high-throughput biology.

4.5 Simulation: A Discrete-Continuous Mixture

In this section, we develop a simulation for which both selection factorization and the pattern mixture factorization models would be inappropriate. First we argue that it would be difficult to posit realistic models with the pattern mixture factorization
models. Then we show, through this simulation, that even the most flexible models with the selection factorization can lead to biased estimates. Lastly, we demonstrate how Tukey’s factorization can provide a reasonable and powerful alternative factorization for modeling non-ignorable missing data.

We start by specifying the true joint distribution of data and indicators via Tukey’s factorization. First, we define the observed data distribution as a discrete-continuous mixture where the continuous component is itself a mixture of normals. The observed data has density

\[
f_{\text{obs}}(\mu, \sigma, p, \alpha) = \alpha \left( \sum_{k=1}^{K} w_k N(y; \mu_k, \sigma_k^2) \right) + (1 - \alpha) \left( \sum_{m=1}^{M} p_m I[y = \gamma_m] \right)
\]

where \( \alpha \) is the fraction in the continuous component and \( p_m \) are the probabilities on the discrete locations \( \gamma_m \in \{-4, -3, \ldots, 4\} \).

In this simulation, we also assume the odds of missingness are non-monotone but linear in the first derivative:

\[
r(y, \beta) = \logit^{-1}(\beta_2(y - \beta_1)^2 + \beta_0)
\]

(4.10)

\( \beta_1 \) corresponds to the value at which an observation is most likely to occur, whereas \( \beta_2 \) controls how quickly the observation probability decays in \( y \). Non-monotone missingness in \( y \) may occur in practice, when extreme values, both small and large, are less likely to be observed.

By Equation 4.9, the missing data distribution (and hence the complete data distri-
bution) for the continuous component is also a normal mixture. In particular,

\[
f_{\text{miss}}(\mu, \sigma, p, \alpha, \beta) = \alpha^* \left( \sum_{k=1}^{K} w_k^* N(y; \frac{2\beta_1}{\sigma_k^2} \frac{\beta_2}{\sigma_k^2} + \mu_k, \frac{\sigma_k^2}{2\sigma_k^2} + 1) \right) + (1 - \alpha^*) \left( \sum_{d=1}^{D} p_d^* I[y = \gamma_d] \right)
\]

(4.11)

This derivation, with specifications for the mixture weights \( \alpha^* \) and \( w_k^* \), is given in the Appendix. Figure 4.2 depicts a histogram of draws from the observed data as well as the true observed, missing and complete data densities for the continuous component.

Note that by the integrability condition, in order for the component variances to be positive, we require that

\[
\beta_2 > \max_k \left( -\frac{1}{2\sigma_k^2} \right)
\]

(4.12)

Intuitively, for the implied complete data distribution to be valid, the tails of the observed data distribution must be decreasing faster than the odds of missingness are increasing. If we have a strong prior belief about the rate at which missingness probabilities are changing, the integrability conditions tells us how we should bound the tails of the model fit to the observed data. Alternatively, given the true observed data distribution, this condition constrains our subjective prior for the missingness mechanism.

With data generated from this process, modeling the data using a standard pattern mixture factorization would be difficult, at best. First, note, the implied missing data distribution in Equation 4.11 does not correspond to a simple location or scale change of the observed data distribution (although individually, each of the continu-
ous mixture components does). Thus, as is common under models specified using this factorization, the missing data cannot be represented with a “between pattern location and scale change”\textsuperscript{21}. In general, when we believe the missing data distribution is complicated, without very strong domain knowledge it would not be possible to propose plausible, scientifically justifiable priors directly for the missing data distribution.

What about modeling the data using the selection factorization? First, as noted in Section 4.4.2, inference under selection factorization models typically requires numerical or Monte Carlo integration to identify the observed data distribution, and hence are computationally more demanding than both the pattern mixture and Tukey’s factorization models. Further, by not directly modeling the observed data, it is more difficult to specify well-fitting models. An argument in favor of Tukey’s factorization over the selection factorization can be made on these grounds alone.

Here we investigate an additional possible pitfall of selection factorization models. As noted in Section 4.2.1 neither the complete data distribution nor the missing data parameters are fully likelihood identifiable. Because the unidentifiability is split across two model factors, it is much harder to use domain knowledge to inform accurate subjective priors.

In this simulation, we assume that we know the “true” prior for the selection function but have little a priori knowledge about the form of the complete data distribution. Under this setting, we demonstrate that different priors for the complete data density in the selection factorization model lead to biased inferences for both the posterior mean and variance of complete data quantities. Conversely, under Tukey’s factorization, we identify nearly unbiased estimate of complete data quantities and acknowledge appropriate posterior uncertainty.

For this example we let $\beta_2 = 0.06$, $\beta_1 = -2$ and $p_{obs} = 0.5$, that is 50% of the
complete data is missing. We find $\beta_0$ numerically given $\beta_2, \beta_1, p_{obs}$. Figure 4.2 depicts this true missingness mechanism, in addition to draws from the prior below:

$$\beta_2 \sim 0.08\text{Beta}(3, 1)$$

$$\beta_1 \sim N(-2, 2)$$

(4.13)

(4.14)

This prior is unbiased for the parameters of the true missing data mechanism. We take this to be the “true” prior, in the sense that the prior mean corresponds to the true data generating parameters, and the prior variance reflects our subjective uncertainty. We also assume that we know $p_{obs}$ and the observed data distribution exactly. As such, any complete data distribution and missingness mechanism which do not imply the true observed data distribution will have zero posterior probability.

In this context, we analyze the sensitivity of our inference for complete data quantities under a set of different priors for the complete data distribution and the selection prior specified in 4.13. In practice, with mixture models it is common to put a global conjugate normal prior on the locations of the mixture components in a normal mixture model. Motivated by this, we explore sensitivity of our results assuming that we the locations of the mixture components have biased normal priors:

$$\mu_k \sim N(\mu_{\text{complete}} + \Delta \mu, 1)$$

(4.15)

Here, $\mu_{\text{complete}}$ represents the true complete data mean, and $\Delta \mu$ is a tuning parameter corresponding to the distance of the prior component mean to the complete data
mean. To sample from the selection factorization model implied by 4.15 and 4.13, we use importance resampling on samples from the Tukey’s factorization model.

Figure 4.3 illustrates the corrupting effect of the selection factorization model when the prior for the complete data distribution is biased. In particular, this prior biases the complete data mean and variance, and the KL-divergence of the inferred discrete component relative to the true complete discrete distribution. Naturally, the bias is larger when the assumed complete data prior mean is further from the truth.

As we have argued, in many cases, standard pattern-mixture models are inappropriate for specifying the complete data distribution. Further, under the selection factorization, by mixing sources of unidentifiable uncertainty, we can bias our results. Tukey’s factorization addresses these issues. As long as we choose a sufficiently flex-
Figure 4.3: Distortion due to different assumed complete data priors. We use the “correct” prior over the missingness mechanism, but the results are skewed by bias in the complete data prior. 

- Boxplot of the complete data posterior mean, as a function of the distance of the prior complete data mean to the true complete data mean ($\Delta \mu$). 
- Boxplot of the posterior complete data standard deviation, as a function of $\Delta \mu$. 
- KL-divergence of the discrete component to the true discrete component distribution as a function of $\Delta \mu$. When we have prior knowledge about the selection mechanism but not the complete data distribution, Tukey’s factorization models are more appropriate.

...model for the observed data and specify a reasonable prior over the missingness mechanism we can conduct accurate posterior inference.

4.6 Application: Analysis of Transcriptomic and Proteomic Data

In experiments involving measurements of transcriptomic and proteomic data, mRNA transcripts and proteins which occur at low levels are less likely to be observed, that is, the data is missing not at random $^{114,100}$. This makes it challenging to infer normalizing constants for absolute protein levels $^{51}$, cluster genes into functionally related sets $^{108}$, infer the degree of coordination between transcription and translation $^{30}$, and determine the ratio of dynamic range inflation from transcript to protein levels $^{19}$.

Here, we demonstrate how Tukey’s factorization can be used to investigate some these issues by assessing the sensitivity of estimands to different assumptions about the missingness mechanism. In this analysis, we use transcriptomic data from Pelechano...
& Pérez-Ortín \textsuperscript{70} (14\% missing data) and protein abundance data from Ghaemmaghami et al. \textsuperscript{37} (34\% missing data).

It is standard to simply assume that both mRNA and protein levels are log-normally distributed\textsuperscript{6,66}, although this assumption may not be justified\textsuperscript{65,68}. Here, we instead model the observed data as a mixture of normals and specify a logistic prior for the missingness mechanism. Together these assumptions imply a prior over the complete data distribution.

As noted in Karpievitch et al. \textsuperscript{50}, missingness can occur for distinct reasons, at different stages of the data collection process. They find that a small fraction of missing proteomic data collected using mass spectrometry is missing completely at random. Consistent with this, we assume missingness has logistic form, but may asymptote at some value less than one. The observed data distribution and missingness function together define the joint distribution:

$$f_{\text{obs}}(y, \beta, \kappa, \mu_k, \sigma_k) \sim \sum_{k=1}^{K} w_k N(y; \mu_k, \sigma_k^2)$$  \hfill (4.16)

$$r(\beta_1, \beta_0, \kappa, y) = \frac{\kappa e^{\beta_1 y + \beta_0}}{1 + e^{\beta_1 y + \beta_0}}$$  \hfill (4.17)

where $0 < (1 - \kappa) < 1$ is the fraction of data that is missing completely at random, and $\beta_1$ describes the rate at which the odds of missingness change in $y$. Under this model, the missing data distribution is
\[ f_{\text{miss}}(y, \beta, \kappa, \mu_k, \sigma_k) = (1 - \kappa^*) f_{\text{obs}}(y, \mu_k, \sigma_k) + \kappa^* \left( \sum_{k=1}^{K} w_k^* N(y; \mu_k + \beta_1 \sigma_k^2, \sigma_k^2) \right) \]

(4.18)

(4.19)

This derivation, with specifications for the mixture weights \( w_k^* \) and \( \kappa^* \), is given in the Appendix.

In our analysis, we assume that there are \( K = 3 \) mixture components. Further, we assume the following prior for the missingness mechanism:

\[ p_{\text{obs}} \sim Beta(0, 0) \]  
\[ \kappa \sim p_{\text{obs}} + Beta(2, 1) \]  
\[ \beta_1 \sim Beta(1, 3) \]  

(4.20)

Note that \( \kappa \) must be greater than \( p_{\text{obs}} \) because the selection probabilities cannot be everywhere less than the population fraction of missing data. Empirically, we find that this prior spans a plausible range of complete data distributions. We implemented the sampler for the observed data normal mixture models using the probabilistic programming language STAN\textsuperscript{104} and ran all missing data sensitivity analyses post hoc, by sampling \( \beta_1 \) and \( p_{\text{obs}} \) from their respective posterior distributions. We numerically solve for \( \beta_0 \) using Equation C.15 (Appendix).

Figure 4.4 depicts the fit to the protein data, assuming the median prior value of \( \beta_1 \). For comparison, we include the selection factorization model used by Franks

83
Figure 4.4: Model fit to proteomic data from Ghaemmaghami et al.\textsuperscript{37} data using two different approaches: the selection factorization and Tukey’s factorization. a) Maximum likelihood fit, assuming normal complete data and logistic missingness. b) Tukey’s factorization with logistic missingness. The observed data is modeled as a three component mixture of normals. Grey lines represent draws of the missingness mechanism from the prior provided in Equation 4.20

et al.\textsuperscript{30}, in which they assume the complete log-data is normally distributed and that the missingness is logistic in $y$. Some model mis-fit is apparent in the selection model. Corresponding figures for the transcript dataset can be found in the appendix.

In Figure 4.5, we compare posterior estimates computed under our prior to estimates implied by an MCAR model and to the complete data log-normal model. In this example, the MCAR estimates and normal estimates bookend those implied by our model under Tukey’s factorization. Under the normal model, the complete data standard deviation is large and the mean is small. This can perhaps in part be explained by the model misfit evident in Figure 4.4. We also use posterior estimates of the complete data standard deviation from models fit to data from both Pelechano & Pérez-Ortín\textsuperscript{79} and Ghaemmaghami et al.\textsuperscript{37} to estimate the variability in the dy-
namic range ratios (Figure 4.6). These results are consistent with those in put forth by Csárdi et al.\textsuperscript{19} who suggest that translational regulation reflects amplification of protein levels.

![Histograms showing Mean and Standard Deviation](image)

**Figure 4.5:** Posterior samples of the complete data mean (right) and complete data standard deviation (left) for protein data\textsuperscript{37}. The MCAR estimates (red) and an estimate assuming normality of the complete data (blue) are shown as vertical lines for comparison. Under the chosen prior (4.20 ), the MCAR and normal complete data models represent opposite extremes of the complete data mean and standard deviation.

Ultimately, when using Tukey’s factorization, the primary modeling choice involves specifying how the odds of missingness change in the molecule abundance, $y$. This is a quantity about which we can usually gain some scientific intuition. Interestingly, in this setting, it may be possible to design experiments which inform better priors on the missingness by estimating missingness probabilities for known protein quantities over multiple experiments.

We make no explicit assumptions about the complete data distributions, and since we directly model the observed data, it is easy to both check and refine the model fit. Lastly, we avoid any difficult Monte Carlo integration that is often needed to conduct Bayesian inference under the selection factorization\textsuperscript{30}. As such, Tukey’s factorization
**Figure 4.6:** Posterior samples of the ratio of complete data standard deviations of protein levels and mRNA levels using Tukey’s factorization. For comparison we include, MCAR missingness (red) and the selection model fit (blue). These results are consistent with the range of estimated inflation values discussed in Csárdi et al.19

is a computationally feasible and often preferable way to encode modeling assumptions when handling missing data in high-throughput biological data.

### 4.7 Discussion

While most of the literature on MNAR data focuses on different applications of the selection factorization and the pattern-mixture factorization, Tukey’s factorization represents a flexible, computationally tractable alternative which facilitates sensitivity analysis and model checking. One of the biggest difficulties with using Tukey’s factorization is constructing models for which the normalizing constant (Equation 4.4) can be computed analytically. In this chapter we focused on exponentially-tilted models in which the missingness is logistic in the response and the observed data can be
well approximated by an exponential family or mixture of exponential families. We demonstrate how Tukey’s factorization can be used to encode non-monotone missingness mechanisms and flexibly model real data with complex distributional forms. The models can be used conduct tipping point analyses\textsuperscript{64} or simply to incorporate subjective model uncertainty via priors on the selection function.

Crucially, Tukey’s factorization is only useful when it is possible to put forth reasonable priors on the missingness mechanism. Translating expert knowledge or subjective belief into general functional forms can be challenging and in many cases, the logistic form may not be flexible enough. Still, there are other flexible functional forms for the missingness function which lead to analytically tractable models. Approaches include quadratic spline models on either the log-odds or directly on the odds space, correctly constrained, and should be further explored. In practice, Tukey’s factorization should be used in concert with strategies for expert prior elicitation\textsuperscript{75,55,76}. This is especially important when extending Tukey’s factorization to models with covariates, where it is even more challenging to elicit accurate covariate dependent missingness models.

In other settings, like the example presented in Section 4.6, we may be able to collect data which partially informs the specification for the missingness mechanism. As such, when possible, we can design experiments to partially estimate \( r(y, \beta) \). Along these lines, Tukey’s factorization may be useful in the context of multiphase inference, which is intimately related to problems in missing data\textsuperscript{8}. In these problems, when preprocessing data, it is often the case that we have strong knowledge (or control) of the selection function yet a weaker understanding of the underlying scientific model. The factorization presented in this chapter can be used by the preprocessor to summarize the composition of the selection function with the preprocessing procedure in a
way that allows a downstream analyst to make valid inferences.

In this chapter we focus on the relatively limited class of problems where the data is univariate and i.i.d. Extending this methodology to a broad class of multivariate missing data problems remains a challenge. However, Tukey’s factorization is easily extensible to montone missing data, since the observed data models can easily be replaced by conditional models. As such, related methods for longitudinal data based on frequentist bounds have been explored. All in all, Tukey’s factorization, which represents a hybrid of the selection and pattern mixture models is an under-researched yet promising alternative for modeling non-ignorable missing data.
A.1 Proofs

A.1.1 Theorem 1

Proof. Let $\psi_{1,2} = \text{Cor}(L_1, L_2)$. Then the covariance between the observed measurements $X_1$ and $X_2$ is $\text{Cov}(X_1, X_2) = G_1 G_2 \psi_{1,2}$ by equations 2.1–2.4. Finally, using
equation 2.8 for the observed data variance, we have

\[
\text{Cor}(X_1, X_2) = \frac{G_1 G_2 \psi_{1,2}}{\sqrt{G_1^2 + \xi_1 + \theta_1 \sqrt{G_2^2 + \xi_2 + \theta_2}}}
\]

(A.1)

\[
= \frac{\psi_{1,2}}{\sqrt{1 + (\xi_1 + \theta_1)/G_1^2} \sqrt{1 + (\xi_2 + \theta_2)/G_2^2}}
\]

(A.2)

\[
< \psi_{1,2}.
\]

(A.3)

\begin{proof}
In the proof, we don’t need to assume that the latent response variances are fixed, they can vary freely. This is because the \( G_k = \tilde{G}_k = 1 \) restriction already ensures identifiability. We denote this new model, without unit variances, by \( \mathcal{M}^p \). Similarly, our new misspecified model without unit variances is denoted by \( \tilde{\mathcal{M}}^p \). Models \( \mathcal{M} \) and \( \tilde{\mathcal{M}} \) will be a special case of the proof.

Bunke & Milhaud\(^{14}\) show, under mild conditions satisfied here, that when the MLE converges to a unique value, the posterior mean converges almost surely to that same value. Thus, instead of working with the posterior mean \( \tilde{\psi}_{1,2}^{\text{PM}} \), it suffices to prove that the inequality holds for the MLE: \( \tilde{\psi}_{1,2}^{\text{MLE}} \leq \psi_{1,2} \), with equality only if \( \xi = 0 \).

Consider multivariate normal data \( \mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_{2n}] \) generated from the true model \( \mathcal{M}^p \). Under \( \mathcal{M}^p \), \( \text{Cov}(\mathbf{X}) = \Sigma \) can be written as

\[
\Sigma = \Sigma_1 \oplus \Sigma_1 + \gamma \mathbf{1}_{2n} \mathbf{1}_{2n}',
\]

(A.4)

\[
\Sigma_1 = (\theta \mathbf{I}_{n/2} + \xi \mathbf{1}_{n/2} \mathbf{1}_{n/2}') \oplus (\theta \mathbf{I}_{n/2} + \xi \mathbf{1}_{n/2} \mathbf{1}_{n/2}') + \phi \mathbf{1}_n \mathbf{1}_n'.
\]
Figure A.1: The structure of the covariance matrix $\Sigma$ of the true model $M^p$ (left) and the covariance matrix $\tilde{\Sigma}$ of the misspecified model $\tilde{M}^p$ (right), from Theorem 2. The marginal variances of the replicates are $\sigma^2 = \theta + \xi + \phi + \gamma$ and $\tilde{\sigma}^2 = \tilde{\theta} + \tilde{\phi} + \tilde{\gamma}$ and the experiment covariances are $\epsilon = \xi + \phi + \gamma$. 
where $\oplus$ is the direct sum operator, and $\mathbf{1}_n$ is the constant one column vector with $n$ rows. In (A.4) we define $\psi_1 = \psi_2 = \gamma + \phi$. As noted at the beginning of the proof, we do not assume that $\psi_1 = \psi_2 = 1$, since fixing $G_k = \tilde{G}_k = 1$ ensures identifiability for both models. We also assume $\gamma \geq 0$, $\theta > 0$, $\xi \geq 0$, $\phi \geq 0$. The correlation between the responses can be written as $\psi_{1,2} = \gamma/(\phi + \gamma)$ now.

The misspecified model $\tilde{M}^p$ has covariance matrix $\tilde{\Sigma}$, with the following structure:

$$\tilde{\Sigma} = (\tilde{\theta} \mathbf{1}_n + \tilde{\phi} \mathbf{1}_n \mathbf{1}'_n) \oplus (\tilde{\theta} \mathbf{1}_n + \tilde{\phi} \mathbf{1}_n \mathbf{1}'_n) + \tilde{\gamma} \mathbf{1}_{2n} \mathbf{1}'_{2n}, \quad (A.5)$$

where $\tilde{\gamma} \geq 0$, $\tilde{\theta} > 0$, $\tilde{\phi} \geq 0$. Again, we do not assume $\tilde{\psi}_1 = \tilde{\psi}_2 = 1$ here, and $\tilde{\psi}_1 = \tilde{\psi}_2 = \tilde{\phi} + \tilde{\gamma}$, and the correlation between the responses is $\tilde{\psi}_{1,2} = \tilde{\gamma}/(\tilde{\phi} + \tilde{\gamma})$.

Figure A.1 shows the structure of both the true $\Sigma$ and the misspecified $\tilde{\Sigma}$.

First, we reparameterize $\tilde{\Sigma}$ in terms of its eigenvalues. We present three properties about the eigenstructure of $\tilde{\Sigma}$:

1. $\tilde{\Sigma} \mathbf{1}_{2n} = (\tilde{\theta} + n\tilde{\phi} + 2n\tilde{\gamma}) \mathbf{1}_{2n}$, so $1/\sqrt{2n} \mathbf{1}_{2n}$ is a normalized eigenvector of $\tilde{\Sigma}$ with eigenvalue $\tilde{\theta} + n\tilde{\phi} + 2n\tilde{\gamma}$. This can be seen easily by performing the matrix-vector product:

$$\tilde{\Sigma} \mathbf{1}_{2n} = \begin{bmatrix} (\tilde{\theta} + \tilde{\phi} \mathbf{1}_n \mathbf{1}'_n) \mathbf{1}_n \\ (\tilde{\theta} + \tilde{\phi} \mathbf{1}_n \mathbf{1}'_n) \mathbf{1}_n \end{bmatrix} + 2n\tilde{\gamma} \mathbf{1}_{2n} = \begin{bmatrix} (\tilde{\theta} + n\tilde{\phi}) \mathbf{1}_n \\ (\tilde{\theta} + n\tilde{\phi}) \mathbf{1}_n \end{bmatrix} + 2n\tilde{\gamma} \mathbf{1}_{2n} = (\tilde{\theta} + n\tilde{\phi} + 2n\tilde{\gamma}) \mathbf{1}_{2n}. \quad (A.6)$$

$$= (\tilde{\theta} + n\tilde{\phi}) \mathbf{1}_{2n} + 2n\tilde{\gamma} \mathbf{1}_{2n} = (\tilde{\theta} + n\tilde{\phi} + 2n\tilde{\gamma}) \mathbf{1}_{2n}. \quad (A.7)$$

2. Let $\mathbf{1}_{2n}^\pm$ be a column vector with $n$ ones on the top and $n$ minus ones on the bottom: $\mathbf{1}_{2n}^\pm = [\mathbf{1}_n| - \mathbf{1}_n]'$. Then $\tilde{\Sigma} \mathbf{1}_{2n}^\pm = (\tilde{\theta} + n\tilde{\phi}) \mathbf{1}_{2n}^\pm$, so $1/\sqrt{2n} \mathbf{1}_{2n}^\pm$ is a normal-
3. The remaining $2n - 2$ eigenvalues are all equal to $\tilde{\theta}$. To see this, we show that if a vector $\mathbf{v}$ is orthogonal to both $1_{2n}$ and $1^{\pm}_{2n}$, then it is an eigenvector of $\tilde{\Sigma}$ with eigenvalue $\tilde{\theta}$. We partition $\mathbf{v}$ into two blocks of equal size: $\mathbf{v} = [\mathbf{v}^\top | \mathbf{v}^\perp]^\top$. If $\mathbf{v}$ is orthogonal to $1_{2n}$, then its elements sum up to zero. If it is orthogonal to $1^{\pm}_{2n}$, then the elements of both $\mathbf{v}^\top$ and $\mathbf{v}^\perp$ sum up to zero as well. So we have

$$\tilde{\Sigma} \mathbf{v} = \begin{bmatrix} (\tilde{\theta} \mathbf{I}_n + \tilde{\phi} \mathbf{1}_n \mathbf{1}_n') \mathbf{v}^\top \\ (\tilde{\theta} \mathbf{I}_n + \tilde{\phi} \mathbf{1}_n \mathbf{1}_n') \mathbf{v}^\perp \end{bmatrix} + 0_{2n} = \begin{bmatrix} (\tilde{\theta} + n \tilde{\phi}) \mathbf{1}_n \\ (\tilde{\theta} + n \tilde{\phi}) \mathbf{1}_n \end{bmatrix} = (\tilde{\theta} + n \tilde{\phi}) \mathbf{1}^\pm_{2n}. \quad (A.9)$$

Thus, our eigendecomposition is $\tilde{\Sigma} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^\top + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^\top + \lambda_3 \mathbf{V}_3$, where

$$\begin{align*}
\lambda_1 &= \tilde{\theta} + n \tilde{\phi} + 2n \tilde{\gamma}, \\
\lambda_2 &= \tilde{\theta} + n \tilde{\phi}, \\
\lambda_3 &= \tilde{\theta}, \\
\mathbf{v}_1 \mathbf{v}_1^\top &= \frac{1}{2n} \mathbf{1}_{2n} \mathbf{1}^\top_{2n}, \\
\mathbf{v}_2 \mathbf{v}_2^\top &= \frac{1}{2n} \mathbf{1}^\pm_{2n} \mathbf{1}^\top_{2n}, \\
\mathbf{V}_3 &= (\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top) \oplus (\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top),
\end{align*} \quad (A.10-12)$$

and we will parameterize $\tilde{\Sigma}$ with $\lambda_1$, $\lambda_2$, $\lambda_3$ instead of $\tilde{\theta}$, $\tilde{\phi}$ and $\tilde{\gamma}$.

The likelihood of a covariance matrix $\Phi$, for $\mathbf{X}$ is given by

$$\mathcal{L}(\Phi | \mathbf{X}) = (2\pi)^{-N(n_1 + n_2)/2} \prod_{i=1}^{N} \left[ \det(\Phi)^{-1/2} \exp\left( -\frac{1}{2} \mathbf{X}_i \Phi^{-1} \mathbf{X}_i \right) \right], \quad (A.13)$$

93
and the log-likelihood is

\[
\log \mathcal{L}(\Phi|X) = -\frac{N(n_1 + n_2)}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \log \det(\Phi) - \frac{1}{2} \sum_{i=1}^{N} (X_i'\Phi^{-1}X_i),
\]

\[
= -\frac{N(n_1 + n_2)}{2} \log(2\pi) - \frac{N}{2} \log \det(\Phi) - \frac{1}{2} \mathrm{tr}\left((N-1)S\Phi^{-1}\right),
\]

(A.14)

where \(S\) is the sample covariance matrix of \(X\).

For a given data set, \(N\) is constant, so we can divide (A.14) by \(N/2\) and omit the constant terms to get

\[
\log \mathcal{L}(\Phi|X) \propto -\log \det(\Phi) - \mathrm{tr}\left(\frac{N-1}{N}S\Phi^{-1}\right).
\]

(A.15)

In the limit as \(N \to \infty\), \((N-1)/N \to 1\) and \(S \to \Sigma\) elementwise with probability 1. Since the likelihood function is continuous, the log-likelihood is simply

\[
\lim_{N \to \infty} \log \mathcal{L}(\Phi|X) \propto \log \tilde{\mathcal{L}}(\Phi|\Sigma) = -\log \det(\Phi) - \mathrm{tr}(\Sigma\Phi^{-1}),
\]

(A.16)

by the continuous mapping theorem. We maximize \(\log \tilde{\mathcal{L}}\) over \(\Phi\), where \(\Phi\) is of the same form as \(\tilde{\Sigma}\) (Eq A.5). For simplicity we use \(\tilde{\Sigma}\) instead of \(\Phi\) in the following.

We find the maximum of \(\log \tilde{\mathcal{L}}\) as a function of \(\lambda_1\), \(\lambda_2\) and \(\lambda_3\). Since the parameter space is not compact we first evaluate the log likelihood at the boundaries. It is easy to see that at each boundary, the log likelihood diverges to negative infinity:

\[
\lim_{\theta \to \infty} \log \tilde{\mathcal{L}} = \lim_{\phi \to 0} \log \tilde{\mathcal{L}} = \lim_{\gamma \to \infty} \log \tilde{\mathcal{L}} = \lim_{\tilde{\phi} \to \infty} \log \tilde{\mathcal{L}} = -\infty.
\]

(A.17)

If at least one of \(\tilde{\theta}\), \(\tilde{\phi}\) or \(\tilde{\gamma}\) goes to infinity, then at least one eigenvalue of \(\tilde{\Sigma}\) goes to infinity (see equations A.10–A.12), so \(-\log \det(\Phi)\) goes to negative infinity and the
second term of the log likelihood is a finite constant. If \( \hat{\theta} \to 0 \), then the first term goes to positive infinity, the second to negative infinity, but because of the logarithm, the second term dominates and the likelihood goes to negative infinity. This means that the likelihood has a global maximum in the interior of the parameter space.

At the MLE, the derivative of the log-likelihood must vanish. Differentiating the log-likelihood in terms of an arbitrary parameter \( p \) gives

\[
\frac{d \log \hat{L}}{dp} = - \text{tr} \left( \Sigma^{-1} \frac{d \hat{\Sigma}}{dp} \right) + \text{tr} \left( \Sigma^{-1} \frac{d \hat{\Sigma}}{dp} \Sigma^{-1} \Sigma \right),
\]

where we use the fact that

\[
\frac{d \det(\Sigma)}{dp} = \text{tr} \left( \Sigma^{-1} \frac{d \Sigma}{dp} \right), \quad \frac{d \Sigma^{-1}}{dp} = - \Sigma^{-1} \frac{d \Sigma}{dp} \Sigma^{-1}.
\]

The derivatives in terms of the three parameters are

\[
\frac{d \hat{\Sigma}}{d \lambda_1} = v_1 v_1', \quad \frac{d \hat{\Sigma}}{d \lambda_2} = v_2 v_2', \quad \frac{d \hat{\Sigma}}{d \lambda_3} = V_3.
\]

In the following, we use the fact that \( \Sigma v_1 = \lambda_1 v_1 \) and \( \Sigma^{-1} v_1 = 1/\lambda_1 v_1 \) and set the first partial derivative to zero.

\[
\frac{d \log \hat{L}}{d \lambda_1} = - \text{tr} \left( \frac{1}{\lambda_1} v_1 v_1' \right) + \text{tr} \left( \frac{1}{\lambda_1^2} v_1 v_1' \Sigma \right) = 0,
\]

which, using \( v_1 v_1' = 1/(2n) 1_{2n} 1_{2n}' \), simplifies to

\[- \frac{1}{\lambda_1} + \frac{1}{2n \lambda_1^2} (4n^2 \gamma + 2n^2 \phi + n^2 \xi + 2n \theta).
\]
From here we can easily see that the MLE of \( \lambda_1 \) is

\[
\lambda_1^{\text{MLE}} = 2n\gamma + n\phi + \frac{1}{2}n\xi + \theta. \tag{A.23}
\]

A similar argument leads to the MLE for \( \lambda_2 \):

\[
\lambda_2^{\text{MLE}} = n\phi + \frac{1}{2}n\xi + \theta. \tag{A.24}
\]

For the third parameter we have

\[
\frac{d \log \hat{L}}{d\lambda_3} = - \text{tr} \left( \frac{1}{\lambda_3} V_3 \right) + \text{tr} \left( \frac{1}{\lambda_3^2} V_3 \Sigma \right) = 0. \tag{A.25}
\]

The second term is the trace of

\[
\frac{1}{\lambda_3^2} V_3 \Sigma = \frac{1}{\lambda_3^2} \left[ \left( I_n - \frac{1}{n} 1_n 1_n' \right) \otimes \left( I_n - \frac{1}{n} 1_n 1_n' \right) \right] \Sigma = \frac{1}{\lambda_3^2} \left[ \Sigma - \frac{1}{n} (1_n 1_n' \otimes 1_n 1_n') \Sigma \right] = \frac{1}{\lambda_3^2} \left[ \Sigma - \frac{1}{n} \left( \frac{1_n 1_n' \Sigma}{n^2 \gamma 1_n 1_n'} \right) \right], \tag{A.26}
\]

and the trace itself is

\[
\text{tr} \left( \frac{1}{\lambda_3^2} V_3 \Sigma \right) = \frac{1}{\lambda_3^2} \left( 2n\theta + 2n\phi + 2n\xi + 2n\gamma - \frac{2}{n} (n^2 \phi + n^2 \gamma + \frac{n^2}{2} \xi + n\theta) \right) = \frac{1}{\lambda_3^2} \left( (2n - 2)\theta + n\xi \right). \tag{A.27}
\]

96
Using this with equation A.25, we get

\[ 2(n - 1) \frac{1}{\lambda_3} = \frac{1}{\lambda_3^2} (2(n - 2)\theta + n\xi), \quad \lambda_3^{\text{MLE}} = \theta + \frac{n}{2n - 2}\xi. \quad (A.30) \]

Going back to the original parameterization is easy:

\[ \tilde{\theta} = \lambda_3, \quad \tilde{\phi} = \frac{\lambda_2 - \lambda_3}{n}, \quad \tilde{\gamma} = \frac{\lambda_1 - \lambda_2}{2n}, \quad (A.31) \]

and yields

\[ \tilde{\theta}^{\text{MLE}} = \theta + \frac{n}{2n - 2}\xi, \quad \tilde{\phi}^{\text{MLE}} = \phi + \frac{n - 2}{2n - 2}\xi, \quad \tilde{\gamma}^{\text{MLE}} = \gamma. \quad (A.32) \]

To show that the posterior mean of the misspecified model underestimates the true correlation, we need to show that

\[ (\tilde{\psi}_{1,2}^{\text{PM}} = \tilde{\psi}_{1,2}^{\text{MLE}} = ) \frac{\tilde{\gamma}^{\text{MLE}}}{\tilde{\phi}^{\text{MLE}} + \tilde{\gamma}^{\text{MLE}}} = \frac{\gamma}{\phi^{\text{MLE}} + \gamma} \leq \frac{\gamma}{\phi + \gamma} = \psi_{1,2}. \quad (A.33) \]

This is equivalent to \( \tilde{\phi}^{\text{MLE}} \geq \phi \), which holds, with equality only if \( \xi = 0 \). This completes the proof. \( \square \)

A.1.3 Theorem 3

**Proof.** Denote \( \rho = \text{Cor}(X, Y) > 0 \) and \( \rho^{\text{obs}} = \text{Cor}(X^{\text{obs}}, Y^{\text{obs}}) \). Assume, without loss of generality, that \( X \) and \( Y \) have mean zero and unit variance. We can write \( Y \) as

\[ Y = \rho X + \sqrt{1 - \rho^2} Z, \quad (A.34) \]
where $Z$ is a standard normal, independent of $X$. By assumption, we only observe $X^{\text{obs}}$, with $\text{Var}(X^{\text{obs}}) = c$, with $0 < c < 1$. By equation A.34, we then have

$$Y^{\text{obs}} = \rho X^{\text{obs}} + \sqrt{1 - \rho^2} Z$$  \hfill (A.35)

$$\text{Var}(Y^{\text{obs}}) = \rho^2 \text{Var}(X^{\text{obs}}) + (1 - \rho^2) \text{Var}(Z)$$ \hfill (A.36)

$$= \rho^2 c + 1 - \rho^2 = (1 - \rho^2)(1 - c) + c > c.$$ \hfill (A.37)

Similarly, $\text{Cov}(X^{\text{obs}}, Y^{\text{obs}}) = \rho c$. Assuming $\rho > 0$, it is true that

$$\rho^{\text{obs}} = \frac{\text{Cov}(X^{\text{obs}}, Y^{\text{obs}})}{\sqrt{\text{Var}(X^{\text{obs}})}\sqrt{\text{Var}(Y^{\text{obs}})}} = \frac{c \rho}{\sqrt{c}\sqrt{\text{Var}(Y^{\text{obs}})}} < \frac{c \rho}{\sqrt{c}\sqrt{c}} = \rho.$$ \hfill (A.38)

\[\square\]

### A.2 Additional figures and tables
Figure A.2: Distribution of observed and imputed mRNA and protein levels, in different experiment, together with the logistic censoring probability. A: LEE protein abundance data, B: LU protein abundance data, C: CAUS mRNA expression data set.

Figure A.3: Principal component analysis of the mRNA replicates. Only mRNAs that were measured in all replicates, are included here, 390 genes in total. It is clear that most of the variation is according to the lab, where the experiment was performed. In the case of the Lipson and Ingolia labs, the two and three batches are also apparent and motivate our choice to treat these as separate experiments.
<table>
<thead>
<tr>
<th># obs</th>
<th># prot.</th>
<th>cumul.</th>
<th># prot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>813</td>
<td>813</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>445</td>
<td>1258</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>249</td>
<td>1507</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>131</td>
<td>1638</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>79</td>
<td>1717</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>72</td>
<td>1789</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>129</td>
<td>1918</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>334</td>
<td>2252</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>689</td>
<td>2941</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>624</td>
<td>3565</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>453</td>
<td>4018</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>342</td>
<td>4360</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>290</td>
<td>4650</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>235</td>
<td>4885</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>180</td>
<td>5065</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>191</td>
<td>5256</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>194</td>
<td>5450</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>204</td>
<td>5654</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>135</td>
<td>5789</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>49</td>
<td>5838</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>16</td>
<td>5854</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># obs</th>
<th># mRNAs</th>
<th>cumul.</th>
<th># mRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>37</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>144</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>24</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>22</td>
<td>190</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>218</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>36</td>
<td>254</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>41</td>
<td>295</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>56</td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>33</td>
<td>384</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>414</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>435</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>30</td>
<td>465</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>488</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>40</td>
<td>528</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>49</td>
<td>577</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>76</td>
<td>653</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>90</td>
<td>743</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>107</td>
<td>850</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>160</td>
<td>1010</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>210</td>
<td>1220</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>342</td>
<td>1562</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>370</td>
<td>1932</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>497</td>
<td>2429</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>685</td>
<td>3114</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>924</td>
<td>4038</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>809</td>
<td>4847</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>617</td>
<td>5464</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>390</td>
<td>5854</td>
<td></td>
</tr>
</tbody>
</table>

**Table A.1:** Details about missing data. The tables show the number of proteins (left) and mRNAs (right) with a given number of observations. The number of observations is in the first columns, the number of proteins/mRNAs with that many observation in the second columns, and the number of proteins/mRNAs with at most that many observations in the third columns.
B

(Chapter 2)
Figure B.1: (a) Histogram for each of 12 different tissues. The between-gene correlations ($R_B$) of measured protein with unscaled protein are generally low (red), but after scaling mRNA by the median PTR correlations are quite high (blue). (b) Within-gene correlations ($R_W$) of measured protein and scaled mRNA are highly variable.
Figure B.2: Summary of rPTR variability, as depicted in panel (a-b), across all tissues and many gene ontology (GO) terms. Metabolic pathways and functional gene-sets that show statistically significant (FDR < 2%) variability in translational efficiency across the 12 tissue types.
C.1 Derivations for Section 4.5

Here, we derive the missing data distribution from simulation study specified in Equation 4.11 of Section 4.5. First, we start assuming the observed data can be fit by a single normal distribution. For the normal distribution expressed as an exponential family, we have:
\((\eta_1, \eta_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})\) \hspace{1cm} (C.1)

\(T(y) = (y, y^2)\) \hspace{1cm} (C.2)

\(g(\eta) = \frac{1}{\sqrt{-2\eta_2}} e^{\frac{\eta_1^2}{2\eta_2}}\) \hspace{1cm} (C.3)

where \(\eta\) are the natural parameters, \(T(y)\) the sufficient statistics and \(g(\eta)\) is the partition function. By Equation 4.10 the odds of missingness are

\[
\frac{1 - r(y, \beta)}{r(y, \beta)} = \beta_2 y^2 - 2\beta_2 \beta_1 y + \beta_1^2 + \beta_0
\]

and by Equation 4.9 the implied missing data distribution is normal with natural parameters

\((\eta_1^*, \eta_2^*) = (\eta_1 - 2\beta_2 \beta_1, \eta_2 + \beta_2)\)

The inverse parameter mapping for the normal specifies that

\((\mu, \sigma^2) = \left(\frac{-\eta_1}{2\eta_2}, \frac{-1}{2\eta_2}\right)\) \hspace{1cm} (C.5)

and thus the moments of the missing data distribution are

\((\mu^*, \sigma^{*2}) = \left(\frac{\mu}{\sigma^2} + 2\beta_2 \beta_1}{2(\frac{1}{2\sigma^2} + \beta_2)}, \frac{1}{2(\frac{1}{2\sigma^2} + \beta_2)}\right)\) \hspace{1cm} (C.6)

\[
= \left(\frac{2\beta_1 \beta_2 \sigma^2 + \mu}{\sigma^2 \beta_2 + 1}, \frac{\sigma^2}{2\sigma^2 \beta_2 + 1}\right)
\]

(C.7)
Next, we extend the results of Section 4.4.1 to mixture models. First, note by Equation 4.9, when the observed data distribution is a mixture of exponential families, the missing data distribution is also a mixture of those same exponential families. Further, using Equation 4.8, applied to mixtures, we have

\[
p(\eta, w, \beta) = \frac{1}{1 + \sum_k w_k h(x)g(\eta_k)e^{\eta_kT(y)}(e^{(\beta_2y^2 - 2\beta_2\beta_1y + \beta_1^2 + \beta_0)})} \tag{C.8}
\]

\[
= \frac{1}{1 + \sum_k w_k \frac{g(\eta_k)}{g(\eta_k + \beta)}} \tag{C.9}
\]

Lastly using Equation 4.7, we can show that the mixture weights are simply

\[
w_k^* = \frac{w_k \frac{g(\eta_k)}{g(\eta_k + \beta)}}{\sum_k w_k \frac{g(\eta_k)}{g(\eta_k + \beta)}} \tag{C.10}
\]

\[
\alpha^* = \frac{\alpha \sum_k w_k \frac{g(\eta_k)}{g(\eta_k + \beta)}}{\alpha \sum_k w_k \frac{g(\eta_k)}{g(\eta_k + \beta)} + (1 - \alpha) \sum_d h(p_d, \gamma_d, \beta)}
\]

with \( h(p_d, \gamma_d, \beta) = e^{\log p_d + \beta_2(\gamma_d - \beta_1)^2 + \beta_0} \). Together, Equations C.10, and C.6 yield the missing data distribution specified in 4.11.

C.2 Derivations for Section 4.6

Here, we derive the missing data distribution from Equation 4.18. In addition to assuming that the observed data can be represented as a mixture of normals, we assume that the odds of missingness can also be represented by a mixture. Specifically, we allow for some missingness to occur completely at random (MCAR). We posit that,
The mechanism is logistic, but asymptotes at some value $\kappa < 1$, where $(1 - \kappa)$ represents the fraction of missing data that is missing completely at random. Under this model, the odds of missingness are then,

\[
\frac{1 - r(\beta_1, \beta_0, \kappa, y)}{r(\beta_1, \beta_0, \kappa, y)} = \left(1 - \frac{\kappa e^{\beta_1 y + \beta_0}}{1 + e^{\beta_1 y + \beta_0}}\right) \frac{1 + e^{\beta_1 y + \beta_0}}{\kappa e^{\beta_1 y + \beta_0}}
\]

\[
= \frac{1 + (1 - \kappa)e^{\beta_1 y + \beta_0}}{\kappa e^{\beta_1 y + \beta_0}}
\]

\[
= \frac{1}{\kappa} e^{-\beta_1 y - \beta_0} + \frac{1 - \kappa}{\kappa}
\]

From Equation 4.8 applied to a mixture of normals,

\[
p(\eta, \beta) = \frac{1}{1 + \int \sum_k w_k h(x) g(\eta) e^{-\beta_1 y - \beta_0} (\frac{1}{\kappa} e^{-\beta_1 y - \beta_0} + \frac{1 - \kappa}{\kappa})}
\]

\[
= \frac{1}{1 + \left(\sum_k w_k \frac{g(\eta)}{g(\beta + \eta)}\right) + \frac{1 - \kappa}{\kappa}}
\]

Finally using Equation 4.7 we find that the missing data is a mixture normals with weights

107
Figure C.1: Model fit to mRNA data from Pelechano & Pérez-Ortín data using two different approaches: the selection factorization and Tukey’s factorization. (a) Maximum likelihood fit, assuming normal complete data and logistic missingness. (b) Tukey’s factorization with logistic missingness. The observed data is modeled as a three component mixture of normals. Grey lines represent draws of the missingness mechanism from the prior provided in Equation 4.20

\[
\kappa^* = \frac{\sum_k w_k g(\eta_k)}{\sum_k w_k g(\beta + \eta_k) + 1 - \kappa}
\]

\[
w_k^* = \frac{w_k g(\eta_k)}{\sum_k w_k g(\eta_k + \beta)}
\]

C.3 Additional results from Section 4.6
Figure C.2: Posterior samples of the complete data mean (right) and complete data standard deviation (left) for mRNA data\textsuperscript{79}. The MCAR estimates (red) and an estimate assuming normality of the complete data (blue) are shown as vertical lines for comparison. Under the chosen prior (4.20 ), the MCAR and normal complete data models represent opposite extremes of the complete data mean and standard deviation.
References


This thesis was typeset using \LaTeX, originally developed by Leslie Lamport and based on Donald Knuth’s \TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont’s humanist typeface. The above illustration, “Science Experiment 02”, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.